Exploring Perceptual Speed, and the effects of Visual Clutter, during Interactive Information Retrieval

Olivia Katherine Foulds
Computer and Information Sciences
University of Strathclyde

A thesis submitted for the degree of
Doctor of Philosophy

Glasgow 2023

Declaration of Authenticity and Author's Rights

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: *Olivia Foulds*

Date: 1st June 2023.

Acknowledgements

"*Nemo sibi nascitur*" —(No one is born unto oneself alone).

This phrase is very true, and I am thankful to everyone who has contributed to my research in some form or another: every supervisor—in both the academic and industrial worlds; colleague; friend; and family member. In particular, I must express my gratitude to my primary academic supervisor, Professor Ian Ruthven: their expertise has been instrumental in shaping my professional growth and development.

In terms of industry, this work was part funded by BAE Systems Maritime and EPSRC as part of an Industrial Cooperative Award in Science & Technology (CASE) Studentship (EP/S513908/1). This collaboration provided numerous opportunities—alongside invaluable guidance and support—and I have been privileged to work with many talented individuals and teams throughout the company.

"Reach the unreached" — Professor Wallace Foulds, CBE

Abstract

Perceptual Speed (PS) is a cognitive ability defined by an individual's accuracy and speed to scan information while completing visual search tasks. Prior studies using PS tests have demonstrated that PS affects multiple factors in Interactive Information Retrieval (IIR), such as users with Low-PS achieving worse search performance, increased time spent completing tasks, and perceiving more negative experiences. However this thesis systematically analyses how PS tests have been used in IIR, and identifies multiple problems with their reliability and validity; both in their measurement, and the overall results known in previous literature. Consequently, from a range of challenges and recommendations discussed, this thesis details the design process for creating new PS tests and implements the new tests alongside an experiment. The experiment was designed to better understand PS during IIR, through manipulating the presence of, and type, of visual clutter visible during a search. Specifically, users were required to find relevant information on an IIR system, across 4 different interfaces which varied: clutter that was absent; congruent with the task; incongruent with the task; or a mixture of both. In all conditions, the type of clutter was operationalised through visible advertisements. The results indicated that users with Low-PS were significantly negatively affected when clutter was present, as opposed to absent. These differences were most prominent when clutter was incongruent. In contrast, users with High-PS performed their best, both objectively and subjectively, when incongruent clutter was visible. Overall, these findings suggest that visual clutter can significantly impact the efficiency of information retrieval depending on a user's PS. These results have implications for the design of displays and interfaces, emphasizing the importance of altering the visible visual clutter to improve the accuracy, speed, and user experience of information processing for users with different perceptual abilities.

Contents

List of Figures

List of Tables

Abbreviations

PS Perceptual Speed.

IR Information Retrieval.

IIR Interactive Information Retrieval.

HCI Human Computer Interaction.

SUI Search User Interface.

Low-PS A user with Low Perceptual Speed.

High-PS A user with High Perceptual Speed.

FA One type of Perceptual Speed test, called *Finding A's*.

NC One type of Perceptual Speed test, called *Number Comparison*.

Low-FA A user with Low Perceptual Speed, as measured by the *Finding A's* test.

High-FA A user with High Perceptual Speed, as measured by the *Finding A's* test.

Low-NC A user with Low Perceptual Speed, as measured by the *Number Comparison* test.

High-NC A user with High Perceptual Speed, as measured by the *Number Comparison* test.

Ad Advert.

Ads Adverts.

NA No-ads.

AA All-ads.

CA Congruent-ads.

IA Incongruent-ads.

MA Mixed-ads.

Part I

Background

Chapter 1

Introduction

1.1 Chapter 1 Overview

This Chapter provides a brief overview of the overarching research area that motivated the present research, followed by the main research questions investigated, and the overall methodological approach undertaken to explore these questions. This then leads onto a short description of the main contributions that the thesis will make to the field, as a result of the results found. The chapter then concludes with a thesis outline, which provides a road-map of the chapters and sections that will follow, in addition to a section which explicitly outlines which chapters have already been published through peer-review.

1.2 Research Motivation

Information Retrieval (IR) is a field that has evolved rapidly in recent years due to the explosion of digital information and the increasing need for effective access and management of this information. Although the performance of retrieval systems has dominated many traditional studies in IR, the field of *Interactive* Information Retrieval (IIR) shifts focus towards the user. Specifically, as explained by Kelly et al. [120]:

"*IIR focuses on users' behaviors and experiences — including physical, cognitive and affective — and the interactions that occur between users and systems, and users and information*".

During IIR, a user must navigate through and process a variety of information in order to achieve their search goal [69]. Goals can range from everyday websearch, to more critical goals such as maritime operators in high-pressured information environments acting upon arising events. However, it has long been known that human brains are limited and susceptible to cognitive overload [114, 117]. This then means that a user cannot process everything they encounter—especially when too much visual clutter is present [198].

However, some people are naturally better able to process visual information and are said to have 'High' "Perceptual speed" (PS), which is one type of cognitive ability [10]. Thus in the context of IIR, if a user has 'Low' PS, then they are at a disadvantage in retrieving

information: they take longer [4, 46]; achieve lower search accuracy [11, 56]; and more negative user experiences are reported [71, 225].

With the long-term goal of creating dynamic search systems that can adapt to a user's unique cognitive ability, the present thesis sought to better understand the overall concept of PS, including how best to measure it, and how different types of visual clutter could help, or hinder users, depending on their PS ability.

1.3 Research Questions and Methodological Approach

This section describes the research questions under investigation and briefly highlights the various methodological approaches undertaken throughout the entire PhD process. For more specific detail about each approach taken, these will be described in each corresponding chapter.

As a result of the initial literature review (which will be presented in Chapter 2, Section 2.9), the overall goal of this research was to advance understanding of the cognitive ability Perceptual Speed (PS), in relation to how it affects Interactive Information Retrieval (IIR) by answering three main research questions:

- **(RQ1) How can a user with Low Perceptual Speed be helped to achieve a more positive online search experience, both subjectively, and objectively?**;
- **(RQ2) What is the relationship between Perceptual Speed and visual clutter in the form of advertisements during an IIR task?;**
- **(RQ3) Does clutter that is congruent with the task improve or worsen the search experience for users with Low Perceptual Speed?**

In order to answer these main questions, various sub research questions were explored using different methodological approaches.

Firstly considering RQ1, a more in-depth evaluation of previous literature was undertaken to identify: ***(1a) How has PS previously been measured during IIR?*** This involved a qualitative approach, where after familiarising oneself with the available literature, themes began to emerge which highlighted that measuring PS encompassed some uncertainties regarding PS test content, administration, analysis, and how results were reported—and consequently, the reliability and validity of PS measurement was questioned (Please refer to Chapter 4).

The uncertainties identified with PS measurement then led onto the next phase of research, where it was speculated that if the PS *tests* lacked reliability and validity, then the concept as a whole needed further investigation. Consequently, the first Systematic Review in the field of PS was conducted, which explored two sub-questions: ***(1b) What claims have occurred regarding PS in Computer Science?*** (considering the significance of different dependent variables including search performance, user experience, search behaviour, physiology, and any

interactions with other variables); and ***(1c) How can the results be explained?*** (Please refer to Chapter 5).

With confirmation that PS was a significant factor during IIR, it became evident that in order to fully answer RQ1, new PS tests needed to be created, leading to the following further sub research questions: ***(1d) How can a digital measurement of PS be created?***; Here, the research approach moved beyond secondary research, where previous literature was examined, and onto primary research, where new PS tests were designed and developed, in accordance with everything learned from the previous chapters. For example, where specific concerns of previous PS tests had previously arisen (E.g. See Chapter 4, Section 4.5), these were addressed and accounted for in the current design phase (Please refer to Chapter 6).

With a clearer understanding of how to measure PS, and having created new measurement tools, this provided a clear framework to answer RQ1. Combined with knowledge gained from previous literature, and results that emerged through the Systematic Review, in order to understand how to help users with Low-PS achieve a more positive search experience, it was concluded that PS needed to be investigated in relation to the effect of visual clutter. This inspired the empirical component of the PhD research, where an experiment was created to determine the effect of visual clutter, in the form of advertisements, on PS (RQ2). Here, effects were explored both quantitatively (through objective measures of search behaviour and performance) and qualitatively (through survey analysis of user experience). Furthermore, given that two different PS tests had been created, analysis considered these separately under a further sub-question: ***(2a) Are there different effects on search outcomes, based on different PS tests?*** (Please refer to Chapter 7).

Finally, with hypotheses generated which implied that different types of clutter may affect users with different levels of PS differently, the same experiment used to answer RQ2 was also used to answer RQ3. Similarly, differences between different PS tests were analysed with the following sub research question: ***(3a) Do different types of clutter impact users with different types of PS, as measured by different tests?*** (Please refer to Chapter 8).

As a summary of the overall thesis approach to answering the research questions, please refer to Figure 1.1.

1.4 Contributions

By answering the research questions, the results from this thesis have multiple implications that can benefit a variety of audiences.

Firstly, by providing the first Systematic Review into the field of PS during IIR, the results contribute to the broader literature on Interactive Information Retrieval. Specifically, these results will then inform future research, theory, and practice in the field, with a particular focus

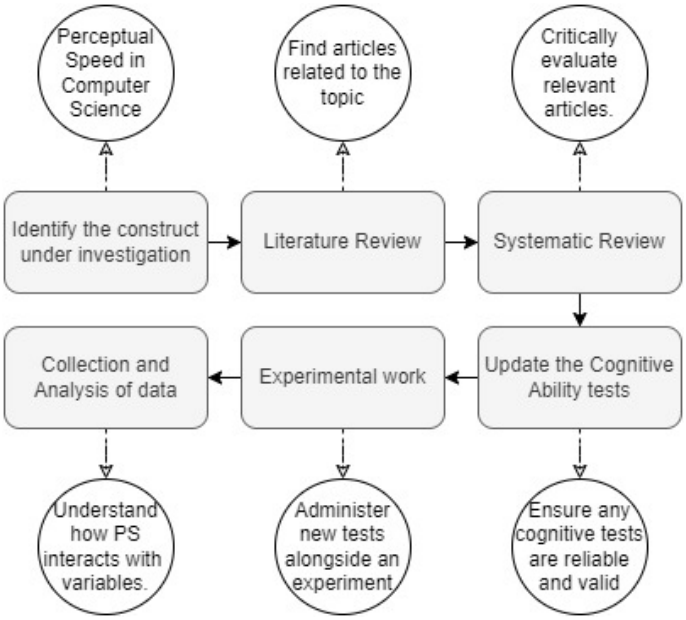


Figure 1.1: A flowchart of the overall thesis process.

having moved beyond the system, and onto expanding understanding of the complexities of how a user's individual cognitive abilities may impact their search experience.

Secondly, having identified problems with pre-existing measurement of PS, and subsequently creating new digital PS tests, these can provide a measurement tool that can be used in many fields beyond IIR, ranging from Psychology to Human-Computer Interaction. With measurement uptake in more fields, a comprehensive and greater understanding of user differences will be achieved, and thus different domains beyond IIR can be benefited.

Finally, whilst it has previously been unknown as to what part of the search system users with lower levels of PS struggled with, through experimentation of different interfaces, the present research has identified that users with Low-PS are able to achieve better searches than High-PS, given the right interface. Furthermore, it has also been established that users with High-PS perform less efficient searches in different interfaces. This therefore provides the first indication that system designers and advertisers must pay attention to the presence of, and type, of visual clutter visible to users of differing PS ability. It is then hoped that the findings of this research will ultimately lead to the design and development of more usable and enjoyable interactive systems that can adapt to meet the needs and preferences of a diverse range of users across a range of contexts; so that everyone can gain the information they desire, in a positive and efficient manner.

1.5 Thesis Outline

For transparency about how this thesis is organised, it was split into the following parts and corresponding chapters:

1.5.1 Part 1: Background.

Chapter 1: Introduction— This Chapter explains the research context, research questions, overall methodological approach taken, research contributions, and provides the thesis outline.

Chapter 2: Literature Review— This Chapter discusses previous work which provided the motivation for the present research, specifically focusing on the areas of Interactive Information Retrieval, Perceptual Speed (PS), and visual clutter.

Chapter 3: Methodological Approach— This Chapter provides a justification for why every research method in the present thesis was selected.

1.5.2 Part 2: Evaluation of Perceptual Speed measurement and the concept overall.

Chapter 4: Test Problems— This Chapter provides a more thorough evaluation of previous literature, where themes emerged for problems regarding PS measurement. Challenges and recommendations for improving the reliability and validity of PS testing were then discussed.

Chapter 5: Systematic Review— This Chapter presents the first Systematic Review into the area, specifically focusing on where significant results have occurred for PS, and how results could be explained.

Chapter 6: New Tests— This Chapter details the process undertaken to develop new digitised PS tests and proposes how participants should be categorised into Low-PS and High-PS ability.

1.5.3 Part 3: Investigating the effect of different interfaces on users with differing PS ability.

Chapter 7: Clutter— This Chapter presents the first results from the user experiment, specifically focusing on how the presence of clutter in the form of advertisements affects users with Low-PS and High-PS (as measured by two different PS tests) in terms of their search performance, behaviour, and experience.

Chapter 8: Congruence— This Chapter presents the remaining results from the user experiment, specifically focusing on how the *type* of visual clutter (whether the advertising is congruent with the task, incongruent, or a mixture of both) affects users with Low-PS and High-PS (as measured by two different PS tests). Explanations of the results are also suggested.

1.5.4 Part 4: Discussion

Chapter 9: Overall Discussion and Conclusion— The final Chapter summarises everything together to provide answers to the main research questions which conclude the thesis.

Additionally, an overall discussion is provided, which merges knowledge learned through multiple chapters together. Overall limitations are then acknowledged, and ideas for future research are proposed.

1.6 Publications

1.6.1 Publications Used in this Thesis

Work presented in this thesis has previously been published at the following peer-reviewed conferences, where for each paper, the corresponding chapter where the content of the paper is included has been stated:

1. Foulds, O., Azzopardi, L., and Halvey, M. (2020) 'Reflecting upon perceptual speed tests in information retrieval: limitations, challenges, and recommendations', In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pp.234-242

    - **Winner of The Association for Computing Machinery (ACM) Conference on Human Information Interaction and Retrieval 2020 Best Paper Award**.
    - The content of this paper comprises Chapter 4. Only the Introduction and Conclusion have been modified to align with the overall thesis story, and the Background removed, to reduce repetition from previous chapters.

2. Foulds, O., Azzopardi, L., and Halvey, M. (2021) 'Investigating the Influence of Ads on User Search Performance, Behaviour, and Experience during Information Seeking', In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pp.107-117

    - **Winner of The University of Strathclyde Euan Minto Prize 2021**.
    - The same methodology of this paper was used in Chapter 7 and Chapter 8.

1.6.2 Additional Publications

Throughout this research, multiple other publications have arisen— both in conferences and book chapters. However, these have not been included in the current thesis for different reasons, which are briefly explained after each reference below.

1. Foulds, O., Suglia, A., Azzopardi, L., and Halvey, M. (2020) 'Predicting perceptual speed from search behaviour', In *The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020).*

    - Although this corresponds with the same PS tests that were created and described in Chapter 6, this specific publication categorised PS into Low and High users using a median split of the sample. It was later established that this would not be the most

optimal way of measurement, and hence, for the purpose of this thesis, participants were grouped differently using both extreme-group analysis, in addition to the whole sample on a spectrum. This new way of categorising participants is outlined later in Chapter 6, Section 6.9.

2. Foulds, O., and Wood, D. (2020) 'The effects of Visual Clutter and Perceptual Speed in high-pressured information environments on the performance of tactical systems operators in the underwater battlespace', In *Underwater Defence Technology Conference (UDT 2020)*.

    - This provided an example of a specific domain that could be benefited from the current research. However, as the thesis focused on the field of IIR, this was not included.

3. Foulds, O. (2020) 'Too many dull words exceed the limits of visual perception: the effects of clutter and colour on learning', In *The 17th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2020)*.

    - This was an extension of work completed in the author's undergraduate degree. Whilst this helped enhance understanding on the topic of clutter, it did not include an exploration of PS.

4. Foulds, O. (2022) 'Investigating How Word Clutter and Colour Impact Upon Learning', *Orchestration of Learning Environments in the Digital World*, Springer, pp.135-151.

    - This was an extension of work completed in the author's undergraduate degree. This further explored the concept of clutter, but again, did not explore PS.

Chapter 2

Literature Review

2.1 Chapter 2 Overview

This Chapter discusses previous work which provided the motivation for the present research. More specifically, this chapter: Presents a brief overview of the field of Interactive Information Retrieval and the relationship with Individual Differences; Explains the importance of one specific individual difference, Perceptual Speed (PS), and it's effects on IIR; Discusses how more understanding is required to understand how to optimize IIR for users of different PS ability; Provides further understanding into visual perception by focusing on the concept of visual clutter; Explores how visual clutter has been operationalised during IIR— narrowing down into a specific focus on advertising— alongside studies which may implicate the interaction between clutter and PS; and summarises the research questions of this thesis in relation to previous work and gaps in literature.

2.2 Interactive Information Retrieval

The field of Information Retrieval (IR) concerns how when an information need arises, a user must issue a query, and then navigate some kind of system where information is stored and organised [24]. The system must then relay relevant information back to the user, in a timely and concise manner, to satisfy the user's information need [24, 78].

Whilst many studies have focused on the technical performance of retrieval systems, such as their *Precision*—the fraction of documents retrieved that are relevant to the user's information need, in relation to a single query [19]—other studies have attempted to model and optimize the entire process as an interactive phenomenon, considering the many different ways a user can potentially interact with a search system [240]. Thus in **Interactive** Information Retrieval (IIR), the focus has shifted from system-oriented to a user-oriented perspective, where user behaviour and experiences are also considered, including the physical, cognitive and affective features of interaction [120].

Now considering both the user and the system simultaneously, the field of IIR has expanded beyond Information Retrieval, to also include many other disciplines such as Psychology, Computer Science, Information and Library Science, and Human-Computer Interaction (HCI) [78]. However, previous researchers have outlined that an interdisciplinary approach can be a challenging area of research: "*How users think, behave, and make decisions when interacting with information retrieval (IR) systems is a fundamental research problem in the area of Interactive Information Retrieval (IIR)*" [150], and "*The inclusion of users into any study necessarily makes IIR, in part, a behavioral science. As a result, appropriate methods for studying interactive IR systems must unite research traditions in two sciences which can be challenging. [...] there is no strong evaluation or experimental framework for IIR evaluations as there is for IR studies. IIR researchers are able to make many choices about how to design and conduct their evaluations, but there is little guidance about how to do this*" [120].

Furthermore, the proliferation of the internet has meant that more and more search systems need to be evaluated. It is only since the 1990s that search engines like AltaVista, Yahoo!, and later Google provided users with powerful tools to search and navigate the rapidly evolving web. For example, when Google launched in 1998, there were around 10,000 queries issued daily [30]. Now in the year 2022, this statistic has exponentially increased to roughly 5.6 billion daily queries [189]. With many other search systems serving even more queries, this highlights that the field of IIR has further expanded. Yet, due to the multidisciplinary nature and expansion of IIR research, a challenge still exists regarding how to thoroughly conduct research studies in this area that fully understand how to evaluate systems that can truly support their users.

2.3 Interfaces in Interactive Information Retrieval

One way that researchers have attempted to evaluate how IIR systems can support their users, has focused on understanding the role of interface design [69]. Specifically, as described by Borlund: "*The overlapping research interest shared between HCI and IR is particularly in relation to the design of IR interfaces and the determination of the functionality and the level of cognitive load of already existing IR interfaces (e.g., Henninger, 1994; Brajnik, Mizzaro & Tasso, 1996; Beaulieu & Jones, 1998)*" [38].

As explained by Wilson, there are certain common features that all searchers expect to see on a Search User Interface (SUI):

- "*input features – which allow the user to express what they are looking for.*
- *control features – which help users to modify or restrict their input.*
- *informational features – which provide results or information about results.*
- *personalizable features – which relate specifically to searchers and their previous interactions.*" [234].

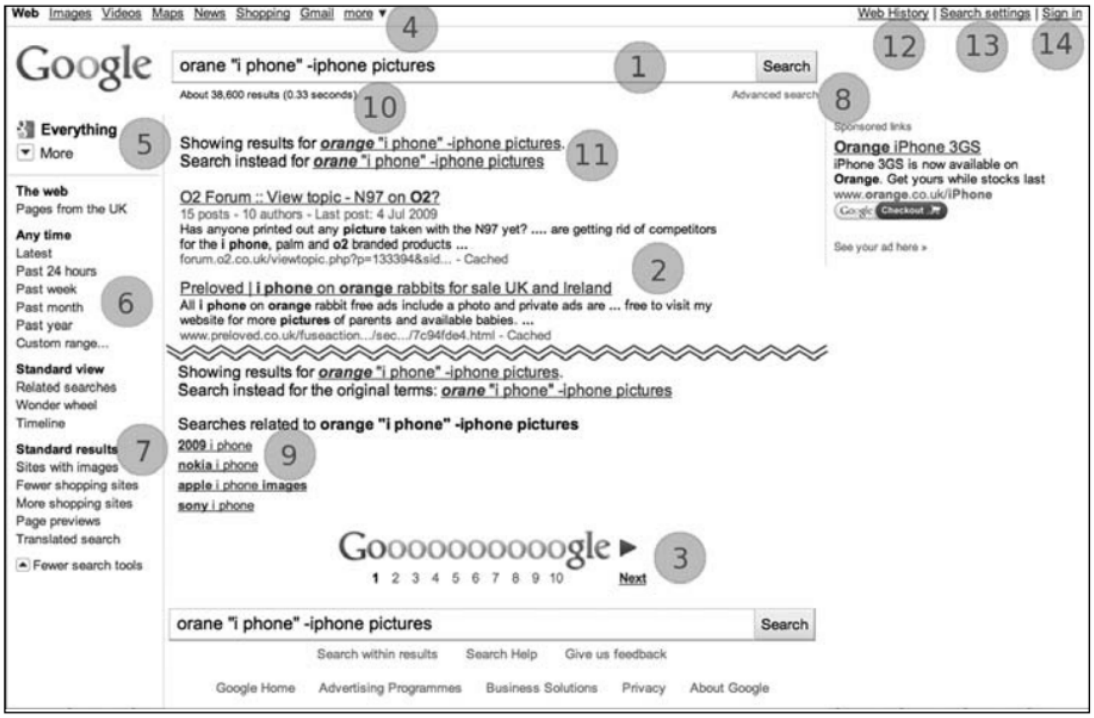


Figure 2.1: A standardised interface. Taken from [234].

Each one of the above features influences how a user perceives and interacts with the interface. For example, as depicted in Figure 2.1—which provides a screenshot of a common interface, Google—the most typical control feature refers to a query box. However, how these features are implemented in each SUI differ on a number of dimensions, ranging from the types of facets visible, to a visible indication of retrieved result relevance, such as TileBars or 2D clustering, to name a few (For a detailed review, please refer to [234]). These differences are ultimately because each interface has a unique aim.

In addition to each interface being designed with a unique aim, each user interacts with an interface differently, depending on their specific information needs and goals. For example, after submitting a query, users make decisions about which search results to click on. They scan and evaluate the titles, snippets, and URLs provided by the search engine to determine which results are likely to contain the desired information. The decision to click on a search result is often influenced by the relevance and quality of the snippet. Otherwise, if the initial search results do not meet their needs, users often make decisions about how to refine their search queries to get better results. This might involve adding or removing keywords, using quotation marks for exact phrases, specifying additional criteria, or navigating various filters and sorting options to refine the search results further; such as by date, language, location, or specific attributes, depending on the user's requirements.

However, whilst many different interfaces have been designed and generated in IIR, their effectiveness, from a user-oriented perspective, still remains unknown: "*Because the majority of*

human factors design guidelines are qualitative, the effectiveness of many design techniques is left to debate because there are no methods to provide numerical analysis or direct comparison between different design proposals" [31]. Whilst this paper was from 2006, the same point has also been identified in more recent years: "*The problem of the guidelines is that they are usually described qualitatively and it is difficult to convert them into a set of strict rules*" [185]. For example, in research which explored the relationship between visual interface aesthetics and task performance and preference, Salimun [200] noted that due to a large variation in preference and performance, it was difficult to predict interface preference precisely. Furthermore, given the costs and time involved in usability testing of every possible interface feature, usability testing is often unfeasible [136]. Consequently, many interfaces are created without true understanding of how they affect a user during IIR.

2.4 Individual Differences

Even if only one interface existed in the world, differences in IIR would still exist because users naturally differ from each other in many aspects, and these individual differences can impact their search behaviour, performance, and experience, when engaging in IIR. One theory that attempts to explain these differences refers to 'The Human Cognition Factor', where individuals have "*preferred ways of seeking, representing, processing and retrieving information, depending on their individual cognitive skills and abilities*" [193]. Thus, web designers have highlighted the need for interfaces to be designed in accordance with knowledge of the user [185]. Consequently, a core component of many online services nowadays involves personalised experiences, which exist in a variety of online services such as social media platforms, e-commerce websites, on-demand video services, and music providers [202]. For example, through identifying demographic attributes of a user such as gender—through their mouse cursor data [15]—targeted recommendations can be created which can improve the relevance of information being returned to satisfy a user's need.

In addition to gender, there are many other individual differences which can impact a user's search, including factors such as: the experience of the user (e.g. domain experts are more successful and employ different search strategies compared to novices [52]); personality traits (e.g. a user with a higher trait of *Need for Cognition* engaged in different search tactics and behaviours during IIR [236]); and cognitive abilities, such as working memory (e.g. users with lower working memory had lower task engagement whilst simultaneously reporting higher levels of workload [18]), to name a few.

2.4.1 Cognitive Abilities

Whilst many individual differences can impact one particular area of online searching, cognitive abilities can impact many areas of IIR. This is because cognitive abilities involve higher mental

functions which involve understanding, problem solving, reasoning, and remembering—and **all** of these factors can impact search behaviour, experience, and outcomes [46]. Overall, it has been observed that people with higher cognitive abilities tend to perform searches better than those with lower levels [46]. Thus, becoming aware of individual differences involving cognitive abilities in particular, can have several important implications for researchers and web designers.

Specifically, research into cognitive abilities since the 1990s was motivated based on the practical considerations of designing information systems. Specifically, *The Cognitive Engineering Approach to Design*, by Allen [9], mandated that individual differences between users should be understood, in order to allow more efficient systems to be designed that could be used by everyone. This was then further explained that the usability of information systems could be enhanced, if differing cognitive processes could be supplanted or augmented with system features.

In his next paper, Allen [11] extended this motivation, by splitting the process of completing a search task into different components: defining the search topic; selecting appropriate search vocabulary; issuing commands; selecting menu choices; viewing retrieved information; and making judgements about something's relevance or usefulness. Simultaneously, navigating these various search elements requires multiple cognitive processes, including visual scanning, attention control, learning, comprehension, and speed in spotting information [4, 18] Consequently, there are many aspects of searching that users with different levels of cognitive abilities could be affected by. By better understanding these components, specific design features, such as deciding the best way to display search results, could be combined with specific user characteristics, to ensure greater system usability.

Now, almost 30 years later, the motivation for researching cognitive abilities remains very similar. As explained by Arguello & Choi [18]: "*Understanding how cognitive abilities influence behaviors and outcomes has several important implications. First, it can help us design systems that are more accessible for users with a wide range of abilities*". However, this article also explained an additional reason for the usefulness of better understanding cognitive abilities during Human-Computer Interaction: "*Second, it can inform the design of personalized presentations, interactions, and assistance tools to support users based on their unique abilities*" [18]. This latter point is important, as designers became aware that if they altered a design feature to accommodate one group of users, it was equally important to make sure that another group of users were not being negatively affected by this: by helping one, such as a user with Low Working Memory, another, such as a user with High Working Memory, may become impeded. Additionally, not all users with lower levels of a particular ability might have problems with the same things [138]. Therefore, having a personalised search system that adapts to enhance the strengths of each user, would be the ideal scenario, and this motivation has been shared by many (*such as* [67, 139, 206, 210, 220].)

However, despite the positive implications of greater understanding cognitive abilities being apparent, there is still a lack of consideration of them amongst IIR research. For example, two years after Arguello & Choi's paper [18], the same authors who had discussed the benefits of cognitive ability research, released new technology to support complex search tasks [60], and yet, there was no consideration of how this tool could be used by users with different cognitive abilities. Consequently, there is still a lack of understanding for: identifying which tasks users with different levels of cognitive abilities struggled with; and designing the most optimal search systems that can support users with differing cognitive abilities.

2.5 Perceptual Speed

Several cognitive abilities have previously been examined in relation to IIR, including Perceptual Speed (e.g. [1,46,66]), Visualisation ability (e.g. [29,46]), Visual Working Memory (e.g. [66,210]), and Verbal Working Memory (e.g. [29,210]), to name a few. However, a review of literature on individual differences identified that Perceptual Speed had been the most investigated—at least, in the domain of how it affected a user's interaction with different visualisations—compared to all other abilities such as spatial ability and memory [151].

2.5.1 Defining Perceptual Speed

Perceptual speed (PS) was first defined in 1938 as one of seven cognitive abilities that intelligence comprised of (Thurstone, 1938, cited in [107]). It originated during World War II, when slow perception was cited as a cause of flight training failure amongst U.S Army Air Forces flight instructors (Guilford & Lacey, 1947, cited in [107]). Shortly thereafter, in 1942, two PS tests were developed and administered to pilot candidates: Dial Reading and Table Reading [75]. Although the Dial Reading Test was eventually dropped from the test battery, Table Reading continues to be administered in the US Air Force pilot selection battery to this day (as of 2022) [75]. This paper-based test involves a large grid of numbers, and the participant must find certain numbers amongst them [75].

In other domains—particularly research orientated, and not industry—different variations of PS tests have been developed. In a Kit of Cognitive Tests that was developed in 1963 as part of the Educational Testing Service, the concept of PS was more formally defined as "*Speed in finding figures, making comparisons, and carrying out other very simple tasks involving visual perception*" [92]. Multiple different PS tests were then created, all of which involved various sub-factors, including: (a) speed of symbol discrimination, as measured by the *Finding A's* test, where participants viewed lists of words and selected any that contained the letter 'a'; (b) speed of making comparisons, as measured by *Number Comparison* through inspecting pairs of multi-digit numbers and indicating whether the pairs were the same or different; and (c) speed of form discrimination, through recognising pretermined but novel configurations, as measured

by the *Identical Pictures* test, where users had to view a geomeotrical figure and then select the same pattern from a choice of 5 in the same row [92]. It was then later established that for a truly valid measure of PS to occur, at least two of these tests should be completed by a participant [83].

Ekstrom's [83] PS Tests are still commonly administered in research today, and their use informs clinical decisions (e.g. identifying associations between cognitive function in fasted and non-fasted states amongst participants with Alzheimer's [37], and dynamic awareness of age-related change in cognitive performance [246]). Beyond psychological fields, in the domain area of this thesis—IIR and Computer Science more generally—Ekstrom's tests of PS also appear to have been the most commonly used in previous research (E.g. [43,71,204,220]). Yet regardless of the individual PS test chosen, the general definition described PS as a cognitive ability, which involves an individual's accuracy and speed to view, scan, and compare information during visual search tasks [10].

2.5.1.1 Other Types of Perceptual Speed

It is important to note that the concept of Perceptual Speed as a cognitive ability differs from the notion of "*perceptual speed regulation*", which instead involves an individual's ability to *regulate* speed, such as in a driving simulator [155]. Similarly, the cognitive ability of Perceptual Speed differs from "*perceptual speed control*", which concerns a driver's ability to *estimate* their physical speed of movement [231].

However in previous literature, there has been some confusion between the concepts of 'Perceptual Speed' and '*Processing Speed*'. For example, some research studies have interchanged the name of these concepts, such as Zimprich & Kurtz [248], who administered the *Number Comparison* Test by Ekstrom, but recorded this as a measure of *Processing Speed*. Whilst some studies may have believed *Processing Speed* and Perceptual Speed were synonyms for the same fundamental concept, having evaluated previous literature, the present thesis has clarified that there are differences between these concepts.

Firstly, Crabb and Hanson [71] explained that whilst *Processing Speed* refers to an individual's "*mental quickness, which requires very little complex thinking*", they also distinguish that Perceptual Speed is a single narrow subset of *Processing Speed*, defined by "*a measure of an individual's ability to search and compare visual symbols or patterns in rapid succession*".

Crabb and Hanson's [71] definition of these concepts was further supported by Christidi et al [61], who described them as different cognitive processes that contribute to overall performance: *Processing speed* relates to more visuomotor processing, and tests that encapsulate tracking of a sequence, motor execution, and divided attention are used. In contrast, to measure just perceptual ability, different tests have been used—such as the WAIS Block Design, which involves matching patterns and speed, similar to Ekstrom's tests—which reaffirms that different constructs are being measured.

Thirdly, Crabb and Hanson [71] additionally demonstrated that Perceptual Speed does not reduce with age, and yet, there is a concretely tested "*Processing Speed theory of cognitive aging*" which states that *Processing Speed* **does** decline with age (Birren & Schaie, 1990; and Salthouse, 1996, cited in [85]).

Finally, the first few hundred literature results that were returned for the term "*processing speed*" in the most common computing database, The Association for Computing Machinery (ACM) Digital Library, were inspected. This confirmed that *processing speed* had a completely different meaning in the field of computing, and instead of referring to a cognitive ability, it tended to refer to the processing speed of a computer - in other words, the number of instructions per second the computer executed (E.g. [41]).

Combining all of the above points together, these differences reaffirmed that Perceptual Speed was different from *Processing Speed*, and in the context of the present thesis, Perceptual Speed was the more appropriate concept to further explore.

2.5.2 Perceptual Speed in Information Retrieval

2.5.2.1 Tasks Explored

In the context of IIR and Computer Science research more generally, the concept of PS has been explored in relation to how it affects a user's search effectiveness in various tasks, including: identifying relevant pages that help the user learn more about a topic [4, 9, 11, 12, 13, 18, 125]; finding a specific answer to a question through web-browsing [10, 46, 71, 103, 174, 204, 205]; analysing various pieces of information to provide a response [18, 43, 46]; being given a range of scenarios and then choosing a preference [64, 66, 67, 139, 140]; or generating something new as a result of searching (such as the creation of a living room design) [46, 225]. Furthermore, a few other very specific Human-Computer Interaction tasks have been investigated, including whether PS can predict performance when using different types of devices in older people when browsing and playing online games [154]; and a simulated military operation search task [84].

2.5.2.2 Search Environment

In addition to different tasks being explored, various search environments have also been utilised, ranging from hierarchical database structures in the 1990s [204], to latterly experimenting with more sophisticated customizable websites with interactive data visualisations in 2020 [67]. In terms of visualisations, many different ones have also been experimented with—to identify their affect on users with different PS levels—including comparisons between single and multi-window formats [12, 13], bar charts versus radar graphs [209, 210, 211, 221, 222], different visual prompts such as highlighting or bolding [56, 94, 218], comparing vertical versus horizontal layouts [64, 140], interactive maps [66, 67, 139], and magazine style effects [219, 220].

2.5.2.3 Search Effectiveness

Regardless of the search task or environment investigated, the way that search effectiveness has been operationalised has changed over time. Originally, search effectiveness was measured in terms of search performance, which was mainly concerned about some form of task accuracy, such as how many relevant documents were found [11], or how much information a user had learned [12]. It has only been in the last decade that research studies have considered the subjective feelings of the user when analysing search effectiveness. For example, in 2011, Al-Maskari & Sanderson [4] examined whether PS impacted how satisfied users were with their search, and this trend continued the following year in 2012, when Fincannon et al. [84] evaluated whether PS impacted how users rated their perception of workload during their search tasks. This corresponded with the developing theory of *The Human Cognition Factor*, which suggested that individuals have preferred ways of seeking, representing, processing, and retrieving information [193]. Consequently, more and more papers in recent years have prioritised the user's feelings during IIR tasks, and this has been reflected in more papers analysing user *experience*, as well as *performance*.

2.5.2.4 Theoretical Framework

Although many different tasks and environments have been explored in relation to PS, these explorations have mainly been atheoretical. For example, one study stated that there was *no* theoretical framework behind their investigation into PS, eye-gaze, and visualisation processing: "*To the best of our knowledge, there is no established comprehensive theory connecting eye gaze patterns and individual user traits that could guide our investigation of gaze patterns during visualization processing*" [222].

For the few papers identified that did mention some kind of theory, these did not help form directional hypotheses for their research. For example, Allen [11] stated that "*Based on learning theory, perceptual speed influences learning, which in turn influences performance*". Yet, no other explanations were given, such as **why** or **how** PS would influence learning. Similarly, Kim & Allen [125] wrote that "*The theoretical foundation of this research is the 'user-oriented IR research approach' identified by Ingwersen (1992). This approach suggests that information system design and evaluation should be based on a firm understanding of how users interact with the information systems... Within the user-oriented approach, our specific theoretical focus was interactionism. We believe that search behaviors (like many other behaviors) are influenced by the interaction between individual characteristics and social contexts*". Yet again, it was not specified, or even indicated, how search behaviours, or which behaviours, may differ between users with different individual characteristics.

For another two papers that also mentioned some form of theory, these were also not related to forming directional hypotheses as to how PS would affect IIR. Instead, the theories related

to explaining some kind of *background* information. For example, Seagull & Walker [204], explained Ackerman's 1988 theory of Skill Acquisition—which theorised why performance on a task increases with practice—rather than considering a theory that helps understand their research aim of understanding how PS would impact search time using different information presentations. Additionally, a later paper by Allen [12] referred to the theory of Cognitive Facilities, and described the following: "*The idea of cognitive facilities developed by Jackendoff provided a foundation in theory for this research, and allows an understanding of the importance of spatial representation of information in digital libraries. The main point of this body of research in linguistics and cognitive science is that there appear to be two separate cognitive facilities: one that deals with space and the objects that are encountered in space, and one that deals with language and other symbols*" [12]. This description was further expanded to discuss different neural regions responsible for cognitive facilities. Although providing an insightful background, the use of this theory also did not directly motivate their study aim of identifying whether PS interacted with different system features designed to enhance learning in IIR [12]. Thus, how learning may differ for users with varying levels of PS in IIR is still unknown.

2.5.2.5 Main Findings

Although previous research has been more exploratory, instead of being based upon clear theories which formed directional hypotheses, many results have been reported for how PS affects a user's search performance, experience, and behaviour during IIR. Amongst all findings, the general consensus is that people with lower levels of PS struggle with their search, in comparison to people with higher levels of PS.

For example, in terms of **search performance**, people with Low Perceptual Speed (Low-PS) have: completed searches with lower precision ratios [9, 56]; learned less during their searches, whilst being overall slower [10]; retrieved fewer relevant documents [11]; and obtained lower recall post-task [13].

In addition to more negative search outcomes being found for users with Low-PS, this also extended onto their **user experience**. Here, people with Low-PS have: reported a lower overall browsing experience [71]; obtained higher levels of perceived disorientation [71]; experienced more self-reported workload [46]; reported interfaces as less usable [225]; and were less satisfied with their search performance, as measured by their opinion of their own search skills [225].

However, there have been some inconsistent results in previous literature. For example, in four comparative IIR tasks—where a user had to retrieve relevant documents, compare various pieces of information, and justify their decision to a particular question—Arguello & Choi [18] identified that users with Low-PS and High Perceptual Speed (High-PS) did *not* differ in their perceptions of workload or engagement. Nonetheless, Arguello & Choi offered possible explanations for these differences: it was speculated that previous studies (E.g. [225]) had administered more difficult tasks, and therefore "*it is possible that our tasks were difficult enough*

for Low-PS participants to exhibit different behaviors but not difficult enough to impact their post-task perceptions" [18]. Therefore, although contrasting results have emerged in previous literature, explanations have implied that Low-PS users would still struggle completing an IIR task, in certain conditions.

With the majority of previous research reporting different experiences and outcomes for users with different PS levels, it is unsurprising that there have also been differences observed in their **search behaviour**. These include Low-PS users: spending significantly more time completing the task [4, 11, 46, 174]; interacting with less features (such as clicking on fewer results) [46]; and engaging with more query abandonment (such as issuing a query that did not result in task completion) [18].

Referring back to the original definition of PS—which referred to an individual's accuracy and speed comparing visual information [10]—it makes sense that a user with Low-PS would overall take longer completing an IIR search task, whilst simultaneously achieving a lower accuracy. To attempt to explain some of these differences, some research studies have incorporated eye-tracking into their experiments. This revealed fundamental **physiological differences** between users, where individuals with High-PS have a higher eye fixation rate, and are thus able to scan what is in front of them more quickly compared to people with Low-PS [209, 211, 222]. Consequently, if users with Low-PS are not physically able to scan as much as their High-PS counterparts, their searching experiences will inevitably differ, both subjectively, and objectively.

2.6 Developing Adaptive Systems

Given such a range of results concerning how users with Low-PS can be negatively affected during IIR, a common aim amongst recent studies referred to the need to develop adaptive systems that can accommodate the unique cognitive abilities of each user. For example, Conati et al. [67] was motivated to develop adaptive systems based on previous research that identified lower levels of PS being linked to lower task performance, and Steichen et al. [209] were motivated since Low-PS led to lower performance, in terms of both speed and accuracy. Both of these studies believed such users would benefit from adaptive interventions, and as such, began to implement user modelling, in an attempt to predict a user's PS. The findings from eye-tracking studies which investigated whether PS affected eye-gaze during various tasks (e.g. [222]) were thus extended to investigate whether PS could be predicted solely from eye gaze (e.g. [209,210]).

Although studies have attempted to predict a user's PS, with the long-term aim of creating adaptive systems that can help these users achieve more positive searches, there is a lack of research which has identified *how* users with Low-PS could be helped. Referring back to the main findings in PS research, if a user with Low-PS completed a search with a lower precision ratio, or they retrieved fewer relevant documents during their search, there is no indication as to

whether it is possible for them to achieve a higher precision ratio, or more relevant documents, using a different interface. Similarly, when negative user experiences have been reported for users with Low-PS, it is unknown whether positive user experiences are even possible.

In the study presented by Turpin et al. [225], the participants with lower levels of PS who significantly rated the interface as less usable attributed factors such as "*distraction*" and "*confusing*" as the reason for worse user engagement. This is unsurprising, given the knowledge of differences in eye-fixation between users. However, how to make an interface less confusing, or less distracting, is still unknown. Instead, the authors concluded that "*Future research might focus on developing interfaces that improve the search experience for people with Low-PS*" [225]. This conclusion was shared by many other studies, where instead of providing a suggestion for how to enable Low-PS users to be more positively affected, only the need for personalised interventions in the future was stated. For example, Naghib et al. [174] concluded: "*It is necessary for information retrieval system designers to design personalized retrieval systems which take into account the users' cognitive features*"; and Lalle & Conati [137] surmised that "*users with low levels of these abilities may benefit from personalized support*".

Whilst many studies only recommended the need for personalised systems, without explaining what kind would be useful, others have provided various suggestions. For example, in a study which involved multiple tasks of varying complexity, Brennan et al. [46] explained that: "*Finding from our work warrants further investigation, possibly toward the goal of developing user-adaptive systems whose features capitalize on the cognitive strengths of users without penalizing their weaknesses. For instance, people with low perceptual speed might benefit from additional tools to help them navigate documents, and keep track of, and integrate, their findings*". Yet, it was not specified what type of tools could be useful, and no subsequent research has examined this. Similarly, in a different search task, Allen [9] also proposed some suggestions for what may help Low-PS users perform better searches: "*It is suggested that information retrieval systems can be made more accessible to users with different levels of cognitive abilities through improvements that will assist users to scan lists of terms, choose appropriate vocabulary for searching, and select useful references*". Yet again, these suggestions had no detailed explanations—such as, what kind of improvements were necessary to help a user have more efficient scanning. ——and they have also not been followed up in later research.

In other research however, different manipulations of the search system have identified significant interactions with PS, which might begin to explain how users with Low-PS could be helped achieve better searches. Specifically, Conati & Maclaren [68] evaluated the effectiveness of two different data visualisation techniques for describing complex environmental changes in an interactive system, and a significant interaction between PS and visualisation type occurred: users with Low-PS had worse task accuracy when using the colored boxes visualisation, and High-PS users had worse task accuracy during the radar graph visualisation. This might imply

that a user with Low-PS could be helped in certain conditions, albeit in this particular context. However, the authors were unable to explain these findings: "*We don't have a conclusive explanation for the direction of the relationships that we found among perceptual speed, accuracy with the radar graph and accuracy with the colored boxes*" [68]. It thus still remains unknown as to which parts of the different search systems allowed users with Low-PS to achieve higher accuracy.

Bringing everything together, it is clear that PS is an important cognitive ability that affects users completing IIR tasks, and thus long-term, it would be ideal if users who struggle with this ability could be helped so that they can achieve searches with better accuracy and user experience. If they could be helped, then dynamic and adaptive search systems could be developed to improve accessibility for all—a desire of many researchers and web designers. However, there is still a lack of understanding for how to help a user with Low-PS succeed in their search, or whether it is even possible for them to achieve search outcomes that assimilate to their High-PS counterparts across different tasks and environments.

2.7 Visual Perception

In order to begin understanding how to help a user with Low-PS succeed in their search, it is important to consider how human visual perception works in general. Perception is a combination of: (a) unique aspects of the individual user (including factors such as prior knowledge, task-relevance, and individual abilities—such as PS); and (b) the physical features present (such as the saliency, colour, and how cluttered the information presented is [42, 149]). However, because human brains are limited and prone to cognitive overload, an individual cannot physically process everything they encounter [114, 117]. This phenomenon can be described by theories on visual clutter.

2.7.1 Defining Clutter

Formally, visual clutter—otherwise known as crowding—has generally been defined as the negative impact of nearby contours that interfere with and reduce visual discrimination when trying to focus on a target [144]. Clutter therefore occurs in a multitude of areas, from everyday errands such as finding something in the refrigerator, to essential visual search tasks like X-ray baggage scanning for dangerous items [2]. Many theories have attempted to explain why clutter causes perceptual problems. Studies have revealed that when people view cluttered compared to uncluttered scenes, the excess of visual stimuli exceeds the limits of attentional resources and short-term memory, which results in a bottleneck that impairs object perception [144]. Even if a person is not directly gazing at different distracting objects, they usually continue to perceive them peripherally, and this can still negatively overload their cognition as they subconsciously process the unnecessary detected information [224].

2.7.2 Effects of Clutter

Defining clutter simply as the number of items around a target, many experiments have discovered a similar theme– increasing clutter impairs efficiency of individuals performing search tasks by increasing response times and the number of errors made [2, 167, 238]. For example, when participants were presented with target words to remember, if these had been initially presented in isolation—in comparison to being viewed amongst distracting words—then participants had higher recognition accuracy, all while being faster at making their decision [87]. However, whilst a plethora of literature involving clutter exists, these studies stem from the areas of Psychology, Cognitive Science, and Engineering. Little research has explicitly investigated clutter during IIR.

2.7.3 Negatives of Increasing Visible Webpage Elements

In the context of IIR, many different factors could hypothetically create an illusion of clutter for a user to process. Firstly, in multiple experiments that manipulated how many result snippets per page were presented, when fewer results were visible, users had more positive perceptions for the search task [180] and revealed the least self-reported workload and difficulty in finding relevant documents [122]. Secondly, in different research investigating aggregated search – where different types of media such as images or news are blended into the same search engine result page (SERP) – users rated the blended interface as significantly less usable [225]. Similarly, when web browser reader views have been implemented –which strips back webpage complexity through excluding items such as menus and images – users perceived the visual appeal of the page to significantly increase and were able to read pages 5% faster [146].

Although these three examples cover a range of visible elements during online search, they all concluded similar results that increased elements or visual complexity on a webpage was associated with more negative perceptions. Yet, different explanations were given for each study. For example, Oulasvirta et al. [180] explained their results using the concept of 'choice overload', where a user cannot exhaustively attend to everything they see, and this then degrades user satisfaction because the user worries they may have attended to the wrong thing. Other explanations came from economic models of search, where users have to physically expend more cognitive effort to search through additional elements. A higher cognitive effort then induces negative emotions (Brajnik and Gabrielli, 2010, cited in [115]) and results in people missing or skipping over various important items, which affects their accuracy of finding what they're looking for [23, 57].

Whilst these studies provide clues for what would happen to users amidst various types of clutter present during IIR, clutter was not formally defined, and explanations of results found did not explicitly draw on theories of clutter. Furthermore, none of these studies explored how users with different levels of PS would interact with the various elements visible. Using the study

by Turpin et al. [225] as an example, it could be speculated that a user with Low-PS may be more negatively affected by different types of media being blended into the SERP, considering they are users who struggle to scan what is present, and a blended interface would require more scanning. However, without further research that combines clutter and PS together, this remains unknown.

2.7.4 Positives of Increasing Visible Webpage Elements

Despite one hypothesis being that users with Low-PS would be more *negatively* affected by increased clutter visible during IIR, users with Low-PS may instead be *positively* affected by certain types of visual clutter. For example, prior studies in aggregated search found that even although users had no reason to look for images, adding some images increased user interaction with the system, which resulted in greater accuracy at completing their search task of finding specific information embedded in a webpage [16]. Additionally, explorations in media communication identified that images stimulate engagement and interest in news stories [91]. This makes sense, as although it has already been identified that when a webpage is too visually complex users find it cognitively taxing to process [224], research since the 1970s has been aware that on the other end of the spectrum, if there are too few elements to process, then users may feel bored [Berlyne, 1970, cited in [186]]. Boredom may then lead users to distraction by other things around them, or cause them to abort their search all together.

Therefore, an inverted U-Shape relationship has been proposed for visual stimuli [186] where users need exposure to a moderate amount of stimuli, as both too much, and too little, can be negative for different reasons: too much to process increases users' cognitive load, inducing frustration and negative emotions, and this may result in a shorter duration of search [115]; but equally, distractions can be beneficial by replenishing mental resources, resulting in lower reported workload and stress [157]. Yet again, none of these studies incorporated individual differences of PS into their research, and thus, it remains unexplored whether certain types of clutter, such as images, could offer potential remedies for users with Low-PS to complete a more positive search experience. For example, perhaps by adding more clutter in the form of images, a user with Low-PS's mental resources could be replenished, and thus they might scan more efficiently.

2.7.5 Other factors affecting Clutter

Although various hypotheses have emerged for how increased—or different types of—elements could affect users with different levels of PS, clutter was not formally defined in any of the cited studies. Despite clutter being defined as the number of items around a target [2, 167, 238]—and thus the previously cited studies in IIR could be assumed to be measuring various types of clutter (e.g. [122, 180])—other research has identified that clutter cannot only be defined as the *amount* of visual information to be processed. In an article which provided a review of

how to measure clutter, it was concluded that "*we lack a clear understanding of what clutter is; what features, attributes, and factors are relevant; why it presents a problem; and how to identify it..[...]..by having an understanding of how clutter plays a role in visual search, we take a significant step toward understanding the role of clutter in many real-world visual tasks*" [198].

Different research has thus identified many other factors that contribute to the perception of clutter, including:

- organisation (e.g. how the clutter is grouped [124, 169]);
- whether clutter is localised immediately around a target, or global and refers to the whole scene (e.g. [33, 148]);
- similarity of the clutter against the target information [59];
- data density [169];
- object occlusion [45];
- colours present [124, 148];
- saliency of information [111, 116];
- and size of visible input [168].

Yet, many of these factors which involve different perceptions of clutter traditionally only relate to the physical features of visual stimuli, without capturing the effects of subjective perceptions of clutter. In a later review of measuring clutter, Moacdieh & Sarter [168] concluded that: "*It is thus arguably the performance and attentional effects that are of primary importance when it comes to assessing clutter..[..].. These performance and attentional costs can be determined only by considering the interaction between display-based factors on one hand and user-based factors, such as task difficulty, workload, and experience, on the other*". It has therefore been acknowledged by multiple studies that fully understanding clutter can only be determined by considering the interaction between physical display-based features *and* user-based factors, such as perceived relevance, workload, prior knowledge, and individual characteristics of the user [6, 111, 116, 184]. Consequently, further investigation of how different forms of clutter interact with a user's cognitive ability of PS, remains an area to explore which could shed light on how to design systems that can accommodate each unique user.

2.7.6 Clutter and Perceptual Speed

The only study that appeared to have combined PS and visual clutter together, did not refer to theories on clutter or crowding explicitly, but instead sought to understand the relationship between PS and task complexity, which was modelled as the "*amount of information that has to be processed at a time*" (Ziefle et al., 2015, cited in [44]). This definition corresponds with

the definition of clutter as defined above, where physical-based features of visual perception involving additional elements, alongside user-based features of PS are combined. Ziefle et al.'s study results found that users with Low-PS had significantly lower task performance, and that users with High-PS could compensate the negative influence of growing complexity better than people with Low-PS—within a task which involved various manipulations of how data could be presented, such as the number of lines per table [247]. Although these results were found within a very specific search context—and therefore the ability to generalise results onto web search more generally is unknown—it does offer one possible avenue to explore for helping users with Low-PS to complete more positive searches, through reducing visible clutter present.

2.8 Advertising Clutter

One area of web search which has explicitly investigated a form of visual clutter refers to advertising. Advertisers and marketers have long been aware of the detrimental impact advertising clutter can have. For example, Lee and Cho [237] showed that when webpages contained 4 extra adverts (ads), user's advert (ad) recognition significantly degraded and user's attitude towards the ads reduced. Similarly, when users perceived clutter to increase with increasing ad salience, a deterioration in ad memory was also found [128]. Consequently, explanations for why ad clutter negatively impacted users related to theories on visual crowding more generally, where a bottleneck in object perception occurs: "*The negative effect of ad clutter can be explained by limited capacity theory..[..].. basically that ad clutter threatens ad effectiveness because an excessive number of advertising weighs heavily upon consumers' memory capacity*" [126].

However, advertising clutter continues to grow. With worldwide digital ad spending in the year 2020 having reached over $332 billion dollars [72], ads now appear on almost all web-pages. A recent survey of 1000 internet users indicated that ads were now dominating the internet by being squeezed into more locations, whilst simultaneously appearing more and more intrusive [34]. In general, the average internet user is exposed to roughly 200 ads daily [3]. As such, online ads have grown increasingly sophisticated and now cover a wide variety of formats including banners, click-bait, personalised ads, and newer designs such as advergames [152,245]. These developments have all been driven from the perspective of the system or marketer, where the aim has been to ameliorate the negative effects of advertising clutter by displaying ads that are more engaging, memorable, and clickable [172]. Yet, all of these developments do not consider the effect of advertising clutter for how they impact a user's search behaviour, performance, and experience. If ad clutter could reduce a user's memory of the ad, then the presence of ad clutter may also overload a user's cognition in general, resulting in poorer search performance and experience.

2.8.1 Banner Blindness and Ad Avoidance

Whilst it is unknown how ad clutter affects a user's search behaviour and performance during IIR tasks, subjectively, many users report ads to be annoying [96], and options exist to remove their presence via ad blockers, browser reader modes, and/or paid subscriptions [181]. Alternatively, although some users continue to experience webpages with ads included, they report mentally skipping or filtering out the ads themselves – often resulting in a phenomenon dubbed *banner blindness* [51] or *ad avoidance* [147] – where ads are seemingly ignored, at least, from the advertiser's point of view.

In a study that examined user's attitudes towards ads, banner blindness was highlighted by many users, with one explaining: "*I'm so used to seeing banner ads I tend to just ignore them*" [230]. Previous research suggests that this concept is so common, that users rarely looked at ads and subsequently had very low recall of the ads that had been visible [51,97,128]. Factors that may affect banner blindness have been proposed such as the ad location [97, 110, 197], intrusiveness [27], or relevance to the task [36], to name a few. However the general consensus for why banner blindness exists appears to relate to how users deliberately avoid the ad to focus on the task at hand, especially when ads are task irrelevant [51, 128]. This is because during visual search, a user cannot attend to every visible element present, and therefore attention becomes selective to avoid information overload [128, 143, 195].

However, with awareness that individuals have different abilities of scanning and viewing information, it is possible that users with Low-PS may struggle identifying which visible elements they should attend to. It has previously been argued that when a user faces cognitive overload on the internet, they reduce the cognitive strain by dumping parts of information, rather than trying to find a way to efficiently process it [237]. Consequently, if a user with Low-PS needed to reduce their cognitive strain more than other users, they may begin to dump relevant information that could help them to achieve their search goal. Alternatively, they may not be able to dump any information at all, and instead process the ads visible. Consequently, the presence of advertising clutter may negatively affect a user with Low-PS more detrimentally, in comparison to a user with High-PS.

2.8.2 Ad perception without awareness

Although users claim to ignore ads—and it is unknown whether users with Low-PS can, or cannot, do this—findings from other studies would suggest that ads are still observed. Firstly, Tangmanee [216] showed that less than 10% of their users were able to correctly recall what ads had been viewed, despite nearly all fixating on at least one. Accordingly, users may *think* they ignore ads because they cannot remember them, but the ads *have* still been looked at. Secondly, Jahanian et al. [113] found that users who did not fixate on ads, but instead had to focus on the central point of a webpage for just 120 milliseconds, were successfully able

to discriminate pages that contained ads from those that did not. This implied that ads are still attended to, even without conscious awareness. Additionally, eye-tracking studies have shown that users often continue to fixate on ads [216], and even without direct gaze ads would still lie in a user's peripheral vision, which may result in them being implicitly processed [224]. Consequently, how the visual system processes peripheral information may mean that regardless of fixation, ads may covertly affect a user. This further emphasizes theories on visual clutter, where information that occurs in peripheral vision can still negatively overload a user's cognition [224, 233]. This then reaffirms the importance of understanding how the presence of ads actually affect a user during IIR: whilst users may believe they ignore them, it is unknown how blind banner blindness truly is, and what the consequences of increased clutter are. Furthermore, it has been suggested for over 10 years that individual differences in perception *will* influence how ads are perceived [102]. Thus, whether differences in ad clutter perception occur for users with different levels of PS remains an open research question to explore.

2.8.3 The Congruence of Advertising Clutter

As was described in Section 2.7.5, clutter cannot *only* be quantified by the number of visible elements present. As one variation, in addition to quantity of ads visible, the congruence of ads has been investigated into how this affects clutter perception. For example: "*we investigated whether or not the relevance of the ads moderates the negative effect of ad clutter on consumers' attitudes toward ads*" [126]. Here, Kim & Sundar [126] manipulated clutter as: a) *ad quantity* (14 ads were visible versus 2 ads visible); and b) *ad relevance*, where a relevant ad contained information that was appropriate to the keyword entered in the search system (e.g. for the query "*tanning*", relevant ads would involve tanning products, whereas irrelevant ads would involve "*mortgages*" or "*house loans*"). Results found that: 1) participants subjectively reported a significantly higher perception of clutter when more ads were visible (which reaffirms that quantity is a good indication of clutter); 2) relevant ads were evaluated as significantly better than irrelevant ads; 3) in addition to relevant ads being perceived as better, the website was also rated significantly more positively when relevant ads were present, in comparison to irrelevant ads; and 4) participants perceived less clutter when the clutter was relevant. The authors thus concluded that perceived ad clutter can be reduced by the relevance between ads and website context, without reducing the number of ads [126].

Different research has also identified that relevant ads—otherwise called "congruent", "fitting", or "topical"—result in: improved user satisfaction [191], increased user attention [51, 52], greater eye fixations [110], and elevated acceptance towards the ad [230]. When understanding why congruent ads have such an affect on a user's perception, it has been proposed that

congruent ads ease information processing as users see more sense, or logic, in congruent associations [36]. In contrast, incongruent ads cause confusion to users which leads to frustration and irritation, because incongruent ads interrupt cognitive processing of target information [27,36,128]. However, all previous research involving the congruence of ads has been driven from the purpose of understanding attention or attitude towards the ad, including Kim & Sundar's [126] study on relevant ad clutter. No research, to the best of our awareness, has explored how the congruence of an ad may affect a user completing their search goal, in terms of general IIR behaviour and task performance, or how theories of visual clutter may be affected.

Furthermore, given the various factors discussed previously—regarding how a user with different levels of PS may be differently affected by the presence of advertising clutter—it remains unknown how the relevance of an ad could impact user's perception of clutter, amidst differing perceptual abilities. For example, if a user with Low-PS is unable to "dump" additional information, and subsequently gets distracted by visible ads, it is possible that relevant ads may provide valuable information and lead searchers directly to their goal, by hinting at related images that inspire users to search for other related queries. Alternatively, users with High-PS, who are supposedly faster and more accurate at scanning information, may be more advantaged if relevant ads are present. Consequently, interactions may occur between PS and the relevance of visible clutter, in terms of the effect on a user during IIR.

2.8.4 Congruence in general web search

Given that the congruence of an ad has not been investigated in previous literature with regards to how it affects a user's IIR, other research on congruence in general provides clues for what might be expected to happen. In the context of aggregated search—where additional elements have been blended into the interface, such as videos, images, and news results, and their effect on a participant's search has been examined—Arguello & Capra [16] illustrated that even although participants had no reason to look for images, when congruent images were present, participants had increased interaction with the system, which resulted in the user being more likely to complete the search correctly compared to when incongruent images were present. Arguello & Capra [17] then later referred to this increased interaction when more congruent images were present as the "spillover effect", where the "*results from one source presented on a SERP can affect user engagement with the results from a different, completely independent source*".

Similarly, work using entity cards—which are a different form of aggregated search that combines various sources into a visual display box—has shown that when topical cards were visible, users also interacted more with the SERP, in addition to perceiving the system as being a higher quality [40]. As other research has identified that increased interaction with search results makes users significantly more likely to make a correct decision in a search task [188], it would be hypothesised that congruent ads may also increase interaction with the search system,

leading to an improvement in search performance during an IIR task. Given that congruent ads and clutter have not been explicitly examined in relation to IIR, it is therefore also possible that when the clutter is relevant, any negative effects of clutter may be ameliorated.

Furthermore, images have the power to convey meaning instantly whilst overcoming language barriers [129]. This means that even if a user struggles with written text that is presented to them, for example in users with dyslexia, then viewing a topical image verifies that the query issued matched the initial concept a user had in mind [208], which reassured them that they were searching successfully [91, 171, 228]. As users with Low-PS struggle with processing information [10] they may be similar to users with dyslexia who also struggle with visual processing [48]. Therefore, if the ads visible were congruent to their task—regardless of whether they create clutter or not—this may improve the search experience, and subsequently performance, for a user with Low-PS. In contrast, if incongruent ads were present, these might provide false information to the Low-PS user that they are performing incorrectly—in addition to requiring a greater cognitive effort to process, because a user must expend energy in working out why they have been shown something off-topic [27, 36, 128]—leading to a degradation in search performance. Consequently, adding images in the form of congruent ads to web-pages may result in positive outcomes for a user, including users with Low-PS. Alternatively, Low-PS users may be so negatively affected by clutter that ads, regardless of congruence, impede a user's search performance and experience. Thus, research is needed to examine the relationship between PS and the congruence of visible clutter, on a user completing an IIR task.

2.9 Motivation for the Current Research Questions

In summary, every day many people search for information online to retrieve relevant information depending on their task. Over time, research has shifted away from a system-oriented to a user-oriented perspective, where the focus of effective interface design now prioritises knowledge of the user. Although users differ on many aspects, cognitive abilities in particular affect all aspects of search behaviour, performance, and experience.

One specific cognitive ability referred to Perceptual Speed (PS), which involves an individual's accuracy and speed to view, scan, and compare information during visual search tasks. This ability has been shown to be an important factor in IIR: users with High-PS tend to have more positive search outcomes, across a range of search tasks and environments. For example, Low-PS users have learned less during their searches, achieved lower search precision ratios, taken longer, experienced more self-reported workload, and were less satisfied with their search performance. Furthermore, eye-tracking research has demonstrated that users with Low-PS are also physiologically different to users with High-PS. In terms of a lower fixation rate, Low-PS users are just not physically able to scan as much as their High-PS counterparts, which explains why their searching experiences will inevitably differ.

Consequently, a common aim shared amongst researchers referred to the need to develop dynamic IIR systems that can identify when a user has Low-PS, and subsequently adapt to accommodate them. Whilst research has begun to predict a user's PS levels—using data from eye-tracking—it still remains unknown what a system that could benefit users with Low-PS could be. Some suggestions have been proposed, such as creating additional tools to help users navigate documents, yet, these tools have never been explained as to what they would involve. Furthermore, in research which identified interactions with PS—such as, Low-PS users having worse accuracy reading data visualisations in coloured boxes, in comparison to radar graphs—no conclusive explanations were given for why one visualisation, over the other, led to Low-PS users achieving the higher accuracy. Consequently, the first research question of the present thesis was:

- ***RQ1) How can a user with Low Perceptual Speed be helped to achieve a more positive online search experience, both subjectively, and objectively?***

With more understanding of how to help a user with Low-PS succeed in their search being an under explored area of research, previous literature on visual perception more generally was examined. In particular, theories on visual clutter were highlighted as another under explored research area in IIR, despite their importance concerning how users become cognitively overloaded during the presence of too much stimuli of a certain type. Although research in IIR appeared to assimilate to theories on visual clutter —such as users having greater difficulty finding relevant documents when more results per page were presented, or interfaces being rated as less usable amidst more images being present—none of these studies ever explored how individual differences in perception—namely, PS—could impact results. Yet, if a user with Low-PS struggles completing a search task, it is possible that when an interface becomes too cluttered—such as through additional images or results being present—then this could negatively affect them even more. This speculation was further supported when research on visual complexity more generally (irrespective of visual clutter, albeit could be considered as a possible manipulation of clutter) identified that users with Low-PS had significantly lower task performance when more data was visible.

However, it cannot be assumed that reducing all visible clutter present would be beneficial to users with Low-PS without further research. This is especially important, considering that other research (irrespective of PS), has found benefits of increased clutter being present. For example, too little stimuli can elicit boredom, and so some form of clutter may actually stimulate engagement and increase interaction, which could result in greater task accuracy. Alternatively, distractions can replenish mental resources and reduce workload, and thus this may offer one remedy for users with Low-PS to complete more positive searches.

Whilst clutter can be defined in many ways, one area of research in web search that has investigated clutter, refers to advertising. However, this has only been driven from the perspective of increasing attention and positive attitudes towards the ad, and not on how it affects a user complete their search more generally. Whilst there is a general agreement that subjectively, users report ads to be annoying and subsequently attempt to ignore them, it is unknown whether users with Low-PS are able to ignore them, or whether they become more cognitively overloaded in their presence, and subsequently complete poorer searches. Furthermore, although users may *think* that they can ignore ads, other research suggests that ads are still peripherally processed. As clutter can create a problem even in peripheral vision, this reaffirms the importance of investigating how cluttered ads truly impact a user during IIR, and what interactions may exist for people with different levels of PS. Therefore, the second research question was:

- ***RQ2) What is the relationship between Perceptual Speed and visual clutter in the form of advertisements during an IIR task?***

Before different forms of advertising clutter can be experimented with in regards to how it affects users with different levels of PS, it was acknowledged that clutter cannot only be defined by the quantity of visible elements. Instead, many other factors can affect the perception of clutter, and in advertising clutter, research has focused on the congruence of the clutter—in other words, whether it provides information that is appropriate to the search task or not. Yet again, these investigations have not focused on the effect of congruent clutter on a user's search experience overall, or whether the congruence may benefit, or hinder users with Low-PS.

Inspired by research on congruence more generally, congruent ads may present an ideal search scenario for a user: they may increase a user's interaction, and subsequently provide additional information that could lead searchers directly to their goal. It is also possible that congruent clutter may ameliorate any negative effects of clutter in general, all while providing the right amount of visual stimuli that can replenish a user's mental resources, in addition to ensuring a user does not get bored. Furthermore, users with Low-PS who struggle with visual processing may be additionally helped by congruent ads, as the congruent images could help the user confirm they have searched for something relevant in a more efficient way than requiring the user to scan every element on the interface. Then, users with Low-PS would be more negatively affected by the presence of incongruent ads: they might provide false information that the user was performing the search incorrectly; and they may require a greater cognitive effort to process, as a user works out why they have been shown something off-topic. If this were the case, then this could offer one suggestion for designing a search interface that can accommodate users with Low-PS. Alternatively, Low-PS users may be so negatively affected by clutter that ads, regardless of congruence, impede a user's search performance and experience.

Thus, research is needed to examine the relationship between PS, clutter, and the congruence of visible clutter, on a user completing an IIR task, leading to the final research question of:

- ***RQ3) Does clutter that is congruent with the task improve or worsen the search experience for users with Low Perceptual Speed?***

2.10 Chapter 2 Summary

Combining the gaps in previous literature together, this Chapter concluded with the following three main research questions which provided the framework of investigation for the present thesis:

- **(RQ1) How can a user with Low Perceptual Speed be helped to achieve a more positive online search experience, both subjectively, and objectively?**;
- **(RQ2) What is the relationship between Perceptual Speed and visual clutter in the form of advertisements during an IIR task?;**
- **(RQ3) Does clutter that is congruent with the task improve or worsen the search experience for users with Low Perceptual Speed?**

In the next chapter, Chapter 3 Methodology, the manner in which these questions were answered are outlined.

Chapter 3

Methodological Approach

3.1 Chapter 3 Overview

This Chapter provides an overview of the different factors considered when deciding upon the overall research methodologies employed in this thesis.

3.2 Rationale

The outset of the research undertaken for this thesis was inspired by a general desire to better understand the overall concept of PS, so that users with Low-PS could be helped to achieve more positive and effective online searches during IIR. However, upon greater familiarisation of previous literature, it became evident that a step was needed before exploration of helping users, as there appeared to be problems with how PS had previously been—and was still continuing to be—measured. The problems identified were divided into themes, where the reliability and validity of PS test content, administration, analysis, and result dissemination were discussed (See Chapter 4). For example, it was identified that different studies were categorising users as having a 'low' PS, based upon the median of their unique sample. This meant that a user with the same score in different studies could be classed as 'low' in one, but 'high' in another.

With the problems in measurement identified, it remained unknown whether the overall concept of PS was reliable or valid, or whether it had been built upon a plethora of flawed measurement. Indeed, through a reiterative process of further reading around the area, other research was then identified which appeared to suggest that there *were* other problems in PS research. This included research which stated it was measuring PS, but then no results were reported (e.g. [137, 206, 219]), or identifying papers which included citations stating users with Low-PS had lower task performance, and yet when following the citation trail, the research never explicitly explored task performance (e.g. [67]). Consequently, although the outset of the present research aimed at helping users with Low-PS to improve their search outcomes, it became evident that firstly, a thorough investigation was needed into the overall concept.

3.3 Choosing a method of Review

Moving beyond literature reviews—which although describe and appraise previous work, do not describe specific methods for how the reviewed studies were identified— a more thorough understanding needed to be generated by evaluating the entire area of PS. Specifically, a guiding research question concerned what claims had actually occurred regarding PS in the field of IIR, and beyond that, it was deemed necessary to evaluate all claims to understand how every result had been explained; it would be pointless to explore how to help a user with Low-PS to achieve a better search, if it actually emerged that Low-PS were not negatively affected during IIR, but instead, the concept had been built upon flaws in citations and unreliable testing. Otherwise, if it was confirmed that Low-PS users *were* negatively affected, then knowing all claims that have occurred would help guide research ideas for what ways might be able to help them.

To answer these questions, different types of exploration were considered, namely: a literature-based discovery, scoping review, narrative or critical/discursive literature review, survey, and free-form discussion. However, these methods were not deemed appropriate for the following reasons.

Firstly, it is commonly acknowledged that in academic literature, "*There is too much information for anyone to read, much less understand*" [106]. This awareness has led to the creation of the Literature-based discovery (LBD) method, which has been defined by Hristovski et al. [2015, cited in [217]] as something which "*generates discoveries, or hypotheses, by combining what is already known in the literature*". This method is especially useful in merging together evidence across multiple domains, where knowledge in one area is not known outside of it [217]. Given that PS has been used in multiple areas beyond IIR—from Medicine to Psychology—the LBD was originally seen as a method that could potentially offer a way to bring together interdisciplinary information sharing, such as by understanding how PS tests have been used and applied in Psychological research. However, on further inspection, LBD is built upon an assumption, named the ABC model, where "*if concept A is associated with a concept B and that concept B is associated with another concept C, then concept A is associated with concept C where the B-concept denotes the association/relationship between the two concepts A and C*" [217]. Consequently, this would not be appropriate to further understand how to explain the results of PS testing in the field of IIR.

As a more general method of understanding every claim that has occurred regarding PS during IIR, a Scoping Review was then considered as an alternative method. Specifically, scoping reviews "*can report on the types of evidence that address and inform practice in the field and the way the research has been conducted*" [173]. However, there is some criticism of this method, with some researchers arguing that "*there is not yet a universal study definition or definitive procedure*" [187] and "*there was no universally recognized definition of scoping reviews nor a commonly acknowledged purpose or indication for conducting them*" [173]. Instead, others have

agreed that their use is for when emerging evidence needs examining, when it remains unclear what specific questions should be posed [173]. Similarly, whilst scoping reviews may synthesize evidence from different studies, there is no evaluation or critical appraisal of this evidence [187]. Yet as the research in the present thesis *was* guided by specific questions, and there was also a need to evaluate this evidence, a scoping review was not selected as the chosen research method.

Various other methods were also considered. Specifically, articles were examined that provided guidance to researchers on choosing the best approach. For example, in a survey of 105 institutions offering PhD programs, there was agreement that narrative or critical/discursive literature reviews should be replaced with Systematic Reviews [190]. This confirmed that the present research needed to move beyond literature reviews. Then, other research reinforced the benefits of a Systematic Review, in comparison to other methods, such as surveys: "*The Systematic Literature Review method [12,49,56] is radically different from conducting a survey or from free-form discussion of related literature in that it follows a rigorous, well-defined procedure guaranteed to produce reliable, reproducible results*" [229]. Consequently, the advantages and disadvantages of a Systematic Review were further explored.

3.3.1 Systematic Review Advantages

After examining various articles, the following advantages of a Systematic Review were identified:

- "*Systematic literature reviews are highly recommended for students who are starting their research and wish to evaluate effectively a particular area and clearly understand how their proposal may contribute considering what has already been published*" [232].
- "*A systematic review may be undertaken to confirm or refute whether or not current practice is based on relevant evidence, to establish the quality of that evidence, and to address any uncertainty or variation in practice that may be occurring*" [173].
- "*A systematic review was chosen as the method of choice for this article as it has the potential to identify all of the relevant scholarly research on a particular topic*" [165].
- "*A systematic review 'uses explicit, systematic methods that are selected with a view to minimizing bias, thus providing more reliable findings from which conclusions can be drawn and decisions made*" [173].
- "*Systematic reviews summarize available literature using specific search parameters followed by critical appraisal and logical synthesis of multiple primary studies*" [166].
- "*Conducting a systematic review may also identify gaps, deficiencies, and trends in the current evidence and can help underpin and inform future research in the area*" [173].

- "*Aveyard and Sharp defined SRs as original empirical research because they 'review, evaluate and synthesise all the available primary data, which can be either quantitative or qualitative' [2]. Therefore, a SR represents a new research contribution to society and is considered the highest level in the hierarchy of evidence*" [190].

Overall, given the rigorous level of evaluation and comparison possible for variation in current practice—all whilst minimising bias, and thus increased reliability of conclusions— a Systematic Review was chosen as the method of choice in the present research for: a) categorising all previous claims of PS in IIR; and b) evaluating these claims.

3.3.2 Systematic Review Disadvantages

Whilst a Systematic Review was selected for the present research, it must also be acknowledged that this method is not without limitations which were considered before the review was undertaken.

Firstly, some researchers have argued that Systematic Reviews lack originality: "*Some individuals believe that a SR is not original research. Indeed, it has been suggested that SRs as 'secondary research' are different than 'primary or original research', implying that they are inferior and lacking in novelty and methodological rigour as compared to studies that are considered primary research*" [190]. However, in a survey of journal editors that were asked their opinion on the originality of Systematic Reviews, the majority (71%) indicated that they should be classed as original studies, and almost all journals (93%) published them (Meerpohl et al., cited in [190]). Nonetheless, creating original research was not the aim of answering the current research questions, and instead the focus was on the usefulness of information that could be generated from the review, and thus this limitation was negated.

Secondly, the time involved in thoroughly completing a Systematic Review has been highlighted as a deterrent by Librarians, where the most time appears to be spent on developing a search strategy [50]. This makes sense, as other research has highlighted three further challenges of developing a Systematic Review search strategy; namely involving formulation, refinement, and documentation [145]. In particular, as search strategies are manually developed, their development can contain biases and subjectivity, and too many studies may be retrieved [145]. However, biases can be reduced through the use of having an expert in the field verify the process [145], and this was achieved in the present research by the current PhD Supervisor. Furthermore, a year of the PhD programme could be dedicated to conducting and completing the Systemic Review thoroughly, and thus although time consuming, it was considered a worthwhile investment into the overall research.

Even with verification of search strategy and documentation by another researcher, the process of categorising information in a Systematic Review is still conducted manually, and thus it is subject to accidental errors of logging. However, to minimise this from happening,

various coding schemes were implemented. These were inspired by other research where a large table was created which contained questions, and various categories comprised the answers, such as: "*Step 1: Does the paper contain a table or figure of mean effectiveness, mean user performance, etc. that appears to deserve significance testing? If YES, select and record the name of one such table or figure. Step 2: Does the paper conduct a significance test? Assign exactly one category from (A)-(I) shown in Table 2*" [199]. Consequently, a similar coding structure was implemented in the present research, which included the creation of an excel document with questions such as: "*Does the paper report a measure of search behaviour?*" and then the answers to these questions were further narrowed down to log behaviour that concerned: a) queries; b) views; c) clicks; and d) time. These results were then colour-coded to reflect variables that returned results that were significant, non-significant, or not-reported. For a full description of the Systematic Review Method, please refer to Chapter 5. However, of importance for justifying the overall methodological approach, is that this systematic process minimised errors in manual reporting, whilst additionally generated a quantitative estimate of the studied phenomenon. Additionally, the main researcher immersed themselves in all available papers and re-read them multiple times, so that when double-checking the written codes, any mistakes would have been obvious and easy to correct. Consequently, the limitations identified with the use of Systematic Reviews were able to be minimised.

3.4 Continuing the Research Cycle

The results of the Systematic Review confirmed that many significant and important effects of PS had occurred during IIR. Yet at the same time, some contrasting findings did emerge which required further investigation. For this reason—and combined with the knowledge from the detailed literature review where problems in PS measurement were identified—the next stage of the present research involved developing new PS tests before it was possible to go back to the first main research question of exploring how a user with Low-PS could be helped to achieve a more positive online search experience. For a detailed explanation of how the new tests were created, please refer to Chapter 6.

3.5 Selecting an empirical research method

Next, to begin exploring how users of Low-PS ability could be helped during IIR, there were two further main research questions to consider: **RQ2) What is the relationship between Perceptual Speed and visual clutter in the form of advertisements during an IIR task?**; and **RQ3) Does clutter that is congruent with the task improve or worsen the search experience for users with Low Perceptual Speed?** Consequently, various research methodologies were again considered. The methodologies considered were also a result of the Systematic Review presented in Chapter 5, which categorised every method implemented in each

paper in the field of PS during IIR. This identified that past studies had mostly administered experiments, and only a few had utilised surveys or interviews.

Firstly considering surveys and interviews, these had been administered to: a) compare whether levels of PS correlated with self-reported ratings of computer proficiency [243]; b) analyse whether a relationship existed between PS and familiarity with technology [178]; and c) exploring consumer satisfaction of online shopping with PS [119]. Whilst all of these studies may have generated new knowledge about PS—and have specifically focused on user experience—they did not directly help to understand what parts of the search process users with Low-PS struggle with, and thus it remains unknown how to help these users. For example, if users with Low-PS had worse satisfaction of online shopping, why this was the case would not be clarified. Of course, different surveys or interviews could be created, such as gathering a group of people with Low-PS and High-PS, and explicitly asking them which aspects of search they struggled with—if any. However, people may be unaware of their own ability or reasons.

Alternatively, surveys could be developed with a more specific focus. For example, focusing on the role of visual clutter impacting PS, an example of a survey question could involve something such as: "*Do you find the presence of adverts distracting?*". However, this could incur the problem that there is often a difference between what people think, and what people do [142]. Furthermore, the results would be limited in that they would be unable to generate confidence as to whether some form of clutter was good, or bad, for a user to complete a more positive search or not. Consequently, the use of a survey or interview—at least, as a stand-alone method—was not chosen in the present research.

Another method considered, which did not appear in previous related literature, involved the use of observation. For example, users with differing PS levels could be observed whilst they undertook a search task, and the researcher could look for similarities or differences between users, such as in their search strategy or behaviour. This could also be implemented with a specific focus on visual clutter, such as through observing users across different types of cluttered, or uncluttered interfaces. This would elicit rich qualitative data that could help to better understand users of differing PS levels. For example, perhaps differences in time spent on certain aspects of the interface would be noticed between users, such as Low-PS taking longer on a screen where more images were visible. However, observation data can be limited, as observer bias could impact the interpretation of the data, leading to inaccuracies and overall validity being compromised. Additionally, due to the time and resource-intensive nature of observation, this tends to minimise the size of sample observed, and thus this makes the results less able to be generalised to others. Furthermore, even if patterns emerged for differences between users, based on their PS, there would be no definitive cause or effect generated. For example, maybe users with Low-PS may take longer in the presence of certain interfaces, but whether this was

the direct effect of clutter or not, would remain unknown. Therefore, the use of observations as a research method was disregarded in the present thesis.

Instead, experiments do allow cause and effect to be established. This has been noted as especially beneficial in the field of IIR, where an evaluation of methods for evaluating interactive information retrieval systems with users stated that: "*Laboratory studies are good with respect to the amount of control researchers have over the study situation. This is particularly useful when trying to isolate the impact of one or more variables*" [120]. As the current research was interested in definitively understanding how clutter impacted PS during IIR—whilst also investigating different types of clutter—an experiment thus appeared to be an appropriate method to implement.

3.6 The use of experiments

In addition to establishing causality, the use of experiments for the present research offered many other advantages and disadvantages, which were all considered before further implementation was initiated.

3.6.1 Experiment Advantages

Firstly, as experiments were the most commonly used approach in previous related research, this would allow the current results to be more easily situated in the context of pre-existing literature. Secondly, with the level of control required to establish causality, the methodology can be more easily reproduced by others, which can help to verify or contradict results, as findings using the same methods across different scenarios can be compared and contrasted. Thirdly, whilst a limitation of experiments in general concerns a lack of ecological validity, given that the current research was situated within IIR, an online experiment would be akin to everyday searching, and thus high realism would be achieved. Fourthly, unlike observations, experiments are less subject to human error: observer bias would not influence the results. Finally, given that an experiment can be deployed online, this means that multiple users can be tested simultaneously. This is advantageous because any potential limitations of lab-space is minimised, and the researcher's time is protected.

3.6.2 Experiment Disadvantages

Like any other research method, experiments are also subject to limitations. Although some common limitations with experiments can be mitigated within the context of IIR—as was just described in the Advantages Section 3.6.1—other limitations must be considered. Specifically, if an experiment is deployed online, then a user completing the experiment may be interrupted by external events, and thus the researcher would have no control and confounding variables may impact results. Yet, this could in fact be regarded as a positive, because it would make

the experiment more ecologically valid and akin to everyday searching where interruptions often occur. However, to ensure a high level of ecological validity—whilst also allowing specific variables to be isolated and studied systematically—the development of an experiment requires extensive training, piloting, and thus a long development time must be anticipated.

3.7 Implementing an Experiment

To ensure the experiment would answer the research questions, an explanatory study was selected. This is a type of experiment which prioritises establishing causality. However, as explained by Kelly et al. [120]: "*Despite the name, it is important to note that not all explanatory studies offer explanations - many just report observations and statistics without offering any explanation*". Instead the focus requires variables of interest to be isolated and studied systematically. Consequently, to answer the main research questions, the experimental framework needed to meet the following requirements:

- Can allow for the modification of different types of visual clutter, in a controlled manner.
- Can record behavioural and performance metrics related to a search task.

Additionally, as well as measuring performance-based measures related to the outcome of the search interaction, the benefits of gathering feedback from users—such as their feelings about the system—has also been noted in IIR research [120]. Thus, although surveys were not chosen as the main research method, they were also implemented after each experimental task, in order to gain richer information about user experience.

Whilst an explanatory experiment could be designed in many different ways, the motivation for selecting a specific search task has been described in detail in the later Method section of the corresponding chapter (Chapter 7, Section 7.5)

3.8 The effect of COVID-19

Given that this research was being undertaken when COVID-19 occurred, the choice of research methods selected were adapted. Specifically, many governments and health authorities around the world imposed restrictions on in-person gatherings and travel during the pandemic. Researchers had to comply with these regulations to prevent the spread of the virus.

Originally, the experiment designed for the present research was piloted as an in-person study during December 2019. However, with changing health guidelines, the experiment had to be converted to an online-only study to stay within legal and ethical boundaries, whilst maintaining the safety and well-being of both the researchers and participants. This meant that some factors of the research were not possible to implement remotely; specifically being, the use of eye-tracking alongside the IIR experiment, and, validating the newly created digital

PS tests with the original paper-based versions. The effects of these two factors were however considered, for how they might have affected the research, and these are described below.

3.8.1 Eye-tracking

Firstly, eye-tracking had originally been considered as a useful tool to incorporate into the research. The benefits of eye-tracking during IIR experiments have previously been explained by Gwizdka & Cole: "*One source of physiological data that can be connected to information acquisition during search interactions is eye movements. In textual IIR primary information acquisition is mediated by eye movement patterns in service of the reading process. Eye movements are known to be cognitively controlled. In human information-interaction, eye movements are sources of information about three types of cognitive processing: attention, semantic processing and decision making*" [101]. By further understanding different types of cognitive processing during IIR, this helps the researcher to examine how individual cognitive abilities affect the search process [98, 100].

Consequently, the researcher originally desired to monitor eye movements whilst participants completed the IIR experiment, and thus throughout 2019, multiple training courses in eye-tracking were undertaken. However, as specialised equipment is required in a controlled environment to implement eye-tracking, this was not possible to use remotely and so when the experiment was deployed online, eye-tracking was not conducted.

Nonetheless, this was deemed not to be detrimental to the overall research, as other research has demonstrated that "*certain cognitive and motor control mechanisms are embodied and reflected, to some extent, in our mouse cursor movements and online interactions*" [15]. Therefore, even without eye-tracking, attention can be inferred through mouse cursor movements, and this has been demonstrated in previous remote research studies that have focused on attention [164]. We thus ensured that in the present experimental infrastructure, mouse hovers were logged to combat the lack of eye-tracking.

3.8.2 Validating the digital PS tests

Whilst not using eye-tracking was deemed to not affect the overall research rigour, there was another part of the research which was affected by remote research. Specifically, digitally created PS tests were designed and created (for the details of this, please refer to Chapter 6). However, it would have been preferable to validate the new digital version to the original paper-based PS tests.

However, in other research which did compare participant's performance in paper-versions of PS tests against computer versions, good reliability was observed between the two: "*Results indicated that these new measures provide both high levels of reliability and substantial validity for performance on the two skill-learning tasks.*" [1]. This would imply that a valid digital test, in comparison to a paper-based version, is possible to achieve. However, these tests were from

the field of Psychology, and therefore they differed to Ekstrom's tests which have been the most commonly used in Computer Science. Therefore, updating the paper-based PS tests that have been used in the domain of Computer Science, and converting them into digital tests, was still necessary.

Although a fully factorial within-subjects design experiment could have been utilised—where 50% of participants completed the paper-based version first and then digital, and the other 50% completed them in the opposite order— this would have meant that an in-person study was required so that the paper-based version could be administered. Similarly, a between-subjects study could also have been implemented—where 50% of participants completed the paper-based, and 50% completed the digital version—yet, this would also have required at least one of the conditions to be an in-person study. However, due to the restrictions of COVID-19, a remote study could not have accommodated a paper-based PS test. Nonetheless, not conducting an in-person study had many advantages, outlined below.

Firstly, an online study was preferred in comparison to an in-person study, to improve the ecological validity of any results found. Given the artificial environment of an in-person study, it is widely agreed that the results gained from such experiments lack ecological validity [28]. As completing an IIR task is something most often done at home, or in a busy office space, it was believed that a controlled laboratory setup would not be appropriate and have less external validity, and thus it would not be possible to generalize any results found. Instead, an online study allows participants more autonomy with regards to the time and location of their participation [93]. Whilst it could be argued that this offers a lack of control over the environment, which may lead to reduced attention on the task—if for example, participants are listening to music or watching TV in the background [182]—it was believed that this kind of environment would be more natural to everyday websearch and IIR, where online searches are usually conducted in busy offices, cafes, or high-pressured information environments.

In addition to less ecological validity being a concern if the study was conducted in-person, the physical presence of a researcher could also have impacted results negatively if a paper-based test had been administered. Not only could the researcher presence produce social pressure on the participant to perform well on the tasks [93], but the role of demand characteristics could also alter behaviour—especially when answering the survey questions that corresponded with the IIR task—as participants may attempt to answer as they believe is expected of them, or adapt their behaviour accordingly [163]. In contrast, online experiments have proven to be successful without the presence of a researcher, whilst also allowing a more representative sample of the internet population to be studied in their natural habitat [182].

There are further benefits of lack of direct physical contact between researcher and participant, which has been highlighted by many different research studies. These include improved

ethical participation where: a) greater participant anonymity can be achieved, which subsequently can improve response rates to any sensitive questions; and b) there is reduced social pressure to continue participation if they wished to withdraw during the experiment if discomfort arose (McCabe, 2004, Birnbaum, 2004, Fricker & Schonlau, 2002, Kaplowitz et al., 2004, cited in [28]). Furthermore, even without the control of a physical researcher being present, the results of well-known in-person experiments have been replicated in online experiments, demonstrating the reliability of online experiments [182]. Consequently, combining all of these reasons together, an online-only study was still regarded as a desirable way to distribute PS tests.

3.8.3 Using Prolific

As described above, eye-tracking and conducting paper-based PS tests were not possible to implement in remote research. However, given the advancement of various remote research platforms in recent years, this allowed high-quality data to still be gathered remotely. Specifically, the present research was administered remotely using the platform Prolific[1]. This specific platform has many advantages which are outlined below.

- **Diverse Participant Pool**: Prolific provides access to a large and diverse pool of participants from various demographic backgrounds. This diversity can be crucial for studies that require a representative sample, and given that the present research was focused on web-search, gathering participants who use the web around the world enabled this representation.
- **Quality Assurance**: Prolific has measures in place to ensure data quality. Participants are incentivized to provide accurate and thoughtful responses, as their profiles and ratings are impacted by their performance. Additionally, researchers can set attention checks and other quality control measures within their studies.
- **Easy Payment**: Prolific handles payments to participants efficiently, simplifying the process for researchers. This can be particularly beneficial for international studies where payment logistics may be complex.
- **Ethical Considerations**: Prolific is committed to ethical participant compensation. Researchers can be confident that they are compensating participants fairly for their time and effort, aligning with ethical research practices.
- **Reduced Recruitment Bias**: Traditional recruitment methods, such as using university subject pools, may introduce bias into the participant sample. Prolific's diverse participant pool helps mitigate this bias, making it suitable for studies aiming for a more representative sample.

[1] `https://www.prolific.co/` – last accessed July, 2020.

- **Cost-Effective**: While there are fees associated with using Prolific, it can be cost-effective when compared to traditional recruitment methods that involve physical resources or time-consuming efforts like in-person recruitment.
- **Flexibility**: Prolific supports various study formats, including both surveys and experiments, making it suitable for a wide range of research methodologies.
- **Data Security**: Prolific has security measures in place to protect participant data and ensure compliance with data protection regulations (e.g., GDPR). This is crucial for studies that handle sensitive information.
- **Researcher Community**: Prolific has an active community of researchers who share insights and best practices, making it easier for new researchers to get started and optimize their studies.
- **Participant Feedback**: Researchers can collect feedback from participants, helping them improve the study design and overall research process.

While Prolific offers numerous advantages for research studies, it is essential to consider the platform's limitations, such as the potential for participant fatigue, the need to compensate participants fairly, and the costs associated with using the service. However, these factors were weighed against the numerous advantages, and it was decided that Prolific provided the best way to adhere to remote research, whilst not compromising the overall research goals.

3.9 Chapter 3 Summary

Overall, given the initial research outset discussed in Section 3.2, a qualitative analysis of how PS had previously been measured created various problematic themes involving the test content, administration, analysis, and how results had been reported. As these problems questioned the overall reliability and validity of PS, this inspired a more thorough analysis of the overall PS concept to be undertaken. Section 3.3 then discussed a number of different types of review that were considered, and this concluded with a justification of the approach chosen—a Systematic Review—based upon various advantages (Section 3.3.1) and disadvantages (Section 3.3.2). Then, in Section 3.4, a brief description for what was required to continue the research was provided. This progressed to again consider alternative approaches that have been used in previous literature, in Section 3.5. However, based upon relative advantages (Section 3.6.1) and disadvantages (Section 3.6.2), alongside choosing a methodology most suited to answering the main research questions outlined in the literature review (Chapter 2), the implementation of an experiment was identified as the most appropriate research methodology to pursue. Section 3.7 then highlighted the requirements needed in the experiment—namely, that it would be explanatory, allow for different types of clutter to be implemented, and behavioural and performance

metrics must be gathered. Additionally, the benefits of also measuring user experience, in the form of surveys, was also discussed. However, details of the exact experiment implementation and survey questions will be discussed later, in the specific Method section of Chapter 7. The next Chapter will instead focus on the qualitative analysis which identified problematic themes in previous PS measurement.

Part II

Evaluation of Perceptual Speed measurement and the concept overall.

Chapter 4

Problems with Perceptual Speed Tests

4.1 Chapter 4 Overview

This Chapter investigated how PS has previously been measured. In doing so, many problems with the PS tests were identified, which were categorised into problems involving the test content, administration, how results have been analysed, and dissemination of findings. Consequently, challenges and recommendations for how PS *should* be measured were discussed and proposed.

4.2 Introduction

Perceptual Speed (PS) is a cognitive ability defined by an individual's accuracy and speed to scan information while completing visual search tasks [10]. Prior studies using PS tests have demonstrated that PS affects multiple factors in IIR. Thus, with greater knowledge of PS, systems could be designed that accommodate users with Low Perceptual Speed (Low-PS) to improve their overall search experience and performance.

In order to answer the first main research question of this thesis, ***(RQ1) How can a user with Low Perceptual Speed be helped to achieve a more positive online search experience, both subjectively, and objectively?***, the first step of the research involved understanding how PS had previously been measured. This was deemed necessary, so that participants with Low-PS and High-PS could be compared completing IIR tasks. Consequently, a sub research question emerged: ***(1a) How has PS previously been measured during IIR?***.

By answering this question, the present chapter also aimed to stir discussion between researchers by drawing awareness to issues identified with PS measurement. As a result, the challenges involved in advancing how future PS tests are used in IIR are further discussed. Finally, the validity and reliability of PS measurement is considered, and recommendations are

proposed for anyone wishing to incorporate PS testing in the future. Therefore overall, three main contributions to the IIR community are explored:

- Firstly, following the advice of [177], a main aim of this chapter is to facilitate dialogue amongst IIR researchers by quantifying and making others aware of the methodological, reliability, and validity issues associated with PS testing administration, analysis, and reporting.
- Secondly, after considering the limitations of PS testing, the current challenges that the IIR community needs to address regarding PS testing are discussed.
- Finally, a series of recommendations are provided for enhancing the quality of PS testing in IIR.

4.3 Perceptual Speed Testing in IIR

With PS being a type of cognitive ability, it's underlying neural mechanisms are thought to be automatic and fairly stable throughout an individual's life [183]. As PS varies between individuals, multiple tests have been developed that attempt to detect this cognitive ability such as: the Minnesota Clerical Test (1965) [127], Ekstrom's (1976) Kit of Factor-Referenced Cognitive Tests [82]; Wechsler's (1981) Digit Symbol Substitution Test (cited in [205]); and Salthouse & Coon's (1994) Letter Comparison Test [86, 201]. Irrespective of the exact test used, they all follow a similar format that involves scanning a list of stimuli and identifying certain targets against a set time period. People who are most accurate at identifying targets in the fastest amount of time are said to have High-PS, while people who make more mistakes and take longer are considered to have Low-PS [83].

In the context of IIR, studies have predominately used tests drawn from or based on *Ekstrom's Kit of Factor Referenced Cognitive Tests* [82]. The kit comprises of three different PS tests that researchers may choose to use. However, Ekstrom suggest that in order to fully deduce a cognitive factor, at least two tests should be administered [83]. The three Ekstrom PS tests to choose from involve numbers, words, or symbols and are shown in Figure 4.1 and described below:

- *Finding A*'s: Participants must effectively scan columns of words and select any that contain a letter "a".
- *Number Comparison*: Participants are given pairs of numbers, and are required to indicate whether the numbers are the same or different by placing a cross on non-identical pairs.
- *Identical Pictures*: Participants are given a symbol and must select the identical image against a choice of five.

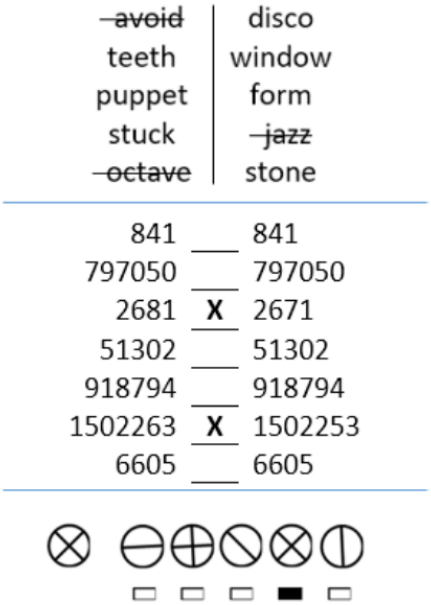


Figure 4.1: Sample PS Tests based on Ekstrom's Kit [82]. Top: *Finding A's* Test. Middle: *Number Comparison* Test. Bottom: *Identical Pictures* Test.

However, although one scoping review exists that analysed over 2100 articles in IIR, and "*perceptual speed*" was used as one of the search terms, there was no explicit discussion or results of PS tests [176]. Instead, an amalgamation of cognitive abilities were merged together to conclude that, as a result of issues around measurement and generalisability, it was unknown how these individual differences truly affected search outcomes [176]. This appears surprising when Ekstrom's PS tests have been widely used since their development in 1976. With such a long time period of use, the reliability of these tests would be thought to be high. However, as O'Brien & McCay-Peet [177] pointed out, items in many studies lack validity and reliability evaluation. Thus, the present chapter aims to evaluate the literature concerning PS and IIR, in order to make researchers aware of any current limitations, and suggest future recommendations for improving PS usage.

4.4 Review Process

To provide the basis for the present analysis and discussion, library searches were performed to identify studies which had used PS tests in the context of IIR. Thus, the search criteria was defined as follows:

Firstly, the Association for Computing Machinery (ACM) Digital Library (DL) was searched– which contains references to core IIR resources, conferences and journals. The initial search for *perceptual speed* returned 19,451 results. Subsequently, inverted commas were added to the query to ensure papers were returned that were not dealing with 'perceptual' and 'speed' as

separate entities. This returned only 12 results, all of which have been used for analysis in this chapter.

This process was further repeated using the same query of *"perceptual speed"* in the university library, which encompasses a huge selection of many databases and returned a much larger result of 6,064 entries. Brief manual scanning revealed that many of these results were predominantly coming from the medical industry. Therefore, to maintain the focus on IIR, the search query was changed to *"perceptual speed" AND "information retrieval"*, with the filter of peer-reviewed items only, which brought back a more manageable 69 results. To ensure that PS was one of the main focuses of the paper, our inclusion criteria involved manually reviewing each of the 69 abstracts to eliminate any that did not indicate the use of PS tests in the context of an IIR study. This left 11 papers.

Finally, 16 more papers were discovered through reference crawling of the 23 already found papers. Although seven of these papers were not directly IIR, but rather originated from a psychological background, they were still included to explain the psychological principles behind the fundamental PS tests.

With an overall corpus comprising of 39 papers published between 1965 and 2019, we began reviewing these papers in search of main themes. In this approach, data analysis is not conducted with pre-specified questions that need answering, but rather themes emerge from the data itself [196]. Consequently, through a reiterative process of paper reading, themes began to emerge regarding PS test content, administration, analysis, and how results were reported. Rather than quantitatively coding all possible themes, instead, we followed a more qualitative approach to accompany this perspectives paper. This involved reporting the main themes that with others awareness, we believe would help improve PS testing for future studies.

Of the 32 papers that used PS tests in IIR studies, 30 used one of Ekstrom's test, while the other two used the Minnesota Clerical test. For the purposes of discussion we will focus on PS tests in light of Ekstrom's tests.

4.5 Main Themes of PS Tests

As a result of letting themes emerge from the literature on PS in IIR, many uncertainties regarding PS test content, administration, analysis, and how results were reported have been identified and split into six main themes below: 1) No standardised thresholds; 2) Inconsistent reporting of results; 3) Unclear marking instructions; 4) Different formats; 5) Limited linguistic reasoning; and 6) Outdated Administration and content.

4.5.1 No Standardised Thresholds

One of the most notable uncertainties with PS tests is that despite being over 40 years old, and many papers have used them and referred to "low" and "high" PS levels, there are no

Table 4.1: Perceptual Speed results reported in the selected studies.

Study	PS Test	Possible Range	Mean (SD)	Median	Min, Max
EKM, cited in Turpin et al. [225]	Finding A's	-	47 (14.9) = Males, 54 (14.9) = Females	-	-
Turpin et al. [225]	Finding A's	0-200	51.94 (10.41)	51	34, 74
Arguello & Choi [18]	Finding A's	0-200	64.16 (12.00)	63	44, 90
USAF, cited in Brennan et al. [46]	Number Comparison	-	47.94 (12.32)	-	-
Brennan et al. [46]	Number Comparison	0-96	44.38 (10.58)	44	25, 73
Crabb & Hanson [71]	Number Comparison	-	46.63 (6.04) = Young 45.08 (6.94) = Old	-	-
Allen [9]	Number Comparison	37	30.1 (8.8)	-	-
Toker et al. [222]	Identical Pictures	-	85.70 (11.64)	-	54, 96
Allen [9]	Identical Pictures	42.5	80.9 (11.4)	-	-

standardised thresholds for what defines Low-PS or High-PS. Rather, only a few papers have even explained how they categorised PS: participants were assigned to a low or high group, based on a median split of PS scores [4, 209, 210, 225]. The problem with reporting low/high based on a median split without providing the scores is that it is not possible to compare across studies, nor can one know what is low or high, or whether there is any statistical difference between the groups.

Table 4.1 presents a summary of the IIR papers that report the median score from the PS test used. With further examination, a huge discrepancy in results can be noted. In Turpin et al. [225], participants were classified as having a Low-PS if they scored between 34 and 51, and a High-PS if they ranged between 51 and 74. On the contrary, Arguello & Choi [18] filtered Low-PS individuals as those scoring between 44 and 63. Therefore, despite the same identical tests being administered, if a participant scored within the range 51-63, one study would classify the participant as having High-PS, whereas the other study would categorise the participant as having a Low-PS. With such discrepancy in analysis depending on the individual sample of participants tested, this greatly reduces the comparability of results across studies.

4.5.2 Inconsistent Reporting of Results

Out of the papers reviewed, only six, or 15.4%, reported exact figures for their PS test results (See Table 4.1). Instead, the majority of existing literature concerning PS tended to only report explicit figures that referred to the significant effects PS has had on another part of an experiment. For example, in a study that examined whether PS affected how long it took for a user to retrieve a relevant document [4], the only PS figures reported were that users had been grouped into Low-PS and High-PS based on an unknown median split, and graphs that detailed how these categories impacted a users time on task were illustrated (See Figure 4.2). Therefore, apart from the six studies mentioned in Table 4.1, in the remaining 84.6% of papers examined, it is not possible to know the PS scores. Consequently, this lack of reporting figures makes it difficult for other researchers to compare and assess the reliability of any results found, which ultimately reduces the academic rigour of many PS studies.

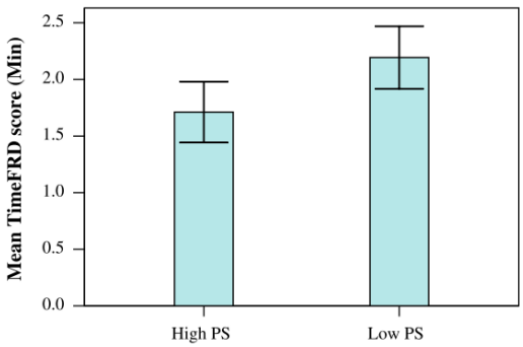


Figure 4.2: Example of how previous studies only report PS effects, and did not define what is high and low. Source: Al-Maskari & Sanderson [4].

Additionally, even in the studies that have reported PS figures, it is questionable whether the PS scores are truly valid. For instance, Ekstrom themselves originally stated that: "*It is strongly recommended that researchers use more than one of these tests in any exploratory endeavour that aims at identifying a factor*" [83]. Yet, many PS studies that have been discussed and claimed to find significant results only administered one of Ekstrom's tests [18,46,222,225,227].

Furthermore, even in the papers that did use more than one PS test, an explanation for how to merge scores from multiple tests is lacking. It is therefore unknown if test scores were weighted equally with an average of the two taken, or if precedence was given to one test over another and if so, which one? For example, although Allen [10] stated that two of Ekstrom's PS tests achieved a moderate Cronbach reliability rating, no explanation for how this was deduced was given. Instead Allen [10] claimed that the two PS tests were assessing different aspects of PS and thus analysed them as separate entities. Similarly, another study claimed that one of Ekstrom's PS tests was too similar to a different cognitive ability test, and therefore they excluded these from their analyses [13]. With so many unknowns with calculating an overall PS score, this reduces the consistency with which PS tests can be analysed throughout the literature.

Lastly, from all of the papers reviewed, only one mentioned that in order to enhance the reliability of their results, participants repeated the PS test approximately 5 days later after their initial test [86]. However, in the work presented in [86], no explanation was provided for how they then calculated the overall PS score. For example, did they take an average between the two separate sessions, or just randomly decide to report only the results from one? Regardless, it is surprising that more PS tests are not repeated across multiple sessions considering PS is thought to be relatively stable in individuals [13]. Because of the stable nature of PS, if a participant was not gaining a similar PS score on both sessions, then this would imply that the PS test was not truly measuring PS [74].

4.5.3 Unclear Marking Instructions

If a researcher wishes to administer a PS test in their study, then they must subsequently be able to analyse the test results correctly to compute a PS score. However, the original marking advice for each individual test lacks clarity and may lead to some confusions by participants completing the tests, but also researchers scoring the tests. Unfortunately, as Ekstrom's PS tests are over 40 years old, the original references that are discussed by Ekstrom are very old and inaccessible, potentially because they have not been digitised. It is therefore unknown how the tests were exactly developed, and which points in the test are the most important factors that need to be considered to deduce the overall PS score. For example, the *Number Comparison* test instructs participants to cross any pair of numbers that are not identical. Results are then calculated by the "*number marked correctly minus number marked incorrectly*" in a given time period [83]. As this test is meant to monitor how many pairs of numbers a participant can scan through in a set time period, yet participants are only indicating the numbers which are non-identical, it is unknown how many pairs of numbers they have successfully acknowledged as identical. Thus, the current advice for scoring this test does not fully correspond to the original instructions given.

Furthermore, the *Finding A's* test also encompasses issues. To reiterate, this test instructs participants to score any words that contain the letter 'a' in them and emphasises that each column has 5 words containing the letter 'a'. Participants are also told: "*Your score on this test will be the number of words marked correctly. Work as quickly as you can without sacrificing accuracy*" [82]. However, there is no explanation given as to how to score a participant's answers if each column is not completed. For example, if a participant was aware that they hadn't identified 5 words containing 'a' in a column, should they delay their time by continuing to repeat a visual scan of the same column or skip to the next column? This lack of understanding in instruction creates a huge gap in deducing overall PS. Participants are different, and their scanning abilities will undoubtedly vary. Thus, one person may get an accuracy score of say 15, but they only completed the first 3 columns thoroughly with no mistakes on page 1. Whereas another person may get the same accuracy score of 15, but rather than finding all correct answers that were immediately in front of them, they got this score from briefly scanning rows across 9 columns in 2 pages. With such opposing possibilities of results, it seems unusual that there is no explanation for how to score these differences, and what these results may mean for a person's true PS levels. If PS involves accuracy and scanning of what's visible [10], then surely there's a difference in PS levels depending on whether a participant can efficiently identify everything that's visible without making any mistakes, compared to finding some correct answers over multiple pages whilst simultaneously missing many others.

4.5.4 Different Formats

Similar to how there are unknowns in whether there is a difference in PS depending on whether systematic or random scanning of answers is employed, it is also unknown how PS tests were exactly formulated, again presumably due to the lack of accessible old references available. For example, Ekstrom's *Number Comparison* presents 24 rows of numbers in 2 columns [83]. Alternatively, another kind of PS test following the same principles, the 'Minnesota Clerical test' for PS, incorporates a number comparison test of 4 columns, each with 50 rows [127]. Yet, neither tests explain why these exact numbers were chosen. Thus, would a test for PS equally measure PS if there were 6 columns, as opposed to 3, visible at any one time?

Additionally, the same study [127] showed that some number comparison tests possibly contain confounding elements because the index of number change was never equally distributed. Likewise, we personally calculated the exact indexes for change in Ekstrom's tests, and found a couple of number pairs in the *Number Comparison* that had more than one difference in them. With no formal explanation as to how these numbers, indexes for change, and columns were formulated, it is unknown whether these are fundamental mistakes in the original design or whether or not these different variations matter for the validity of PS. However, as other research exists that demonstrates how visual perception changes depending on layout, it would make sense that the format of the PS test is important. For example, one eye-fixation can process 24 letters in a vertical position, compared to 12 letters in a horizontal position [143, 179]. Yet, Michalski and Grobelny [2015, cited in [130]] found that individuals better perceive horizontal layouts more than vertical ones.

Furthermore, although many studies stated that they used Ekstrom's PS tests, officially these tests require a licence to use [82], and yet none of the papers reviewed mentioned how, or even if, they obtained licensing. Therefore, this may suggest that researchers have instead used Ekstrom's PS tests as a guide to make their own test. Although this point is just speculation, if it is the case, then the exact format of how the PS test was visibly administered in many tests is unknown. This again makes the comparability of PS studies challenging.

4.5.5 Limited Linguistic Reasoning

In the *Finding A's* PS test, each page contains 5 columns, with 41 words per column, and thus 205 words per page [83]. With 4 pages per part, that totals 820 words. As there are 2 equivalent parts, this means there were 1640 words overall, out of which 200 (or 12.195%) contained the letter 'a'. After an analysis of the words used, we observed that the number of letters in all the words appeared to be quite equal, ranging from equivalent words that have 4,5,6,7, and 8 letters in length, and that there was a fairly equal balance between 1 or 2 syllables used. It was also noted that some of the words were repeated in Ekstrom's *Finding A's*.

Yet despite such a large set of word stimuli, there is no explanation for how the words were selected for the test. Additionally, it is not documented whether the words: contain the same frequency in the English language; elicit similar sentiment; were positioned in any particular order; are processed differently by a Native English speaker; and other factors that are important in linguistics such as the distribution of nouns and verbs [141, 175, 194]. These points do not necessarily reduce the reliability and validity of the PS tests to date, as they have been successfully used over many years to elicit significant results. However, it is worth being aware of these factors for any future researcher that may wish to expand upon and develop their own PS test to ensure that the stimuli chosen contain the same components that produce valid results for PS. For example, studies in Neuropsychology have long recognised that emotional words are perceived stronger than non-emotional words [213]. Thus, if all the words that contained the letter 'a' were more common or emotionally sentimental in the English language compared to the words that didn't contain a letter 'a', then perhaps individuals would automatically identify them, regardless of their PS levels.

4.5.6 Outdated Administration and Content

If a researcher wanted to administer Ekstrom's PS tests, then a licence must first be sought from ETS Research [82], who then distribute PDF copies of the specific tests requested. Yet, the administration of the tests remains the same as 40 years ago when they were first devised: a paper-pen version, which ultimately requires manual scoring. In fact, psychological researchers have described how scoring PS tests takes longer than the participant completing the actual test, described in [1] as: "*the scoring process turns into somewhat of a PS test for the individual scorer, as he or she attempts to count correct, missed, and incorrect responses using a template to match to the examinee's responses*". However, with many experiments now run online, it makes it impractical to use paper based surveys. This motivates the question, how do we computerise the PS tests such that they are reliable and valid instruments?

Furthermore, caution must be taken when using a cognitive test that dates back so many decades because over time, attention evolves. For example individuals now "*have to fight to stay focused on long pieces of writing*" as a result of information technology [57]. Likewise, a recent study by Mark et al. [156] reaffirmed that individuals currently have a limited capacity for attentional resources, and that this is not helped by current information workers experiencing increasing levels of distractions. In relation to PS, attention is fundamental to cognition, and PS is a type of cognitive ability. Keeping this in mind, the authors of this paper conducted a pilot study of the *Finding A's* test, and discovered that it took over 10 minutes for some participants to complete the test. Yet, the original *Finding A's* was meant to only take 2 minutes to complete 4 pages with 820 words. Therefore, it is worthwhile making new researchers aware of these differences, to ensure that the current PS tests contain the right amount of stimuli and time necessary for current states of individual attention.

4.6 Discussion

Although PS testing has been used in many IIR studies, as a result of analysing the measurement tools used in these studies, many uncertainties have been described that provide interesting debate for current researchers to consider. Consequently, the above themes identified have provoked challenges and recommendations for future administration and analysis of PS tests.

4.6.1 Challenges

The key challenges researchers face with PS testing appear to concern the content and administration.

Regarding administration mainly from the above theme of *Outdated Administration and Content*, an obvious next step for furthering PS testing may seem to be converting the old paper-pen format into a modern, computerised test. This would then theoretically resolve the problems identified of being old-fashioned and difficult for researchers to score and analyse, which may have even put some researchers off from considering using PS tests. Consequently, if the PS test was administered online, then it might be easier for researchers to integrate into their studies where the main part is already administered online, and thus the hassle of switching between paper and computer would be eliminated. With a more effortless form of administration and automatic scoring from a computer, more researchers might be encouraged to involve PS testing into their research, which would in turn increase the reliability of results if more studies were able to be compared. Although these points are just hypothetical, other researchers have stipulated the benefits of computerising PS tests with the main reason being that software could be dynamically used to adapt the screen to counter the negative effects for Low-PS users [68].

However, it is not as simple as taking the same paper PS tests and converting them to an online format for many reasons. Firstly, there's a difference between how stimuli are perceived depending on whether they are viewed on paper or a computer. For example, completing 41 words on a column on A4 paper may differ to how many words you can physically see at once in a column on a different sized computer screen. This difference was reaffirmed by [207] who compared participant's responses to a PS test conducted on paper, a video display terminal (VDT), and a combination of switching between both. The exact PS test used was not one of Ekstrom's, but similarly involved 200 number comparisons taken from the Minnesota Clerical Test. As can be seen in Table 4.2, people score a lot less when conducting the tests online compared to paper. Therefore, more research is needed that explains these differences, in order to develop an online PS test that is truly measuring PS.

Secondly, a few studies have attempted computerised PS tests, but with no explanation as to what measures were taken to account for the above problems surrounding converting PS tests from paper to online. For example, Zimprich & Kurtz [248] took 60 numbers from

Table 4.2: PS differences depending on administration type in Silver & Bennett [207]

Format of Test	Mean	Standard Deviation
Paper	119.29	32.42
Online	85.24	21.15
Both	67.32	15.98

Ekstrom's *Number Comparison* and administered it online within a 90 second time period. This is in comparison to Ekstrom's original 96 number comparisons over three minutes [83]. Yet, Zimprich & Kurtz [248] provided no explanation for: why only 60/90 items were taken; how they chose those particular 60 items over the remaining 30 that were not picked; why the time limit was halved; or how the content was visually divided and presented on a screen in columns or rows. Similarly, Silver & Bennett [207] and Fisk & Warr [86] attempted computerised PS tests, but again, no justification for their content or explanation for how they were presented was given.

Additionally, from the literature reviewed on computerised PS tests, many other factors were also not discussed that may influence the validity of PS tests. These include: how participants physically select the answers on a screen such as whether selected items are scored out or change colour; whether all stimuli are presented in individual boxes, grid-lines, or blank backgrounds; if words/numbers are aligned to the left, middle, or right of the screen; what font is used; and what is the inter-letter spacing or spacing between items. This list is not exhaustive, and of course it may be that these factors are incidental in affecting a PS score. However, although not specifically examining PS, other psychological research has identified that inter-letter spacing is a perceptual factor that modulates visual word recognition performance: decreased spacing resulted in slower identification thereby confirming the interference between close proximity of stimuli and visual perception [170]. Thus, if inter-letting spacing affects perception in reading, it may also affect how PS tests are designed. Consequently, the above list of factors described may affect PS online test validity. Yet with so many variables apparent, much more research is clearly needed that investigates and accounts for these components before a precise and valid PS test can be assured.

If time was invested into developing a new computerised PS test, then it would appear worthwhile for researchers to consider, and account for, some of the other themes that this paper identified regarding the content of current PS tests. Namely, the different formats, linguistic reasoning, and attentional structure all ignite discussion for researchers to consider.

As one of the themes in this present paper identified that there is variation between different PS tests concerning the format of stimuli, such that Ekstrom's *Number Comparison* presented stimuli in 2 columns of 24 rows [83] while the Minnesota Number Comparison presented 4 columns of 50 rows [127], it is unknown what the optimal layout for PS tests should be. Moreover, details about how the original PS tests were developed have never been specified, causing

unresolved questions as to whether different formats of visual presentation were even tested on people to gauge any possible differences in PS response. With other research having identified that visual perception is influenced by horizontal/vertical layouts [130, 143, 179], and calls for computerising PS tests have determined the need for reconsideration of PS test displays [1, 207], further PS development is needed. Experiments should manipulate multiple different ways at physically viewing the PS stimuli such as different variations of columns and rows. Although time consuming to design and test, these manipulations are necessary to ensure that any new computerised PS tests are still valid and effectively measuring PS.

Before research can consider the layout of stimuli, the correct kind of stimuli that will equally elicit valid PS results must first be deduced. Our theme of 'limited linguistic reasoning' discussed how the meaning and structure behind the stimuli chosen for the PS test that contained words was unknown. Therefore, further research is required to make sure there are no confounding variables, such as certain words containing too highly emotional meanings and thus making perception easier [141, 175, 194], negatively influencing PS results. Thus, when selecting word stimuli, new researchers may wish to make use of databases such as The English Lexicon Project [26], where words can be chosen, filtered and equalled for specific lexical characteristics.

Furthermore, as PS tests are effectively measuring how accurate and fast an individual is at identifying some kind of perceptual change [83], such as a word that contains an 'a' or a number that doesn't equate with it's pair, more investigation is required as to where the index of change is positioned, and how many changes there are. For example, in the *Finding A's* PS test, there are 41 rows of words where 5 contain a letter 'a'. Firstly, questions to consider include whether it is necessary that there are always exactly 5 changes to be identified, as opposed to another specified or random number. Secondly, the spacing between target answers requires deliberation. For instance, does it matter how close together the words containing 'a's are? Are they all clustered together in the centre of the column, equally distributed throughout, or randomly dispersed such that some end up close together while other columns are sparse? Again, these questions aim to stir discussion with researchers who wish to develop new PS tests to ensure the structure is still reliably measuring PS.

Finally, this current paper identified a main theme which involved the current PS tests being outdated. Beyond the outdated administration of paper/pen formatting, the notion of human attention changing over 50 years was discussed. As PS is a type of cognitive ability, and attention is a key component of cognition, it is crucial that future PS tests do not overload people's limited attentional capacities. Thus, perhaps new PS tests may need to be shorter, contain fewer overall stimuli, or the length of time to complete the test should be extended. Before these revisions can be achieved, all components of the PS test that may affect attention need re-examination. This is essential to ensure future PS tests are still validly measuring PS, whilst simultaneously accounting for the fact that attention may have evolved over time. Lastly,

reconsidering attention limits of participants is necessary to ensure that they are not overtired as a result of PS testing, as this may adversely confound any results found in subsequent tests they complete in the main studies.

4.6.2 Recommendations

In the *challenges* section of this review, many areas requiring a lot of further research have been discussed. Yet practically speaking, it would take a considerable amount of time before any of the results from this research could be implemented in future PS tests, where reliability and validity of PS is still guaranteed. Nonetheless, although we have explored the need for PS tests to be revised and computerised, the original paper/pen format has still proven to be useful, with many studies finding significant results. However, the themes identified in this current paper have ignited some recommendations for currently available PS tests that researchers might wish to follow in order to improve their administration and overall reliability and validity of results found.

Firstly, the theme of *unclear marking instructions* identified that the *Number Comparison* PS test score is calculated as a result of the items participants marked correctly *and* incorrectly. Yet, participants are only instructed to cross out non-identical number pairs, which leaves it unknown how many identical pairs they have correctly scanned through. Thus, a perhaps better way of administering this PS test would be for participants to 'tick' for same, and 'cross' for different pairs of numbers. That way, the researcher would be able to exactly quantify how many pairs a participant is efficiently scanning through. As PS concerns an individual's accuracy and speed to view, scan, and compare information during visual search tasks [10], we would hypothesise that having a more robust way of quantifying how many items a participant is processing would return a more valid measure of PS.

Another recommendation that we propose would increase the PS test validity regards how many times the PS test is administered on the same participants. Realistically, it may be difficult to recruit participants on multiple occasions. However, as PS is meant to be a stable cognitive ability [183], if a participant wasn't getting a similar score on the same test at different times, this would reduce the validity of results [74]. Thus, if a researcher wanted to reaffirm that the PS test they were administering was truly measuring PS, we would advise taking a small sample of participants and administering the PS on two separate occasions to ensure similar results were being obtained.

Additionally, another theme established a breach in PS validity as many studies claiming to have assessed PS only used one PS test, which contradicts original guidelines that stated more than one PS test was required to fully identify a cognitive ability [83]. Accordingly, we encourage future researchers to avoid this problem by always administering at least two PS tests. Unfortunately, there are no explicit guidelines on how to merge multiple test results together. However, Allen [9] utilised Cronbach reliability testing between 2 of Ekstrom's tests.

This measures the internal consistency, otherwise known as how closely related a set of items are as a group [215]. Thus, we also encourage researchers who use more than one PS test to run reliability analyses between their PS tests to increase the reliability of their overall PS measure.

The theme of *inconsistent reporting of results* identified a consistent trend which involved how many previous IIR studies involving PS failed to report the exact results or distribution of the PS tests used. Therefore, we strongly recommend that all researchers should avoid this unknown and instead report as many exact results as possible such as: the median; mean; standard deviation; and a graph that contained all possible PS scores with how many participants achieved each score. If unknown figures were made to be known and explicit in future PS tests, then we would predict that this would improve the reliability of results obtained and make it easier for other researchers to compare their studies to. Additionally, if there was then a large sample of multiple studies who had used and reported their PS scores, an analysis would be possible that could compute average standardised thresholds. Having an exact threshold for what was considered 'low' and 'high' would then benefit future studies to ensure a consistency in results, regardless of the sample of participants used.

Finally, beyond unknown exact figures, there are other factors that many studies failed to report, which if they had, would have increased the robustness of results obtained. For instance, in Arguello & Choi [18], they stated that the *Finding A's* PS test incorporated a possible range of 0-200. Yet, no units were given for these figures or explanation for what those figures exactly meant. It is therefore unknown where these numbers came from which leads to new researchers being left unable to compare these figures into their own work. Furthermore, many studies never expressed the format of the PS test used: although they quoted that the PS test originated from Ekstrom, which as we know is paper-based, only few studies explicitly state whether the administration of the test was done on paper. Hence, there is no guarantee that other studies have all administered their PS test in paper/pen format, and perhaps instead taken the Ekstrom stimuli as a guide and computerised it. If this was the case, then this would interfere with the comparability of PS testing between studies. Consequently, researchers of future PS tests should ensure that all details and aspects of their PS test are always reported, to allow for easier reviewing and reliability assessments of the overall test usage to be made by others.

4.7 Chapter 4 Summary

Overall, although PS has been known to affect IIR, the current chapter has identified many areas that could be improved upon to make the PS tests more reliable and valid. Regarding the **content** of the tests, more understanding is needed for: the linguistic structure of words used as stimuli; where changes are positioned, and how many there are; a reconsideration of current human attention and how this may affect how many stimuli are visible; and a further exploration

for how the format of stimuli should be visually presented. Concerning **analysis** of the tests, further research is needed for clearer marking instructions and setting standardised thresholds. Additionally, the **administration** of PS tests needs some refinement such that they: should be computerised; more than one PS test should be administered; and the same tests should be completed by the same participant on two separate occasions. However, challenges were noted that explained the difficulty of converting a paper test into an online format with appropriate stimuli. Finally, to increase the comparability of PS studies, researchers should follow certain conventions for **reporting results**. Specifically, actual figures and specific details about their test should be reported so that transparency is increased and comparisons between different research samples is enabled.

All of these recommendations and challenges summed together provide many avenues and questions for future research. Consequently, it is hoped that this chapter has made researchers aware of the limitations in order to stir discussion and ignite debate to advance future PS test usage. However, as it is now clear that PS tests have a lot of refinement and further research needed—and the reliability and validity of previous measurements has been questioned—this led onto another important question to consider: if previous measurement of PS has lacked reliability and validity, then can the results of studies that have incorporated PS be trusted? The answer to this question was explored in the subsequent Chapter 5 through a Systematic Review of what claims have occurred regarding PS, and an analysis of how these results could be explained.

Chapter 5

Systematic Review

5.1 Chapter 5 Overview

This Chapter reinforces the motivation for conducting a Systematic Review into the area of PS in Computer Science. This is followed by the details of how this method was undertaken, including the review protocol followed. The results of the Systematic Review are then provided and split into answering two research questions: "*What claims have occurred regarding PS in Computer Science?*" and "*How can the results be explained?*".

5.2 Introduction

Given that the previous chapter found many problems in the use of PS testing, this identified the need to conduct a larger review of the literature, to ensure that the general concept of PS was valid and reliable: if the PS *tests* lacked validity and reliability, then this raised the question as to whether the concept as a whole could have been investigated poorly.

However, when beginning to analyse previous literature more thoroughly, two papers in particular further exemplified the need for a greater, more systematic, examination of PS. Firstly, a recent peer-reviewed journal article, published in 2020 by Conati et al. [67], aimed to predict PS using eye-tracking. It was argued that if PS could be predicted, then this could drive adaptive systems to accommodate a user's individual differences. This motivation was driven solely from one cited previous study, where lower levels of PS were linked to lower task performance. However, when the original source that claimed Low-PS lowered task performance was checked, specifically the reference which referred to Toker et al. [221], this paper did not actually mention anything about Low-PS lowering task performance. It therefore remained unknown as to whether PS did, or did not, affect task performance.

Secondly, another peer-reviewed publication investigated how different user characteristics, including PS, influenced the effectiveness of different visualisations [206]. Yet despite thoroughly detailing how PS was measured, including the range of participant scores collected, PS was never referred to again after the methodology section. Consequently, no results were given

or discussed in relation to PS and task performance, and instead it completely disappeared from the remaining paper. This was not unique, as other studies also followed this pattern, where PS was supposedly investigated, but then no results or discussion ever materialised (*such as* [137,219]). Therefore, in addition to the concept of PS potentially being investigated poorly, it was also speculated as to whether the concept in itself was still worthwhile to investigate.

Consequently, as peer-reviewed studies in reputable venues have been based upon studies that either do not claim what the citation states, or PS results have completely disappeared from the paper, it was deemed a paramount time to systematize knowledge about PS in a rigorous way. The community that investigates PS will then be able to clearly identify what claims have been made about how PS affects a user, whilst also being mindful of the reliability and validity of such claims.

Although many other publications involving PS contain a literature review, the field is still lacking a Systematic Review. Unlike a traditional literature review, Systematic Reviews follow a more thorough, transparent, explicit, and reproducible methodology with a predefined review protocol. This approach minimises result bias and enables more reliable findings to be found [217]. With the intention of filling this gap, a Systematic Review that evaluated how PS has been used within the field of Computer Science was conducted. For the purpose of this review, the domain of IIR was broadened to Computer Science more generally for two reasons. Firstly, it aimed to provide a thorough understanding of where problems in citations could have arisen from, and thus narrowing down the scope to just IIR may have not identified this. Then secondly, as was identified by previous research and described in the literature review of this thesis (See Chapter 2, Section 2.2), an interdisciplinary approach to IIR is needed, which incorporates Human-Computer Interaction and Computer Science more generally [78].

In essence, the major contributions of this work are: (1) being the first Systematic Review in the PS discipline; (2) providing a comprehensive classification of what claims exist regarding PS in Computer Science; and (3) critically evaluating possible effects of PS.

5.3 Method

As explained in Chapter 3 (Section 3.3), a Systematic Review was selected as the research method, as opposed to another method such as a Scoping Review, through the definition from a recent journal article cited by almost 3,000 researchers, which aligned with the current research aims: "*A systematic review may be undertaken to confirm or refute whether or not current practice is based on relevant evidence, to establish the quality of that evidence, and to address any uncertainty or variation in practice that may be occurring*" [173]. Additionally, Systematic Reviews have been widely used in a variety of similar Computer Science domains, such as Social Network Analysis Tools [20], Imbalanced Data Challenges [118], Human-Computer Interaction studies that also aim to design systems that are intuitive of users' behaviour [229], and analysis

of statistical significance, power, and sample sizes across papers published in SIGIR, a world-leading conference in Information Retrieval [199].

Instead of creating new data through primary research, Systematic Reviews are a type of secondary research, where analysis is undertaken of primary data that has already been collected and completed [190]. In order to provide a rigorous method, a practical guide to completing a Systematic Review in Computer Science was followed (*See Weidt-Neiva & Silva, 2016* [232]). This approach has been used in other Systematic Reviews (*e.g. Thilakaratne et al.* [217]) as it details a well-defined procedure of various steps that must be completed to ensure reliable and reproducible results are obtained. This approach was also combined with Kelly & Sugimoto's [123] description of Systematic Reviews, as recommended by other Systematic Reviews in Computer Science (*e.g. Sakai* [199])– researchers must: plan how studies will be gathered; adhere to strict scientific guidelines; exhaustively examine the retrieved literature; and conduct analysis with a neutral position and open mindset to minimise potential selection and interpretation biases to ensure reproducibility, and hence reliability.

5.3.1 The Main Research Questions

The first stage in conducting a Systematic Review in Computer Science, as identified in the practical guide by Weidt-Neiva & Silva [232], involved identifying the questions that guide what is wanted to be understood during the research. The aim of this paper was to better understand how the concept of PS had generally been studied and applied in Computer Science, whilst critically evaluating any effects of PS. The present review therefore explored two main research questions: ***1) What claims have occurred regarding PS in Computer Science?***; and ***2) How can the results be explained?***

5.3.2 Selecting Databases

To gather relevant research, The Association for Computing Machinery (ACM) Digital Library was firstly searched, given that this provided all PDFs for another Systematic Review in Computer Science (See [199]). In this database, The ACM Guide to Computing Literature was selected, which comprised of 3,144,110 records, as opposed to the ACM Full-Text Collection, which contained just 661,685 records.

Furthermore, other Systematic Reviews in Computer Science tended to utilise between 4-5 databases, in order to improve the coverage of retrieved papers [20, 118, 229]. Consequently, the recurring databases in these studies were also selected for the current review, including: Science Direct, Scopus, Springer Link, and Web of Science. Although IEE Xplore was also used in the previous studies, after trialling the search terms for the present review there, it appeared that the results were more applicable to the field of Neuroscience and Engineering, such as "*Functional Transcranial Doppler Ultrasound for Measurement of Hemispheric Lateralization*

During Visual Memory and Visual Search Cognitive Tasks". Consequently, IEE Xplore was not selected as an appropriate database.

5.3.3 Defining keywords and search strings

The search term used was *"perceptual speed"*, in inverted commas. This was to ensure that papers that only mentioned 'perceptual' or 'speed' separately were not influencing results found.

Other combinations of search strings were also explored, including phrases which included specific domains of Computer Science such as: *"perceptual speed" AND "information retrieval" OR "perceptual speed" AND "information seeking"*. However, these phrases were considered too restrictive, as there are many areas that PS has been investigated in, and this review sought to provide a general understanding of how PS had been used in the field of Computer Science.

5.3.4 Defining Search Strategy

All databases were searched on two occasions: firstly, in May 2021, and secondly, the search was repeated in December 2021. Table 5.1 presents the raw number of results returned, as a result of the search string *"perceptual speed"*, on both occasions, for each database.

Table 5.1: The number of results returned from each database for each search string and date of search.

Search String and Date	**ACM**	**Science Direct**	**Scopus**	**Springer Link**	**Web of Science**	**Total**
"perceptual speed" (May 2021)	187	2,063	2,755	1,369	699	*7,073*
"perceptual speed" [in Computer Science] (May 2021)	187	51	112	85	18	453
"perceptual speed" (Dec 2021)	195	2,106	2,847	1,448	715	*7,311*

Although the ACM library was naturally searching papers with computing relevance, the other databases covered a large variety of subjects. In each database, the search results indicated which subjects had the highest presence of PS. For example, in Science Direct in May 2021, the largest proportion of papers (1,013) were in the field of *Psychology*, followed by *Neuroscience* (816 papers), and *Medicine and Dentistry* (578 papers). This was a similar pattern in the other databases: Scopus's largest amount of papers being retrieved were in the field of *Psychology* (1,362), *Medicine* (1,127), and *Neuroscience* (652); Springer Link's top three most retrieved papers were also *Psychology* (494), *Medicine and Public Health* (296) and *Biomedicine* (161); and Web of Science's largest number of papers were in *Neurosciences* (153), *Psychology Multidisciplinary* (103) and *Gerontology* (99).

Given that many of these subjects were not relevant to the present review, the search results were filtered by subject to only include those relevant to Computer Science. Since every

database had a unique indexing system, the subject filters had to be slightly adapted. Science Direct, Scopus, and Springer Link were all filtered to only include papers in the subject of "*Computer Science*". However, Web of Science did not have a discreet category of Computer Science. Instead, every possible subject was manually examined for relevance and then the following were selected as inclusion filters: *Computer Science Artificial Intelligence, Computer Science Theory Methods, Computer Science Interdisciplinary Applications, Computer Science Cybernetics, Computer Science Information Systems*, and *Information Science Library Science.* Table 5.1 also exhibits how many papers were then returned, after subjects had been selected.

Despite the search being repeated in December 2021, the increase in papers being returned from May to December were not in the field of Computer Science. For example, whilst there were 85 Computer Science papers returned in Springer Link in May 2021, in December 2021, it was 87. Despite this increasing by 2 papers, on further inspection, they appeared to be categorised incorrectly, and were in fact from the field of Engineering Psychology. Similarly, the few additional papers retrieved in the ACM database were also not deemed relevant, as they instead appeared to be more related to medicine or Psychology, such as '*Automatic Speech Classifier for Mild Cognitive Impairment and Early Dementia*'. Consequently, the search in December 2021 did not retrieve any more papers that were included in the final sample for the present review.

5.3.5 Defining Inclusion and Exclusion Criteria

With Systematic Reviews, it is common to only include papers from a recent time period, such as the previous ten years. However, for the purpose of this review, it was decided best not to disregard any date, given the initial problem identified that papers cited other papers that did not claim what they were supposed to claim. Consequently, there may have been a vicious cycle of misreporting, and thus going back to the original sources, no matter how old, was necessary to understand PS as a whole. However, it is worth noting that the databases used will still contain implicit time limits, such as the ACM DL, where the earliest paper on record is dated from 1951.

All articles from each database that had been returned as a result of the search *"perceptual speed"*, and filtered by *Computer Science*—which totalled 453 articles—were exported into the reference manager, Mendeley Desktop, and after duplicates were removed, this left 313 unique papers. Unfortunately, 6 papers were unavailable for full-text download, even after searching many other websites. This left **307 papers** that were downloaded in full.

Although every paper downloaded was supposed to be in the domain of Computer Science, it appeared that some papers were not. For example, papers titled '*Personality, Working Memory Capacity and Expert Manual Annotation of German Spontaneous Speech*', '*Aging and Financial Decision Making*', and '*Improving selection for psychomotor skills in dentistry*' were

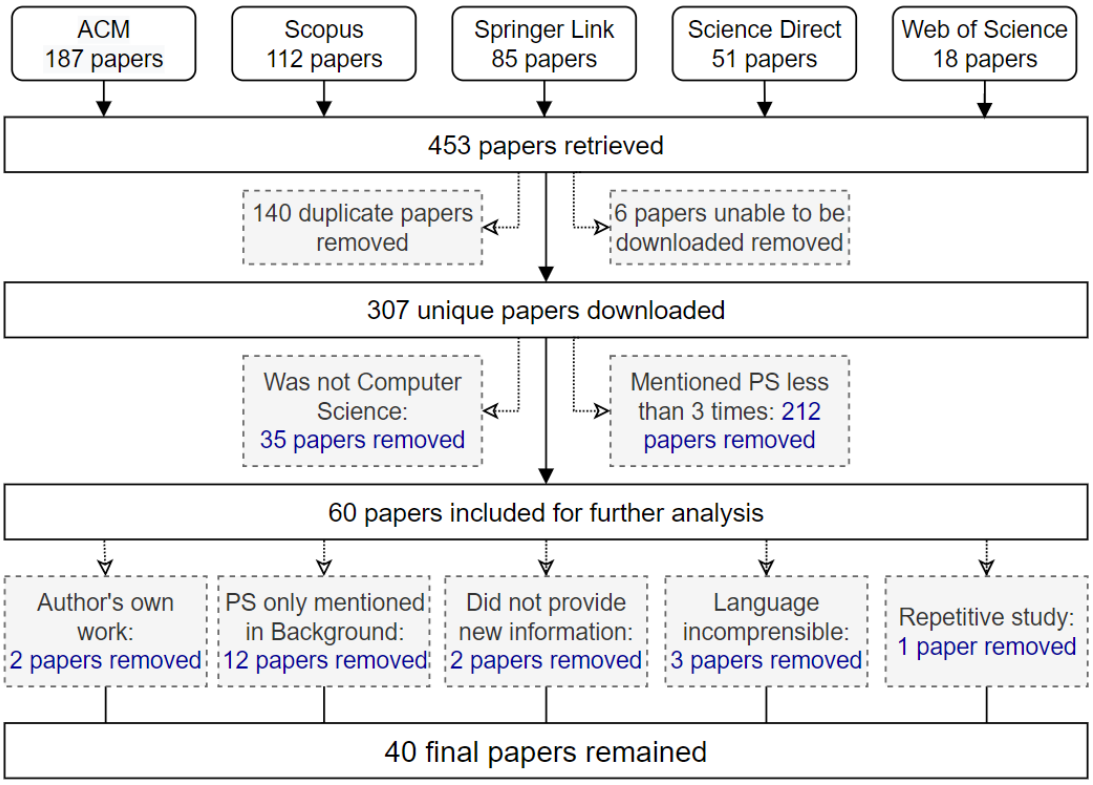


Figure 5.1: The Systematic Review inclusion and exclusion criteria, where PS refers to Perceptual Speed.

not considered topical to Computer Science. Therefore, if the paper did not appear to relate to any aspect of Computer Science, then it was not included in further analysis.

Furthermore, for a large percentage of papers retrieved, there was only a passing reference to PS, or it instead only appeared in the full text reference section. As a brief mention of PS does not help further understand the concept of PS as a whole, only papers that had more than two mentions of PS were included for further screening. This had to be manually checked, as often a paper referred to PS, and then abbreviated it to just "PS" for the remaining paper. Alternatively, the paper continued to discuss it under the name of the PS test used, such as "Identical Pictures" or "IP". Consequently, any PS test used also had to be checked in each paper, and then variations of the different terms were searched throughout the remaining paper.

In order for the paper to be further analysed, a number of exclusion criteria were then implemented (*See Figure* 5.1).

Firstly, two papers were excluded from the search as they were from the current author's own work. This was to ensure that the Systematic Review was unbiased, as outlined in Kelly & Sugimoto's [123] recommendation for Systematic Reviews to minimise interpretation biases.

Secondly, if PS was mentioned many times—except only in the background of the paper—then this was removed. This was to ensure that the main focus of the paper included PS as primary research.

Thirdly, to further emphasise the focus on primary research, despite PS being mentioned throughout the whole article, a further two papers were also excluded for not providing new information, but instead: a) acting as a literature review [151]; and b) providing a summary of multiple papers by the same author [65].

Fourthly, a further exclusion criterion that was implemented involved how comprehensible the language in the paper was. It was decided not to restrict results to only peer-reviewed articles, to identify where the flaws in previous citations may have arisen from. However, whilst every paper was written in English, after examination, three papers (two of which were from the same conference) did not make grammatical sense to read, and these were thus excluded for lacking clear explanations.

Of the remaining 41 papers, it appeared that some were using data from the same user study. Specifically, 19 papers had been published using data from 7 user studies (See Table 5.2). To minimise repetitive results, these papers were manually examined for repetition. Despite having multiple publications from the same study, the different papers still appeared to be unique: they explored different aspects in each paper. For example, one paper focused on analyzing how eye fixations differed between users with Low-PS and High-PS across different visualisations [222], whilst another paper from the same study provided details on differences in length of search task completion time between Low-PS and High-PS [221]. The only exception to this, was a paper that was initially published as an 8 page conference paper [70], but was then extended as a 26 page journal article [71]. Consequently, the initial smaller conference paper was removed, so that results were not duplicated.

Table 5.2: Identifying which articles appeared to share an identical user study that generated the data, ordered from earliest published, to most recent.

User Study, described by the Interface investigated	**Reference of paper**
Word map versus Multiwindow visualisation.	[12, 13]
Bar chart versus Radar graph visualisation.	[209, 210, 211, 221, 222]
Visual prompts for bar graph visualisations.	[56, 94, 218]
Vertical and Horizontal Value Charts.	[64, 140]
Various websites accessed through Google.	[70, 71]
MetroQuest visualisation map.	[66, 67, 139]
Magazine visualisation.	[219, 220]

All criteria discussed in this section were defined and verified by multiple researchers to ensure they aligned with the overall research aim. After all of the exclusion criteria had been implemented, 40 papers remained. The complete list of both included and excluded studies, categorised with the reason for exclusion, has been provided in Appendix A.A, as recommended by Puljak & Sapunar [190].

5.3.6 Extraction of Answers

To answer the main research questions, the 40 papers were read in full. A spreadsheet was created where each row corresponded to a unique paper. The first 6 columns included referencing information: an identifier, title, authors, publishing date, publishing location, and DOI. The remaining columns of the spreadsheet were then created to extract information from each paper that was thought to be directly relevant to the research questions.

To answer Research Question 1 of **'What claims have occurred regarding PS in Computer Science?'**, there were columns for different dependent variables: two for search ***Performance*** (referring to whether a measure of overall search *success* had been reported, and how much a user had *learned*); three for subjective ***Experience*** (namely, how PS affected *workload*, *belief of search success*, and system *ease of use*); three for search ***Time*** (which corresponded to the time it took for a user to *view* something, *click* on something, and their *overall* search completion time); three for search ***Behaviour*** (including information about *querying*, *views*, and *clicks*); and one for ***Physiology*** (which concerned whether PS had impacted features such as eye gaze, fixations, or saccades). Additionally, an ***Interactions*** column was created, where instead of an overall effect of PS having been reported, any interactions of PS with another variable, such as Task Difficulty, were noted. Furthermore, another column was created which contained a description of the ***Interface*** used in each study.

All of the columns were created through an iterative process. For example, at first, only one column referred to Behaviour, but as more papers were read, it became apparent that Behaviour could be sub-divided into three different types of Behaviour, and thus every paper was re-categorised accordingly. For each cell in the spreadsheet, the information was colour-coded as to whether the finding had been reported as significant (green), non-significant (red), or unknown (blue). There was an additional column to indicate if *no results* for any variable had been presented in a paper.

For Research Question 2, **'How can the results be explained?'**, the columns sought to gather information relevant to how the concept of PS had been experimented with. This included: whether a *theoretical framework* had been followed, and if so, which one; the type of search *task(s)* administered; what PS *definition* had been used; the *source* of the definition; the PS *test(s)* administered; the *format* of administration; any PS *score(s)* reported; PS *thresholds* for how Low-PS and High-PS had been calculated; the *sample* of participants used; and what *analysis* method had been undertaken.

Given that every column created for Research Question 2 was in parallel with the columns created for Research Question 1, this allowed comparisons to occur, such as whether more or less significant results in a certain variable, such as *performance*, occurred when a different factor, such as the type of PS test administered, had been used (E.g. comparing *Finding A's* against *Number Comparison*).

Having directly extracted answers from each paper and onto the spreadsheet, a thorough categorisation for answering the two main research questions had occurred. From this, visible patterns instantly emerged where differences between papers could be seen, or gaps in the literature appeared, such as papers that did not report a PS definition, and consequently had blank cells in the spreadsheet. Consequently, the following results sections describe the spreadsheet in detail, both through categorisation of all effects found, as well as identifying possible themes that could have affected the overall results.

5.4 Previous Claims of Perceptual Speed in Computer Science

As identified in Section 5.2, there was some confusion as to whether PS did, or did not affect users in the domain of Computer Science. Therefore, before the concept as a whole could be evaluated or assessed for result reliability and validity, it was necessary to answer the first research question of, **'What claims have occurred regarding PS in Computer Science?'**.

To answer this main research question, it was identified which papers, in the sample of 40, provided some kind of effect of PS. This involved removing any papers which: a) did not consider PS as an independent variable, but instead used machine learning to see whether PS could be automatically predicted using pre-existing data (9 *papers*); and b) provided no results (6 *papers*).

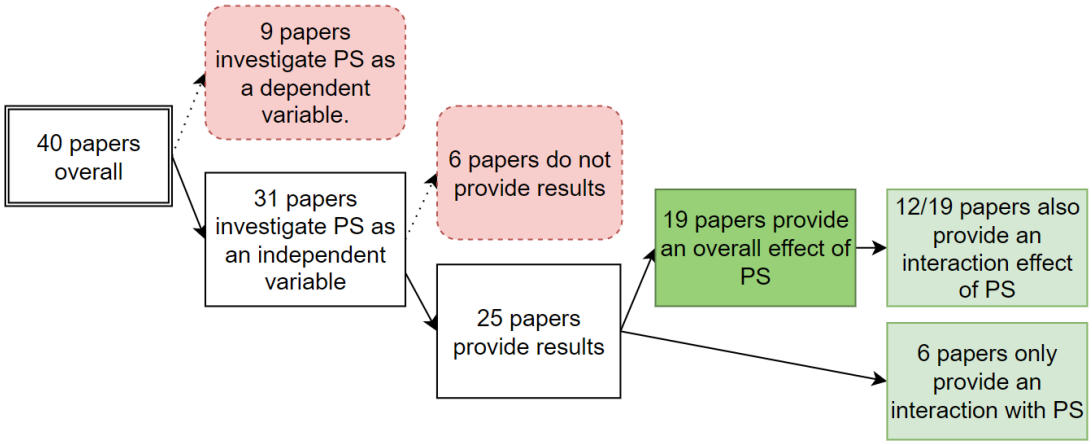


Figure 5.2: A flowchart that represents how many papers, out of the sample of 40, provided a direct effect, or interaction, with Perceptual Speed (PS).

Out of the 25 papers which provided results (please refer to Figure 5.2), these were divided into ones that claimed an *overall* effect of PS (19 papers), and others that only reported *interactions* with other variables (6 papers). For example, in Conati & Maclaren [68], search performance was investigated, and yet this was not considered as a paper that investigated search performance *overall* - only interactions were reported, such as finding that High-PS users were more accurate using colored box visualisations, whereas Low-PS were more accurate

with radar graphs. It remained unknown whether overall Low-PS, or High-PS users achieved better search accuracy. Therefore, (1) overall effects, and (2) interactions, were considered in separate subsections below.

5.4.1 Overall effects of Perceptual Speed

In order to examine what claims regarding PS had been found in the literature overall, an overview was created which categorised how many papers had investigated different dependent variables (*Search Performance, User Experience, Search Time, Search Behaviour, and Physiology*), and of these, which papers had returned significant results (*See Table* 5.3) [1].

Table 5.3: The number of papers that investigated different dependent variables, and whether results were significant (sig) or not.

Variable Analysed	**Number of papers**	**Sig**	**Sig but no direction given.**	**Both sig and non-sig**	**Trend but sig unknown.**	**Unknown**	**Non-sig**
Performance	**11**	4	2	0	1	0	4
Experience	**10**	3	0	1	0	2	4
Time	**9**	5	0	0	0	0	4
Behaviour	**6**	1	1	2	0	1	1
Physiology	**2**	2	0	0	0	0	0
Total	**38**	**15**	**3**	**3**	**1**	**3**	**13**

As can be seen from Table 5.3, there were a few occasions of results being unknown. This was despite the fact that papers that did not provide a concrete effect of PS were removed prior to this analysis. However, as some papers investigated multiple dependent variables in the same article, for a few of these, no results were provided, and instead the results focused on another dependent variable. These papers were kept in the sample for further analysis, but this explains why some '*Unknown*' results continued to occur. Nonetheless, from Table 5.3, it was apparent that the most papers investigated a measure of search ***Performance*** (11 papers), and the least concerned ***Physiology*** (2 papers).

Additionally, the results could be broken down further: whilst only six papers may have investigated some form of search ***Behaviour***, many more measures of behaviour were analysed and reported in each paper. For example, in the same article, Naghib et al. [174] measured: 1) *query length*; 2) *reformulation of the search question*; 3) *number of webpages observed*; and 4) *number of links observed*. Thus, Naghib examined 4 separate measures of behaviour, and of these, some may have been significant, whilst others may have returned non-significant differences with PS. Consequently, the number of overall measures that were examined, and of these which were significant, unknown, and non-significant were also reported in Table 5.4. This showed a slightly different pattern of results, where measures of user ***Experience*** provided the most data, instead of search ***Performance***.

[1] Please note, the total number of papers in Table 5.3, (38), is larger than the overall number of papers examined, (19), because some papers investigated multiple variables (e.g. both *Performance* and *Behaviour*).

Table 5.4: The total number of variables investigated.

Variable Analysed	Total	Sig	Unknown	Non-sig
Performance	**14**	7	1	6
Experience	**26**	15	2	9
Time	**11**	5	0	6
Behaviour	**21**	14	2	5
Physiology	**6**	6	0	0
Total	**78**	**47**	**5**	**26**

5.4.1.1 Perceptual Speed and Search Performance

As was observed in Table 5.3, 11 papers examined how PS had affected some kind of search performance, regardless of how performance had been quantified in each paper. For some, this meant the number of relevant documents retrieved for a specific topic [4], while for others, performance was quantified through target identification [84], correct/incorrect responses [43], knowledge learned [11], or comprehension accuracy [220].

Given these differences, it was thought best to organise the results based on natural themes that emerged which connected some papers, such as multiple papers that quantified performance in the same way. However, to be explicit and keep track of which papers returned significant, or non-significant results, and with how many variables, please refer to Table 5.5.

Table 5.5: The reference for every paper that investigated a measure of ***Performance*** (and how many measures of performance), against whether the result returned significant, unknown, or non-significant results

Result found	Performance Papers (and number of measures analysed)
Significant	Allen [9] (1), Allen [11] (2), Allen [13] (1), Carenini et al. [56] (1), Conati et al. [64] (1), Naghib et al. [174] (1)
Unknown	Lalle et al. [139] (1)
Non-significant	Al-Maskari & Sanderson [4] (1), Fincannon et al. [84] (3), Brauner et al. [43] (1) Toker et al. [220] (1)

Firstly, for the one paper where results were unknown, this was due to Lalle et al. [139] stating the following: "*We found that users with low levels of these four abilities compared both the maps and the charts less extensively across transit scenarios than their counterparts*". Here, PS had been grouped alongside other cognitive abilities, including spatial memory, visual scanning, and visualization literacy, and therefore disentangling the effects of just PS was not possible.

Whilst Lalle et al. [139] did not provide a clear performance effect, other papers provided multiple performance effects. For example, Fincannon et al. [84] monitored three measures of performance, in a simulated military operation search task: ability to stop an unmanned vehicle; correctly identifying a pedestrian target; and competence at re-routing the vehicle. However despite multiple performance measures investigated in this paper, all of them returned non-significant results that suggested PS did not affect any form of search performance.

In contrast, despite other papers reporting a non-significant result, this did not mean that a difference between Low-PS and High-PS was not present. For example, Al-Maskari & Sanderson [4] found that on average, Low-PS users retrieved less relevant documents when completing a TREC Search Task (4.39 *documents*), compared to High-PS users (4.79 *documents*). Although not significant, when rounding to the nearest whole number, this showed that High-PS users were able to identify and retrieve an entire extra relevant document, compared to users with Low-PS.

Al-Maskari & Sanderson's [4] non-significant results assimilated to significant results found in three other papers by Allen ([9], [11], and [13]) who all identified that users with High-PS retrieved significantly more relevant documents in comparison to Low-PS users (For example, $p<.01$ in [9]). Here, a relevant document was classed as those viewed by, and judged useful, by more than half of participants who considered them relevant to the topic, and therefore Allen related document relevance to a precision ratio: High-PS achieved searches with significantly higher precision ratios.

For other types of search performance, specifically, correct/incorrect answers, results were more juxtaposing between studies. For example, Carenini et al. [56]—whose search task involved participants observing different visualisations and answering textual questions such as "*in how many courses are both Andrea and Spencer below the class average?*"—reported that High-PS users were "*significantly better at completing tasks*" compared to users with Low-PS and Average-PS ($p<.01$). Similarly, Naghib et al. [174], whose search task involved students searching educational websites to find answers to specific questions such as "*What is the primary source of oil?*", stated that PS had significant linear relationships with searching success at a 95% confidence level – where search success was calculated based on correct answers. Although the direction was not stated—and thus it was unknown whether users with Low-PS or High-PS had higher search accuracy—a significant difference between the two still emerged. This was also the case in Conati et al. [64], who in another task that involved participants answering specific questions about visible data, another significant difference between Low-PS and High-PS users ($p < .001$) was reported, and yet the direction of results was not specified. Yet in comparison, Brauner et al. [43], who also quantified search accuracy as correct/incorrect responses (albeit in a different task: a simulated business game of reading stock levels and deciding whether there was enough based on projected customer demands), found that PS did *not* influence search performance.

Although performance was generally quantified as some kind of active performance *during* the search task, two studies additionally measured performance *post-task*. Yet, both results were also contradicting. In one task which involved participants identifying relevant documents in a specific topic, High-PS users learned significantly more vocabulary as a result of their search

($p < .05$) [11]. In comparison, Toker et al. [220] identified no difference in comprehension accuracy between Low-PS and High-PS users when conducting visualisation tasks.

5.4.1.2 Perceptual Speed and User Experience

Similar to how performance could be quantified in different ways, measures of experience could also be further broken down into the *type* of user experience that was investigated (*See Table* 5.6). This included: 1) **workload**, as measured using the NASA TLX; 2) **system usability**, such as reported ease of use, user engagement, aesthetic appeal, and ease of understanding; and 3) **belief of search success**, which considered whether users believed they had searched well, or if they thought the task was more difficult.

Table 5.6: The reference for every paper that investigated a measure of ***Experience*** (and how many measures of Experience), against whether the result returned significant, unknown, or non-significant results.

Result found	Experience (Workload)	Experience (Usability)	Experience (Success)
Significant	Fincannon et al. [84] (1), Brennan et al. [46] (6)	Turpin et al. [225] (2)	Turpin et al. [225] (4), Crabb & Hanson [71] (2)
Unknown	0	Toker et al. [221] (1)	Allen [9] (1)
Non-significant	Arguello & Choi [18] (1)	Crabb & Hanson [71] (1), Arguello & Choi [18] (4), Toker et al. [220] (1)	Al-Maskari & Sanderson [4] (1), Conati et al. [64] (1)

Workload.

Firstly considering workload, although Fincannon et al. [84] found no differences between users with Low-PS and High-PS during search *performance* in a simulated military operation search task (*See Section* 5.4.1.1), significant differences were identified for user *experience*: High-PS experienced less self-reported workload, which was significant in every factor of the NASA TLX. Although Brennan et al. [46] investigated PS using different search tasks—specifically, users had to find a specific piece of information, analyze different sources before explaining them, and create a novel solution for a problem identified during their search—significant differences in all factors of NASA TLX were also found: High-PS experienced less self-reported workload. In other words, if a user had Low-PS, then overall they appeared to find the search tasks more physically and mentally demanding.

However, in Arguello & Choi's [18] research—which involved participants completing comparative information tasks, such as comparing different water purification methods for eliminating bacteria and saving relevant websites that helped them make decisions—the opposite pattern occurred for self-reported workload: there were no differences between Low-PS and High-PS using NASA TLX measurements.

System Usability

Continuing with the research example of Arguello & Choi [18], in addition to monitoring perceived workload, various measures of system usability were also investigated, including: perceived usability, aesthetic appeal, focused attention, and reward – all of which found no

PS differences. Similarly, in other research, specifically Toker et al.'s [220] visualisation tasks, no differences in PS were found for participant's subjective ratings of the system's ease of understanding.

Yet, in a different search task implemented by Crabb & Hanson [71]—specifically, where users were instructed to follow a path through various websites, such as NHS 24, to answer specific questions, such as "*What groups of people are eligible for a seasonal flu jab?*"—High-PS users reported higher system ease-of-use in comparison to Low-PS, although this was also not significant. Nonetheless, in another study—where users were to generate something informed from a specific search, such as creating an exercise program—High-PS users also rated *a)* the ease of use and *b)* the perceived usability of every system investigated, as higher, in comparison to Low-PS, and these differences *were* both significant [225].

Belief of Search Success

Next, when aggregating how users rated the experience of their overall search success, more contradicting results occurred. In Turpin et al.'s research [225], just as a significant result was found for perceived system usability, additional significant differences were found between Low-PS and High-PS users in terms of search success belief. In comparison to Low-PS, High-PS users: rated their searches as significantly more successful; believed their own search skills were significantly better; assessed the system's ability to retrieve relevant documents as significantly higher; and thought they had found significantly more relevant documents overall [225]. Furthermore, whilst Crabb & Hanson's [71] research did not find significant differences between PS and system ease of use, significant differences in belief of search success were found, where High-PS had a significantly higher browsing experience and lower levels of perceived disorientation.

In contrast, in Conati et al.'s [64] research involving data visualisations, questionnaires elicited no significant effects in terms of cognitive abilities and a user's search belief. Similarly, in Al-Maskari & Sanderson's [4] research—where users were to retrieve relevant documents about a specific topic—Low-PS and High-PS were equally satisfied with their searches. Therefore, just as was the case with workload and usability, both significant and non-significant results were also obtained for how PS impacted belief of search success.

5.4.1.3 Perceptual Speed and Search Time

When analysing papers which measured some form of time, which was 9 papers in the sample, these considered the **overall** length of search session, or the amount of time taken to **click**, or **save**, a particular item (*See Table* 5.7). Of these 9 papers, slightly more (5 papers) reported significant results, compared to non-significant results (4 papers). Yet, breaking the papers down into number of time measures overall, the opposite occurred, with slightly more results found being non-significant (6) in comparison to significant (5).

Overall Search Time

Table 5.7: The reference for every paper that investigated a measure of ***Time*** (and how many measures of Time), against whether the result returned significant, unknown, or non-significant results.

Result found	Time (overall)	Time (clicking)	Time (saving)
Significant	Allen [11] (1), Toker et al. [221] (1), Brennan et al. [46] (1), Naghib et al. [174] (1)	0	Al-Maskari & Sanderson [4] (1)
Unknown	0	0	0
Non-significant	Turpin et al. [225] (1), Arguello & Choi [18] (1), Toker et al. [220] (1)	Arguello & Choi [18] (1)	Brauner et al. [43] (1), Arguello & Choi [18] (1)

Firstly examining overall length of search session, which 7/9 papers investigated, four reported significant differences between Low-PS and High-PS users. These all followed the same direction: users with Low-PS took significantly longer completing search tasks, across a variety of different scenarios: identifying relevant documents for a specific topic [11]; evaluating student performance across different visualisations [221]; finding a specific piece of information, analysing different sources before explaining them, and creating a novel solution for a problem identified during their search [46]; and searching educational websites to find answers to specific questions [174].

Yet, for the non-significant findings reported in other research, the opposite trend occurred: in Arguello & Choi's research [18], Low-PS users actually took *less* time completing search tasks. For the remaining two papers that also reported non-significant differences in search time, no trends were reported which indicated whether Low-PS, or High-PS had equivalent search completion times, or if one was slightly faster than another.

Time spent Clicking and Saving

As the study by Arguello & Choi [18] found no *significant* difference between Low-PS and High-PS users regarding overall search time, it was unsurprising that they also found no significant differences for other, more specific aspects of search time. Yet, the same trend did occur, that Low-PS were slightly faster in comparison to High-PS users, with both the time taken until their first **click** on the search system, and the time taken until their first relevant bookmark was **saved** [18]. Unfortunately, no other studies monitored the time taken for a user before they clicked on something, and therefore comparisons cannot be observed for this metric. However, two other studies did examine the time taken before something was saved, yet contradictory results were also found: in Brauner et al. [43], PS did not influence time in a simulated business game; but in Al-Maskari & Sanderson [4], Low-PS took more time making a decision, as High-PS users spent significantly less time until they retrieved their first relevant document during TREC search tasks (High-PS: 1.71 minutes, Low-PS: 2.19 minutes). There therefore appeared to be no consistency in results found between studies: in one, no differences were found [43]; in another, Low-PS made faster decisions [18]; and yet in another, Low-PS made longer decisions [4].

5.4.1.4 Perceptual Speed and Search Behaviour

With a smaller sample of papers (6) that considered search behaviour as a dependent variable in their research, a diverse distribution of results was still found: two papers identified significant results [12, 174], although one of these did not specify the direction [174]; two papers identified both significant and non-significant results [18, 46]; one returned non-significant findings [225]; and another was unknown [9]. However, breaking the results down into individual aspects of behaviour, including behaviour related to *querying*, *viewing*, and *clicking* separately, a much larger proportion of results found significant findings (14) compared to non-significant findings (5) (*See Table* 5.8).

Table 5.8: The reference for every paper that investigated a measure of ***Behaviour*** (and how many measures of Behaviour), against whether the result returned significant, unknown, or non-significant results.

Result found	**Behaviour (query)**	**Behaviour (view)**	**Behaviour (click)**
Significant	Brennan et al. [46] (1), Arguello & Choi [18] (5), Naghib et al. [174] (2)	Brennan et al. [46] (2), Allen [12] (1), Naghib et al. [174] (2)	Brennan et al. [46] (1)
Unknown	0	Allen [9] (2)	0
Non-significant	Brennan et al. [46] (1), Turpin et al. [225] (1)	0	Arguello & Choi [18] (1), Turpin et al. [225] (2).

Viewing Behaviour

Although there was no consensus for one specific factor producing a consistent result in measures of user *Experience* or *Time*, this was not the case for *Behaviour*. Here, a more visible pattern emerged for behaviour regarding what users *viewed*. Metrics involving behaviour views were gathered in four of the papers, although one did not report their results [9]. For the remaining three papers, all found significant results. While the directions were not reported in Naghib et al. [174], the other two papers were consistent: Allen [12] found that users with High-PS viewed significantly more records than Low-PS in a TREC search task ($p < .01$); and 16 years later in different search tasks—which involved finding specific information, analysing different sources before explaining them, and creating a novel solution for a problem identified during their search—Brennan et al. [46] identified two more significant findings: in all search tasks, High-PS users viewed significantly more URLS, both per-query, and overall.

Querying Behaviour

Although two significant results were found regarding viewing behaviour in Brennan et al.'s [46] research, in terms of querying behaviour, both significant and non-significant results occurred. Firstly, when the search task began, an immediate difference between PS users occurred: High-PS users issued significantly longer **queries** [46]. This would imply an additional aspect of PS, where it does not only relate to processing information, but also impacts how users *seek* information. However, although High-PS users also appeared to issue more queries in each search task, this difference was not significant [46]. As this was non-significant, the

results regarding number of queries may have been due to chance. This could make sense, as in Turpin et al.'s [225] research—in a task that required users to generate something novel as a result of their search—Low-PS and High-PS users had similar numbers of queries overall.

Yet, in Arguello & Choi's [18] research, another contrasting finding occurred, where Low-PS users issued significantly *more queries* overall. Furthermore, Low-PS users also: *a)* had significantly more quick query reformulation; *b)* had significantly more queries without scrolls; *c)* had significantly more queries with repeated intent; and *d)* issued significantly more queries that did not result in saving a bookmark.

Finally, in Naghib et al.'s [174] research—which involved students searching educational websites to find answers to specific questions such as "*What is the primary source of oil?*"—there were again significant differences in querying behaviour - both in query length, and reformulation of the search question. However, the direction of these results were not reported.

Clicking Behaviour

In addition to contrasting results concerning querying behaviour, opposing differences also emerged regarding clicking behaviour. Firstly, Brennan et al. [46] identified that High-PS users clicked on significantly more search engine result page (SERP) links. Similarly, Arguello & Choi [18] identified that High-PS issued slightly more clicks, although this was not-significant. Then, Turpin et al. [225] also reported that High-PS clicked on more vertical results. However, observing clicks overall, Turpin et al. [225] found the opposite: Low-PS clicked on *more* things. Whilst Turpin et al.'s [225] two results concerning clicking behaviour were non-significant, they still indicate trends that show there was no clear consensus for clicking behaviour across users with different levels of PS.

5.4.1.5 Perceptual Speed and Eye Fixation

Only two papers studied a raw overall difference in eye-gaze between users of different PS levels, and it is important to note that these both came from the same user study. Whilst this represents the smallest sample of all dependent variables analysed, it also demonstrates the only variable where only significant results occurred. In both papers, three separate measures of eye-gaze were reported, and all of these returned significant differences (*See Table* 5.9).

Table 5.9: The reference for every paper that investigated a measure of ***Physiology*** (and how many measures of Physiology), against whether the result returned significant, unknown, or non-significant results.

Result found	**Physiology**
Significant	Toker et al. [222] (3), Steichen et al. [211] (3)
Unknown	0
Non-significant	0

The search task involved users evaluating student performance in eight different courses, using both bar graph and radar graph visualisations. The overall results can be summarised as

follows, where Low-PS users significantly: spent more time in the legend of the visualisations; transitioned their gaze to the legend more often; but overall had less fixation-rate [222]. Additionally, in Steichen et al.'s paper [211], who investigated the sequential nature of user eye gaze patterns through differential sequence mining, further differences were reported. Firstly, users with High-PS had more fixations in the data labels, both before and after looking at multiple values in the visualisation. Secondly, "*'High' AOI to 'Label' AOI transitions that are broken up by an intermediate fixation at the 'Low' AOI (i.e., Hi-Lo-La) occur more frequently for low PS users.*" While this can be difficult to interpret out of context, Steichen et al. [211] provided one possible interpretation of this finding, which referred to how Low-PS users may be less precise when trying to locate a small 'Label' that was associated with a value on the visualisation. Thirdly, Low-PS users had many more repeated fixations within the 'Text' AOI, which Steichen suggested may signify that Low-PS users require more effort to process the larger textual components of visualisations. While one user study can be difficult to generalise the results from, it still indicates that there are differences in physiology with users of different PS levels.

5.4.1.6 Summary of Overall Claims found

Bringing all overall claims together, it was evident that there was a large variation regarding how PS affected users:

- In terms of performance, exactly 50% of variables investigated reported a significant result, 42% reported a non-significant result, and around 7% were unknown.
- Similarly, for all 3 categories of user experience (Workload, Usability, and Success), results were contradicting in each - there was no one variable that consistently produced a significant, or non-significant results, or in a consistent direction (E.g. one study claimed that workload did [84], whilst another claimed that it did not [18], differ between users of different PS levels).
- Likewise, the results regarding how PS affected search time were also contrasting: some studies claimed that Low-PS users took significantly longer completing a task [221], others claimed that Low-PS users were faster [18], and others reported no differences [43].
- Furthermore, although there appeared to be a consistent trend that the behavioural aspect of viewing search items significantly differed between users with different levels of PS—particularly, the fact that High-PS users viewed more than Low-PS users [12, 46]—there were a lot of contrasting differences in the other behavioural measures: in some studies, Low-PS clicked on more things [225], whereas in others, High-PS clicked on more things [18, 46]. Similarly, for querying, some studies reported no differences between users with

different PS levels [225], others reported that Low-PS issued more queries [18], whereas another study found that High-PS issued more queries [46].

- The only variable which reported fully significant results—which came from two papers that used data from the same user study—referred to differences in eye-gaze found between Low-PS and High-PS users.

With such contrasting overall claims, this emphasised the importance of further exploring how the claims interacted with other variables investigated. Consequently, all interactions reported were categorised below in Section 5.4.2.

5.4.2 Perceptual Speed Interactions

To continue answering the first research question of ***'What claims have occurred regarding PS in Computer Science?'***, in addition to reporting all overall effects that have been claimed regarding PS, any interactions with other variables were also categorised. In addition to reporting an overall effect of PS, 12 out of the 19 papers also reported interaction effects. Additionally, 6 other papers in the original sample of 40 papers also reported interaction-only results. As some studies reported multiple interactions, overall, there were 29 interactions to analyse, of which, 24 were significant and 5 non-significant.

Having examined the interactions, natural themes emerged for what kind of other variables interactions had been reported with, including: another independent variable, such as task type or interface; or an external variable, such as age or sex. Consequently, these variables have been described separately below. However, to be explicit as to where the significant and non-significant interactions occurred, and from whether this stemmed from an interaction-only or interaction and main-effect paper, please refer to Figure 5.3.

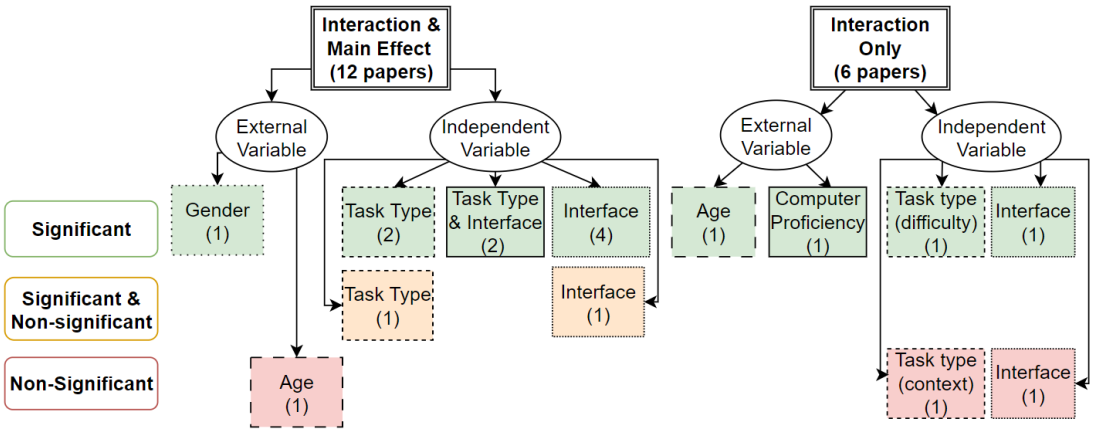


Figure 5.3: The different types of variables that reported an interaction with Perceptual Speed, where the number in brackets refers to how many papers were in that category.

5.4.2.1 External Variables

Examining external variables, a few interactions were noted outwith the search environment studied. Firstly, in Brauner et al.'s [43] experiment involving an analysis of online stock levels and making a decision, a correlation analysis revealed a strong significant relationship between ***gender*** and PS, where women had a much higher PS score in comparison to men ($p<.01$). Secondly, using a survey as a research methodology, Zhang et al. [243] claimed that users with higher levels of PS were positively correlated with higher scores on a ***Computer Proficiency*** Questionnaire (CPQ). Similarly in another survey, O'brien et al. [178] claimed that ***age*** affected PS, where users with higher levels of PS were younger, or an older adult who was competent with technology, in comparison to a much lower PS score being found in older adults who were less familiar with technology ($p<.05$). However in contrast, in Crabb & Hanson's study [71]—which involved an experiment with 30 search tasks that required users to navigate various websites, in order to answer specific questions, such as *"What groups of people are eligible for a seasonal flu jab?"*—no age-related differences regarding PS were identified.

5.4.2.2 Task Type

Task Context.

In the earliest study that manipulated task type (by Kim & Allen, 2002 [125]), the discussion given that related to PS stated the following: *"The hypothesis was that different tasks would be associated with different levels of search activities and outcomes, but only for searchers with higher levels of cognitive abilities. In the case of PS, no interaction effect was found"* [125]. When observing what the different tasks investigated were, this was operationalised based upon ***context***. For both tasks, university students had to read the same article, and were then asked to search and retrieve relevant documents about that topic. In one task, the aim given to students was to search and find relevant information in order to write a detailed *10-page term paper*; in the other task, the aim given was to search and find relevant information in order to write a *student newspaper article*. Thus, both tasks involved the same topic and search interface, but based upon subtle context differences, these did not appear to impact the results found.

Task Complexity.

With such subtle differences in Kim & Allen's [125] research, it is unsurprising that no interactions were found, as the task differences were much less obvious than different studies, such as Brennan et al. [46]. Here, the tasks differed more obviously, based upon the number of actions required to produce an outcome. For example, an experiment was conducted with three different search tasks, which varied in terms of the complexity involved in completing the websearch: *1) Remember* - find a specific answer; *2) Analyze* - generate a list of items and explain them; and *3) Create* - generate a novel solution [46]. While overall it was reported

that Low-PS took significantly longer than High-PS users when completing the search task (*See Section* 5.4.1.3), an interaction effect identified that this was only for the '*Remember*' and '*Create*' tasks. In contrast, Low-PS users were actually faster than High-PS users when completing the '*Analyze*' search tasks.

Task Difficulty.

Similar to task complexity, which was defined by the number of actions required to produce an outcome, task difficulty concerns the amount of cognitive effort required to complete a task [46]. In Toker & Conati's research [218]—which involved a visualisation task using a bar graph alongside textual questions about the data presented—a significant interaction was found between task difficulty and PS: using eye-tracking, Low-PS users navigated to the labels of the bar graph more, when the task was ***more difficult*** ($R^2 = .009$). Here, task difficulty was quantified by task type: '*Retrieve Value*' was considered a simple task, whereas '*Compute Derived Value*' was considered a difficult task.

Similarly, another paper by Carenini et al. [56]—which just happens to have used data from the same user study as Toker & Conati [218]—reported further interactions, alongside main effects. As identified in Section 5.4.1.1, Low-PS users performed similarly to Average-PS users, but both were significantly worse in comparison to High-PS users. However, when looking at the interaction of performance against *Task type* and PS, whilst there was always a trend that High-PS performed better, this was only significant for the most complex task—*Compute a derived value*—again reaffirming that when a task is difficult, Low-PS users struggled the most.

Furthermore, in a different study—which was conducted the same year as Carenini et al. [56] and Toker & Conati [218]—Conati et al. [64] experimented with 5 different types of tasks that all involved users answering specific questions about data given in a visualisation. The 5 types (*Retrieve value, Find extremum, Sort, Compute derived value 1*, and *Compute derived value 2*) also ranged in order of difficulty, where the hardest task involved more cognitive effort to answer. For example, for *'Retrieve value'*, a typical question was *"Is the value of 'skytrain-distance' of hotel3 less than hotel6?"*. In comparison, for *'Compute value 2'*, a question involved *"List the top 3 homes (in descending order) according to the aggregated value of 'cost' and 'space'"*. Whilst a significant main effect of PS was stated for performance, no detail or direction was given (*See Section* 5.4.1.1). However, three interactions were reported: Low-PS users were significantly slower than both Average-PS and High-PS users during *Retrieve* tasks; Low-PS users were also slower in *Sort* tasks, albeit this was not significant; but as the tasks increased with difficulty, Low-PS and Average-PS were significantly slower than High-PS during *Find Extremum, Compute value 1*, and *Compute value 2*.

For the remaining two papers that also reported interactions with task difficulty, these came from a 2012 paper by Toker et al. [221] and 2013 paper by Toker et al. [222], which both stemmed from the same user study: performing a battery of visualisation tasks. There were two

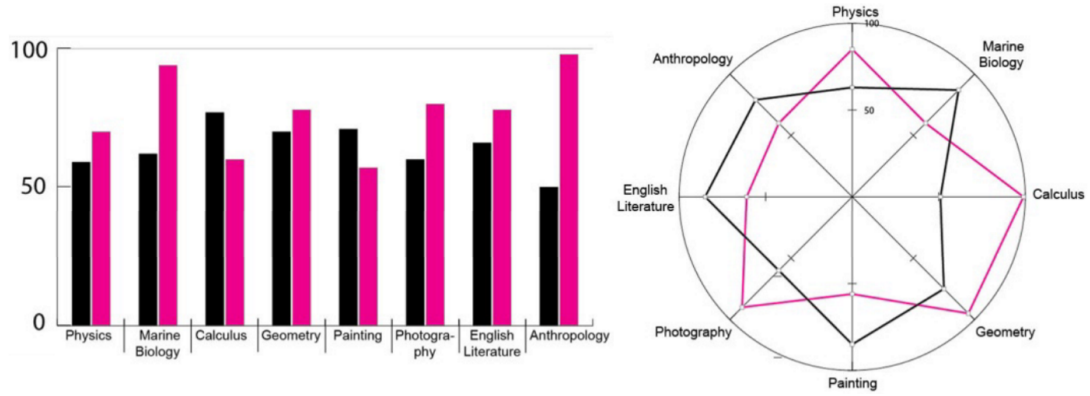


Figure 5.4: The interfaces used in Toker et al.'s research [221], where Low-PS took longer completing tasks using a radar graph (right) as opposed to a bar graph (left). (Taken from: [221]).

types of search task administered in this study: *1) Single scenario* - which involved a question such as *"in how many courses is Maria below the class average?"* ; and *2) Double scenario* - where the question involved two factors, such as *"Find the courses in which Andrea is below the class average and Diana is above it?"*. In Toker et al. (2012, [221]), a significant interaction was noted, where Low-PS took significantly longer completing tasks during the *single scenario* type. However for the *double-scenario* task, no PS effects were observed. Then, in Toker et al. (2013, [222]), which reported results from eye-tracking during the user study, two further significant interactions were noted: during an easy task (the *single-scenario*), Low-PS users: a) transitioned to the label more; and b) transitioned to the legend less. However, as Low-PS did not take longer completing the search in more difficult tasks in these papers, this result is in contrast to Brennan et al. [46] and Conati et al. [64], where Low-PS did take longer completing searches in more difficult tasks.

5.4.2.3 Interface

Radar graph versus Bar-graph.

In addition to just examining *Task Type* in Toker et al.'s 2012 and 2013 research [221, 222], different interfaces were also investigated, where the main visualisation was either: *1) a Bar-graph*; or *2) a Radar graph* (*See Figure* 5.4). Whilst an overall main effect of PS was found for time, where High-PS users were significantly faster completing search tasks in comparison to Low-PS users (*See Section* 5.4.1.3), a significant interaction was also identified: Low-PS users could complete tasks using a *Bar-graph* faster than High-PS users completing a task using a *Radar graph* [221]. When a *Radar graph* was used, Low-PS performed the slowest. Additionally, Toker et al. (2013, [222]) identified another significant interaction: Low-PS users had more gaze activity within a specific Area Of Interest (AOI) with *Radar graphs*, in comparison to *Bar-graphs*. Combining these two papers together, this demonstrated that given a particular

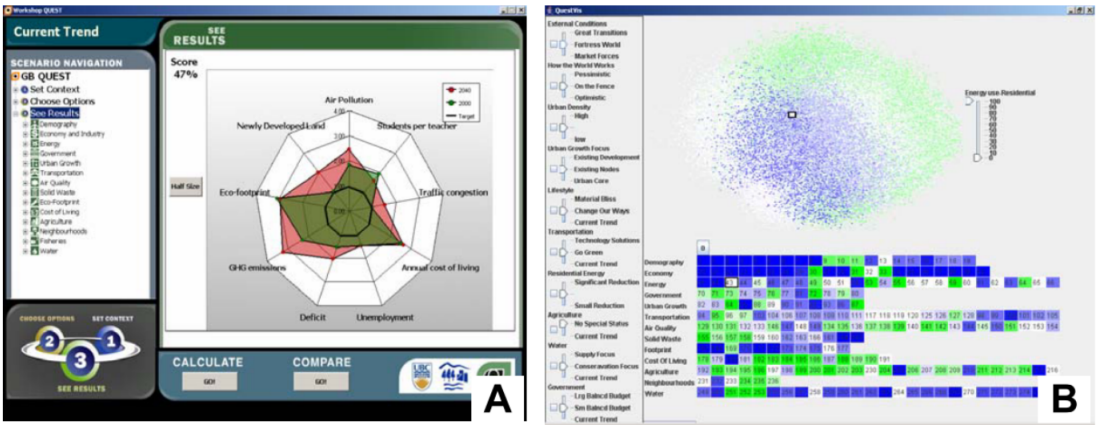


Figure 5.5: The interfaces used in Conati & Maclaren's research [68]: A) radar graph; and B) coloured boxes. Low-PS were more accurate with A, whereas High-PS were more accurate with B. (Taken from [68]).

visualisation, Low-PS users struggled to process it, namely *Radar graphs*, and thus had to re-scan the visual interface more, which ultimately increased their overall search time.

Radar graph versus Coloured boxes.

In other research, Conati & Maclaren [68] identified that Low-PS users performed searches with *higher* accuracy when using *Radar graphs*. However, in this research, *Radar graphs* were not compared to *Bar-graphs*, but instead a very cluttered display of many *coloured boxes* (*See Figure* 5.5). Specifically, a significant interaction of visualisation type with PS occurred ($p = 0.035$), where High-PS users were more accurate in their search performance when using colored boxes rather than radar graphs, whereas Low-PS were the opposite, and had higher accuracy using radar graphs. However, the authors were unable to explain these differences: *"We don't have a conclusive explanation for the direction of the relationships that we found among perceptual speed, accuracy with the radar graph and accuracy with the colored boxes"* [68].

Order of Information.

Although Conati & Maclaren [68] found differences between interfaces interacting with PS, Allen [10] also experimented with different interfaces, but found *"no significant interaction between results on either of the tests of perceptual speed and the interface type used in influencing search precision, $p = .223$"*. In Allen's work where no differences were found, users had to scan a list of over 700 subject headings, and select any that they believed represented a specific topic. A form of querying was possible, where participants could enter a search expression, and the list would jump to the most closely related heading. One interface presented 23 main headings per screen, in alphabetical order, where under each heading, there were sub-headings. The other interface displayed more main headings, 39, however, sub-headings were not visible unless the main heading was explicitly clicked on to expand (*See Figure* 5.6). It was hypothesised that the ability to expand sections would minimise the amount of scanning required, which could benefit users with Low-PS. Yet as no interaction was found, this hypothesis was rejected [10].

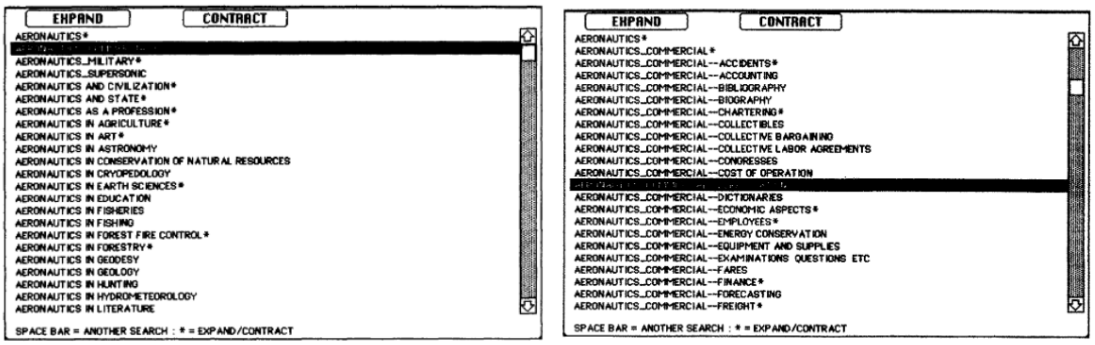


Fig. 3. The expand/contract interface before expansion.

Fig. 4. The expand/contract interface after one expansion.

Figure 5.6: The expand/contract interface before and after expansion used in Allen's 1994 research [10]. No image was given for the other interface. (Taken from [10]).

However, in other work by the same author where the order of information also varied, albeit in a different manner, significant interactions did occur. Here, users had to identify relevant documents for a specific topic using different interfaces: *1)* presented the documents in a *usual order (Author, Title, Source, Subjects Covered, Abstract)*; and *2)* the documents were presented with the *subject heading first (Subjects Covered, Author, Title, Source, Abstract)* [11]. It was anticipated that displaying the *subject heading first* would result in faster scanning and subsequently greater learning. Although three significant main effects were previously identified—which described how High-PS users were faster, achieved better search performance, and learned more, in comparison to Low-PS users (*See Section* 5.4)—these had all been from when people used the second system, designed to enable fast scanning using the *subject heading first* interface. In contrast, a clear interaction was found, that when people used the *usual* system, there was no effect of PS on: a) the amount of learning, b) search performance; and c) search speed. These results would imply that in some interfaces, Low-PS and High-PS users do not differ in their search abilities, and yet in a system designed to enable faster scanning, Low-PS struggled to engage in fast scanning, and subsequently performed more poorly.

Single Window versus Multiwindow.

A few years later, Allen [12] created another experiment, where again, users had to identify relevant documents in relation to a specific topic. However, instead of comparing interfaces that differed based on the order of information, two other interfaces were experimented with: *1) a single window word map-* where clicking on a word in the map caused a box to be drawn around the word, and the term list would scroll to a term associated with that word; and *2) a multiwindow display-* where each element of the bibliographic record appeared in a separate box, all of which were visible on the screen concurrently (*See Figure* 5.7).

Although Allen [12] only reported a main effect of behaviour, and not performance, the interaction reported concerned performance in terms of how much vocabularly had been learned during the search: High-PS learned more vocabularly using the *single window* display, whereas Low-PS learned more using a *multiwindow display* ($p < .03$). Allen [12] proposed one possible

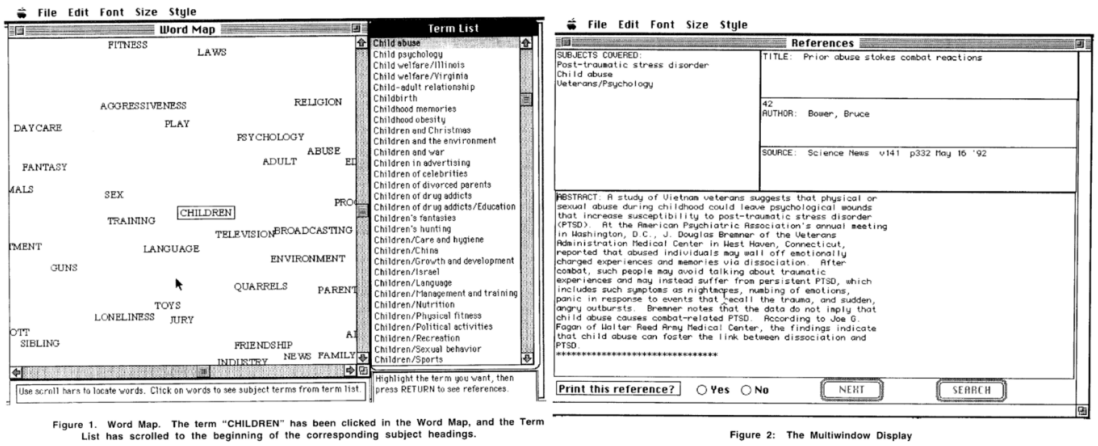


Figure 1. Word Map. The term "CHILDREN" has been clicked in the Word Map, and the Term List has scrolled to the beginning of the corresponding subject headings.

Figure 2: The Multiwindow Display

Figure 5.7: The interfaces used in Allen's 1998 research [12], showing the single window word map on the left, and the multiwindow display on the right. (Taken from [12]).

explanation of their results: *"A partial scan strategy likely to be employed by users with lower levels of perceptual speed would be facilitated by the clear separation of data elements in the multiwindow display. Use of the multiwindow display by users with higher levels of perceptual speed seems to impair the amount of learning. A top-down, sequential scanning strategy likely to be used by individuals with higher levels of perceptual speed would be impeded by the breaking up of the display into separate windows"*.

However, in a further paper by Allen [13], which reported different results from the same user study that was published two years previously in 1998 [12], another interaction revealed a different pattern of search behaviour: individuals with Low-PS achieved higher recall (as indicated by the number of references viewed) when they used the *word map*; individuals with higher levels of PS achieved higher recall when they used the *multiwindow display* ($p<.01$). Consequently, a complex interaction appeared to be occurring, where Low-PS may perform better searches during the *word-map* interface, but simultaneously remember less of what they have learned during this search scenario. Unfortunately, Allen [13] proposed no explanation for these results.

Blended versus Non-blended.

Fast forward almost 20 years, the bibliographic retrieval systems that Allen had been experimenting with had been replaced with more complex Information Retrieval systems, such as the likes of Google, where videos, images, and news articles could also be viewed in addition to just textual results. In 2016, Turpin et al. [225] created an experiment utilising this kind of retrieval system, where users had to navigate online in order to create something new, such as an exercise programme. Overall, it was identified that there were no differences in search time, between users with Low-PS and High-PS completing this task (*See Section* 5.4.1.3). However, in Turpin et al.'s research [225], two different interfaces had been experimented with: *1) Blended* - this blended vertical results—which included images, videos, news results, and shopping—into the

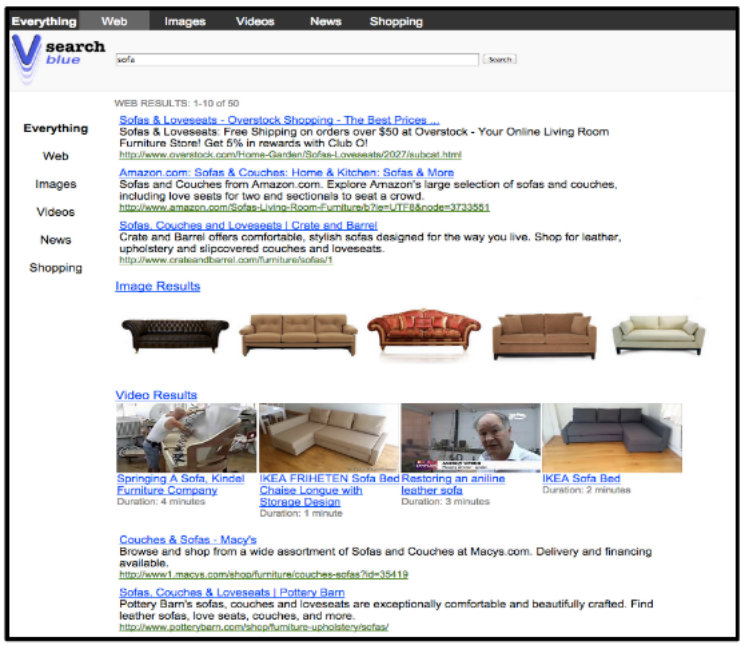


Figure 5.8: The blended interface used in Turpin et al.'s research [225]. The non-blended interface was not shown in the original paper. (Taken from [225]).

main web result page (See Figure 5.8); and *2) Non-Blended* - the vertical results could only be accessed individually through separate tabs. Here, a significant interaction was reported, where Low-PS spent significantly longer completing tasks with the *blended* interface, and significantly less time completing tasks with the *non-blended* interface, while High-PS spent similar amounts of time completing tasks regardless of interface. It therefore makes sense, that overall, no differences between PS users in terms of time were noted, because the two different conditions balanced out the scores of Low-PS users. Yet, when different types of content were merged together, such as images and videos all being visible at once, Low-PS took much longer to process the information. However, it must be noted that similar to Toker et al. [221], no measure of performance was reported, and therefore it is unknown whether longer searching was a good, or bad thing, in terms of search accuracy.

Interleaved versus Blocked.

For the final paper that reported interaction effects, these stemmed from Arguello & Choi [18], who implemented comparative information tasks, such as asking users to compare different water purification methods for eliminating bacteria, and saving relevant websites that helped them make decisions. Similar to Turpin et al. [225], in Arguello & Choi's research [18], two different interfaces were used, which differed based on the verticals present (See Figure 5.9): *1) the Interleaved* condition- which combined results from different verticals in an unconstrained manner (web, images, news, shopping, videos); and *2) the Blocked* condition - which presented results from the same vertical as a group using visual cues, such as borders with shadow effects, which were always positioned in the same regions on the SERP, to help distinguish between

(a) Interleaved Interface (truncated)

(b) Blocked Interface

Figure 5.9: The interfaces used in Arguello et al.'s research [18], where the left shows the Interleaved interface, and the right depicts the Blocked interface. (Taken from [18]).

results from different verticals. Overall, many main effects were reported between Low-PS and High-PS users (*See Section* 5.4). However most of these effects were not significant, such as the number of clicks, or overall time spent searching.

In contrast, many significant interactions occurred in Arguello et al.'s research [18], showing that in the *Interleaved condition*, Low-PS users had: more queries; more clicks; and more queries without mouseovers [18]. Arguello then expanded these results by stating: *"Low-PS participants were less effective at finding relevant results on SERPs with the interleaved interface. Possibly, the blocked interface allowed low-PS participants to perform less visual scanning by allowing them to focus their attention on relevant regions of the SERP...One perspective is that high-PS participants were equally effective at finding relevant results with both interfaces. Another perspective is that the blocked layout helped low-PS participants find relevant results as effectively as high-PS participants"* [18]. However, although there were behavioural differences reported, it is unclear how Arguello & Choi concluded search efficiency, based on the variables investigated, as no interactions were reported for search performance or search time. For example, Low-PS users querying and clicking more in the *Interleaved* condition could mean that they were engaging in the search more here, and thus being *more efficient* in their search. A lot more clarification is thus needed with future experiments, but nonetheless, behavioural differences still emerged, implying that different interfaces do impact users with different levels of PS.

5.4.2.4 Summary of Interactions found

Summarising together, similar to the contrasting claims identified for the overall effects of PS, the many different interactions observed also demonstrated a variety of contrasting findings:

- Firstly, a few external variables had been investigated in relation to PS, but no conclusive results emerged. For example, whilst one study claimed PS differed by age [178], another

study reported no age-related differences in PS [71].

- For an independent variable, the type of task, for some areas such as task context, no differences were observed in relation to PS [125]. However for other areas, such as the task difficulty, many significant interactions were reported. Whilst most of these claimed that Low-PS users performed less efficient searches when the task was more difficult (E.g. Low-PS users took longer searching in more difficult tasks [64]), other research found opposing results (E.g. Low-PS users took longer in easier tasks [221]).

- Considering another independent variable, the interface used, patterns emerged where differences in the type of information present seemed to affect PS (E.g. Low-PS took longer searching when videos and images were merged together [225], and were more effective at finding relevant results when videos and images were clearly separated [18]). However, further mixed results were reported, where although one study that manipulated the order of information found that Low-PS and High-PS significantly differed in terms of search accuracy and search time [11], other similar research found no interactions [10].

Consequently, understanding *why* there was such disparity between the overall results of different studies remained a prominent question to investigate, which was explored next in Section 5.5, which critically evaluated possible effects of PS.

5.5 Results: Evaluation of Perceptual Speed in Computer Science

Although some papers in the Introduction (See Section 5.2) had been motivated based upon incorrect citations, whilst others provided no results regarding PS—and this cast doubt as to whether PS did, or did not, affect users in the domain of Computer Science—the previous sections demonstrated that many claims have been made in regards to PS, which would imply that PS *is* an important cognitive ability to explore in terms of designing efficient search systems that accommodate the individual user. However, with some doubt about the reliability and validity of PS tests (from earlier Chapter 4), in addition to a large mixture of both significant and non-significant contrasting PS results being claimed in the present chapter, it was thought necessary to critically evaluate the concept of PS, in order to understand how the differing results could be explained. For example, perhaps significant results always occurred on papers that administered one type of PS test, such as *Finding A's*, in comparison to papers that used a different test, such as *Number Comparison*. If this were the case, then this might shed light on possible best practice for future research with PS. Consequently, for various factors of PS that could be compared between papers—including the theoretical framework, search tasks, definition, measurement, scores, sample, and analysis undertaken—these were observed

for whether any obvious explanation for results emerged, in order to answer the third research question of **'How can the results be explained?'**.

5.5.1 Different Search Tasks

Overall, a large variety of search tasks were studied, including papers which required participants to:

- interpret a visualisation (14 papers: [56,68,94,137,138,206,209,210,211,218,219,220,221,222]);
- find a relevant document (7 papers: [4,9,11,12,13,18,125]);
- find a specific answer (7 papers: [10,46,71,103,174,204,205]);
- choose a preference (5 papers: [64,66,67,139,140]);
- complete a survey (3 papers: [119,178,243]);
- analyze information (3 papers: [18,43,46]);
- create something new (2 papers: [46,225]);
- browsing and playing games (1 paper: [154]);
- or engage in human-robot interaction (1 paper: [84]). [1]

With such different search tasks examined, it was speculated whether one task, over another, may have produced more, or less significant results, and could thus explain the overall discrepancy in findings. Therefore, every group of search task and dependent variable category (*Performance, Experience, Time, Behaviour*, and *Physiology*) were observed for any apparent differences, to which none were found. For example, in the *Performance* category, both significant and non-significant results were found across different tasks, such as finding a relevant document (e.g. significant: [11] ; non-significant: [4]); and a visualisation task (e.g. significant: [56] ; non-significant: [220]). For the other variables, no patterns were similarly identified, but to avoid repetition, details and examples are not reported here. For the full table of results, please refer to Appendix A.B. Thus, despite the differences in tasks analysed, this did not appear to account for the overall findings, or indicate that one task produced more reliable findings than another.

[1] Please note, the number of papers equates to more than 40, because some papers administered multiple search tasks in their user study.

5.5.2 Perceptual Speed definitions

Whilst research into PS may have considered many different search tasks, it should be thought that the definition of PS would still remain constant across all studies. However, looking across the papers in the sample, many interesting themes emerged regarding how PS had been defined, involving how: many papers did not provide a definition; references for other definitions were lacking; and the source of definition varied greatly, to the extent that mistakes were evident. Consequently, it was speculated whether the varying definitions could have been impacting the overall results, or alternatively, could be indicative of more, or less, reliable studies, to which the following subsections explored further.

5.5.2.1 No definition provided

After extracting the definition of PS given from each paper, it appeared that 7 out of the 40 papers did *not* contain any form of definition. As one of these papers [221] was from the same user study as other papers that did provide a definition [209, 210, 211, 222], the definition here could be deduced. However for the remaining 6 other papers that incorporated PS in their work ([4,43,119,178,205,243], this was not defined in any way. As some of these papers had provided overall effects of PS, it was speculated whether providing no definition may have indicated a poorer study, and perhaps resulted in more non-significant results found. However, this did not appear to be the case, as the papers without definitions resulted in both significant (e.g. [4]), and non-significant findings (e.g. [43]), and papers *with* a definition also found non-significant results (e.g. [84]).

On further inspection, for some papers, PS appeared in the method, without any prior discussion or explanation as to why it was analyzed. For example, in a 27 page peer-reviewed journal article, the participant section contained a table of user characteristics, of which PS was among them, and numbers were provided for whether this differed between younger participants, older adults who were familiar with technology, and older adults who were less familiar with technology [178]. However, this table was the first time that PS was mentioned in this paper, and therefore it appeared from nowhere. Although the results section from this study was included in the Interaction section (see Section 5.4.2.1)—and significant interactions had occurred—given the lack of definition for the concept, the ability to assess whether the researchers had operationalised the concept reliably remains questionable. Consequently, an assessment of the paper's worth and results becomes more difficult to achieve.

5.5.2.2 Different Sources of Definition

Whilst 33 papers *did* provide a definition of PS, and therefore automatically it might be assumed that they could be more reliable and represent better scientific practice, 15 of these papers did not provide any form of reference or source to their definition. This could be problematic, as

despite PS being one concept, it is possible that there were slight differences in how it was operationalised between subject areas. For example, one of the papers in the present review provided a specific definition, where PS was "*the cognitive ability to quickly and accurately find target information in literal, digital, or figural forms*" (French et al. 1963, cited in [103]). However, in other research areas, such as Psychology, PS has been defined very abstractly, where PS involved "*encoding, perception, central processing and response*" [95]. Therefore, if a definition was not referenced in a paper in the current review, it is possible that the definition could have been sourced from a different field such as Psychology. For example, one paper defined PS as: "*the speed in finding a known pattern in a visual field*" [84]. Whether this was inspired by the psychological processes of encoding or central processing, as done in Psychology, or whether the 'pattern' was referring to a literal, digital, or figural comparison, as had been done previously in Computer Science, remains unspecified. With potential differences in the conceptual representation of PS, this would imply that the results found in Computer Science must be interpreted with caution, as although the same concept has been supposedly investigated, differences in how the concept has been operationalised could be impacting results.

However, for 12 papers out of the 15 that did not provide a definition source, the source was able to be deduced as a result of comparison to other papers through this Systematic Review. For example, despite no citation, both Allen [9] and Conati & Maclaren [68] used an identical definition, word for word, as was defined and cited in seven other papers: "*speed in comparing figures or symbols, scanning to find figures or symbols, or carrying out other very simple tasks involving visual perception*" [11, 13, 18, 67, 206, 219, 220]. Similarly, nine other papers which did not reference their PS definition ([56, 64, 138, 139, 140, 209, 210, 211, 222]) all directly copied the same definition used in [218] and [66], which *was* referenced. Furthermore, although another paper with no reference was not *identical* to any other paper [125], it was a *shorter* version of a definition that had been defined and cited elsewhere [11, 18, 206]. For a full categorisation of the different definitions, and associated articles and sources, please refer to Table 5.14 (which can be found at the end of this chapter, given that it occupies one page). Consequently, although it might have been assumed that non-cited definitions meant the concept was more difficult to interpret, categorisation of the definitions in the present review was able to identify that the definition sources were similar, across all papers—except for just a small sample of 3 papers—which overall improves the reliability of results found.

5.5.2.3 Incorrect Sources of Definition

As just discussed, many papers shared an identical definition. However, even for papers which did provide a definition source, for the same identical definition, the claimed source differed between publications. Using the example of Allen [13], the reference for their definition came from the Ekstrom Manual, published in 1976. This was also the source provided in 6 other papers that used this definition [11, 18, 67, 206, 219, 220]. Having checked this original source

(specifically, [83]), it was verified that this was the correct reference, as the exact definition could be found on page 123. However, for another two papers that quoted the exact same definition ([154, 225]), the reference provided for this definition came from elsewhere. Specifically, they referred to the Ekstrom Kit, also published in 1976 [82]. Yet, after checking this source, no definition of PS was mentioned. Thus, the reference to the definition in these two papers ([154, 225]) was incorrect. Given that Turpin et al.'s [225] research provided 13 separate results regarding PS, and yet there were evidently mistakes in how the concept had been referenced, this raises doubt as to the reliability of this overall study, and accordingly the results found.

Similarly, another definition that was used by thirteen papers, ("*a measure of speed when performing simple perceptual tasks*"), appeared to be a shorter version of the original Ekstrom Manual definition. Yet, only two of the thirteen papers directed their reference to the Ekstrom Manual as their source [66, 218]. While nine of the papers did not provide any reference for this definition, a further two papers cited a completely different source: the Ekstrom *Manual for Kit*, published in 1996, 20 years after the Ekstrom manual. It could be possible that the 1996 version was reissued from the 1976 Manual, however, even after contacting a faculty librarian, the 1996 version was unable to be found anywhere. Therefore, the reference to 1996 is likely to be a misprint, and should in-fact refer to 1976. This either demonstrates another example of errors in the field, or, if the 1996 version is genuine, then the field has incorporated references that are inaccessible to other researchers, which may explain why mistakes have occurred in citations.

5.5.2.4 Summary of definitions

Ultimately, the varying definitions and sources between papers indicate the differences in how PS has been conceptualised in Computer Science. Whilst it is evident that there are discrepancies, this Systematic Review has allowed a thorough categorisation of the differences. Nonetheless, as mistakes in referencing definitions have been identified, this adds another layer to the already known mistakes in referencing performance effects (See Section 5.2). This ultimately emphasises the importance of treating the claims of PS found with caution, before a clearer understanding of results can be achieved.

5.5.3 Measuring Perceptual Speed

In addition to considering how PS had been defined, many differences also emerged regarding how PS had been measured, including: the type of test; the quantity of tests administered; and the format of test administration (paper or digitally). Consequently, these areas were examined for whether any patterns emerged that could explain the overall results, such as whether one PS test appeared to be associated with more significant findings, in comparison to another test, or whether different forms of measurement could be impacting knowledge about PS test reliability or validity.

5.5.3.1 Different Perceptual Speed Tests

Similar to the lack of PS definitions provided in many studies, 12/40 papers also did not provide any information for how PS was measured. Therefore, for 30% of papers which claimed to investigate some aspect of PS, it is unknown how this was achieved. However, for the other papers, the majority utilised a combination of the same three tests of PS, developed by Ekstrom: *Finding A's*; *Number Comparison*; and *Identical Pictures*. For example, *Number Comparison* was administered in the first paper, published in 1992 by Seagull & Walker [204], and also in one of the last papers, published in 2020 by Naghib et al. [174].

For just 2 papers in the sample, a different test altogether was administered: Digit Symbol Substitution, created by Wechsler [178, 205]. Although the papers provided no description of these tests, a recent different publication clearly described what was involved: "*Development of the Digit Symbol Substitution Test (DSST) was initiated over a century ago...The DSST is a paper-and-pencil cognitive test presented on a single sheet of paper that requires a subject to match symbols to numbers according to a key located on the top of the page. The subject copies the symbol into spaces below a row of numbers. The number of correct symbols within the allowed time, usually 90 to 120 seconds, constitutes the score*" [112].

It is therefore apparent that the Digit Symbol test was a very different test in comparison to *Finding A's*, *Number Comparison*, and *Identical Pictures*- all of which required the user to select a target, rather than selecting *and* reproducing a target. Consequently, whether these different tests were measuring the same aspect of PS remains unknown. This then creates the question as to whether the PS claims are comparable between studies. For example, a significant interaction was reported in O'Brien et al.'s research [178], which claimed differences in PS based on age after using the Digit Symbol Substitution Test. In contrast, Crabb & Hanson's research [70], who administered the *Number Comparison* PS test, claimed no interactions with age and PS.

Nonetheless, for each paper that provided an overall claim regarding PS, each result was inspected for whether an individual, or specific combination of PS test used related with the results found. As can be seen from Table 5.10, a trend did appear that if the combination of *Identical Pictures* and *Number Comparison* was administered (IP + NC), then significant results always occurred. In contrast, when only *Identical Pictures* (IP) or *Number Comparison* (NC) was used, non-significant results also occurred. This therefore might imply that using a combination of tests may be optimal for the most meaningful results to be obtained.

5.5.3.2 Perceptual Speed Quantity

In addition to Table 5.10 demonstrating that a combination of PS tests may be associated with more significant results, this would also correspond with the original guidelines by Ekstrom which stated that "*It is strongly recommended that researchers use more than one of these tests in any exploratory endeavour that aims at identifying a factor*" [83]. However, Table 5.11

Table 5.10: The different combinations of Perceptual Speed tests used, and how many returned significant, non-significant, or unknown results. IP = *Identical Pictures*, NC = *Number Comparison*, and FA = *Finding A's*.

	IP	**IP+NC**	**NC**	**FA**	**All 3**	**Unknown**
Sig	1	4	1	0	0	2
Both	0	0	2	3	1	1
Non-sig	1	0	1	0	0	0
Unknown	1	0	0	0	0	1

shows that only 7/40 papers followed this guidance, where a mixture of Ekstrom's PS test combinations occurred. For the other papers which used Ekstrom's tests, 19 of these only administered one test. As was previously identified in Chapter 4 (Specifically, Section 4.5.2), it is questionable whether these papers are fully valid.

Table 5.11: The type, or combination, of Perceptual Speed test(s) used in all 40 papers in the sample.

Perceptual Speed Test Used	**Paper**
Identical Pictures	[125, 154, 209, 210, 221, 222] [67, 138, 140, 220]
Number Comparison	[43, 46, 71, 174, 204, 243]
Finding A's	[18, 84, 225]
Identical Pictures and Number Comparison	[9, 10, 11, 12, 13]
Number Comparison and Finding A's	[103]
Identical Pictures, Number Comparison, and Finding A's	[4]
Digit Symbol Substitution	[178, 205]
Unknown	[56, 68, 94, 119, 211, 218] [64, 66, 137, 139, 206, 219]

Yet, even when some papers administered more than one PS test, the results from each test were treated as separate entities. For example in Allen [10] who administered both *Identical Pictures* and *Number Comparison*, the following was reported: "*Although analysis showed that the two perceptual speed tests achieved moderate reliability (Cronbach's cx = .69), it seemed likely that the two tests were assessing somewhat different aspects of perceptual speed. Accordingly, they were included in the analysis as separate independent variables.*" In Allen's other papers, both *Number Comparison* and *Identical Pictures* were also analysed separately, albeit in different orders. For example, in both Allen (1992, [9]) and Allen (2000, [11]), it was observed that only *Number Comparison* showed significant results compared against their dependent variables, and *Identical Pictures* were thus not given results listed. Yet in Allen (2000, [13]), they reported that: "*To avoid ambiguity, scores on the number comparison test were not used in the analyses reported here*". Consequently, it appeared that even if multiple tests were used, the results from one were sometimes disregarded.

Despite the lack of consistency in PS testing in the current sample of papers, significant results were still found in papers that used just one PS test (e.g. [174]). This would therefore

imply that although multiple tests may be the most valid measurement of PS, one PS test can still provide insightful findings.

5.5.3.3 Perceptual Speed Administration

Another point to note that differs between papers in how PS was measured, concerns the format of administration. Again identified in Chapter 4 (Section 4.6.1), whether a test was administered on paper, or digitally, could impact results. Yet, only 12/40 papers reported the administration method, of which 11 stated it was conducted on paper. However as one paper by Haapalainen et al. stated it was a computerised test [103], it is very possible that many other papers that did not report their format also administered computerised tests. Unfortunately, without contacting the original authors, which remains difficult when some papers date back to 1992, this information will remain an unknown. Thus, it is important to be aware that without this information, these differences in PS test administration could have impacted the overall results found.

5.5.4 Perceptual Speed Scores

Regardless of the different tests or formats used to measure PS, it might be expected that the PS scores would be reported in each paper, to enable a clear categorisation of what defined a Low-PS, or High-PS user. However despite PS being a main part of every paper in the current sample of 40 papers, more than half of these papers (22) did not report any PS score, which corresponded with Chapter 4 (See Section 4.5.2). Given the dichotomy of significant results found—such as in the Performance category, around half of the papers reported significant findings, and the other half non-significant findings (See Table 5.5 in Section 5.4.1.1)—whether these results could be explained based upon the PS score was considered. For example, it was hypothesised whether the papers that reported the PS score were the papers that found significant findings. However, after comparing all results for whether significant claims were associated with reporting an exact score, no obvious patterns emerged. For example, out of the 7 papers that found entirely significant results, 5 of these did not report any kind of PS score. Nonetheless, there were further aspects of PS scores that may have been impacting the reliability of overall results to consider, including: the various types of scores reported; the different ranges of score available for the supposed same test; and overall thresholds of what defined Low-PS and High-PS—all of which are discussed below.

5.5.4.1 Different dimensions of scores

For the papers that did report PS scores (See Table 5.12), this demonstrated that different studies reported different areas of results. For example, some reported the *mean* PS score, whilst others reported the *median*. Furthermore, although many studies reported a *minimum* PS score, it was unclear as to whether this was the minimum score possible on the test, or

Table 5.12: The Perceptual Speed (PS) scores reported in each test, where IP = *Identical Pictures*, NC = *Number Comparison*, FA = *Finding A's*, DS = *Digit Symbol Substitution*, L = Low-PS, and H = High-PS.

PS Test	Identifier	Min	Max	Out of	Mean	Median
IP	Mahmud et al. [154]			96	51	
	Toker et al. [221]	54	96		85.7	
	Lalle et al. [138]			60	46.5	
	Toker et al. [220]	25	66	72	45.2	
	Conati et al. [67]	L:13, H:41.	L: 40, H:63.		L:32.2, H:47.1.	
IP & NC	Allen [9]				IP: 80.9, NC: 30.1.	
NC	Brennan et al. [46]	25	73	96	44.38	44
	Brauner et al. [43]	18	39		24.7	
	Crabb & Hanson [71]				Young: 46.63, Old: 45.08.	
FA	Turpin et al. [225]	34	74	200	51.94	51
	Arguello & Choi [18]	44	90	200	64.16	
DS	O'brien et al. [178]				Young: 71.5, Old (high-tech): 56.8, Old (low-tech): 45.	
Not given	Carenini et al. [56]	31	63	96	45.37	
	Conati et al. [64]	25	67		46.7	
	Lalle & Conati [137]	24	66		45	
	Sheidin et al. [206]	17	53		35.3	

the minimum score achieved in that particular sample. Consequently, with different studies reporting different dimensions, this makes an overall comparison between studies difficult.

5.5.4.2 Different scores within the same test

As can be seen from Table 5.12, in addition to different score ranges *between* tests, there also appeared to be different score ranges possible *within* the same test. For example, Mahmud et al. [154] stated that in *Identical Pictures*, the maximum score that could be achieved was 96. However, in Toker et al. [220] who also used *Identical Pictures*, whilst their highest scoring participant achieved 66, this was apparently out of a maximum of 72. Thus, it appeared that although some studies administered the 'same' PS test, there were still differences between them. It is unknown why this was, especially when Toker et al. [220] cited Ekstrom, and when referring to this original source, the maximum score that could have been achieved should have been 96. Therefore, it is possible that variations of Ekstrom's tests were used. One explanation may relate to the fact that some more recent studies may have computerised the test, as discussed in Section 5.5.3.3, without explicitly stating this. This could further explain why such different scores were achieved in each test, for example, the mean score in *Number Comparison* was 24.7 for Brauner et al. [43] and yet almost double at 46.63 in Crabb & Hanson [71]. Regardless, without further understanding of these differences, it is difficult to interpret the overall claims of PS, with so many ranges of the same concept being possible.

5.5.4.3 Perceptual Speed Thresholds

As insightful meaning of PS scores cannot be generated from comparisons between studies, this reaffirms the importance of Chapter 4 (Section 4.5.1), which explained how there were no thresholds of what defines a Low-PS and High-PS user. In addition to no generic standardised thresholds of PS, there was also unknown as to the thresholds used within each study. For example, in Arguello & Choi's research [18] that stated "*We decided to group participants into Low and High cognitive ability groups using a median split*", the median PS score in that study was not reported. This lack of data was not unique to [18]: as can be seen from Table 5.12, overall, there were many cells with unknown figures. With so many unknowns, understanding whether how participants were grouped affected overall results was not possible to thoroughly examine.

Nonetheless, as Arguello & Choi [18] explained their participant division into Low-PS and High-PS users was achieved using a median split, many other papers also used this method (including [4, 46, 67, 209, 218, 221]). Whilst many papers did not report how they classified users, other papers used very different methods. For example, Carenini et al. [56] divided users into three types of PS: Low, Average, and High. This was achieved using ranges, where "*Low represents the bottom quartile of the values distribution (i.e. lower 25%), average represents the values within the interquartile range (i.e., middle 50%), and high represents the upper quartile (top 25%)*". Conati et al. [64] similarly followed this method of three types of PS being calculated.

While the median split and quartile ranges were possible to quantify when analysing data from one PS test, as identified in Section 5.5.3.2, some papers utilised multiple PS tests. Thus, different ways to categorise users here was conducted. In Al-Maskari & Sanderson [4], a measure was formulated called "*Overall Perceptual Speed*", which incorporated scores from all three PS tests including *Finding A's*, *Number Comparison*, and *Identical Pictures*. However, the details for how scores were computed was unknown. As other papers using multiple PS tests decided to treat their results as separate entities (*as described in Section* 5.5.3.2), there appeared to have been no consistency for why one PS test was chosen to group users by, in comparison to another test. Therefore, to properly compare PS scores and have a clearer picture of what defines someone as having either Low-PS or High-PS, more data and consistency is ultimately required. However given the evident differences present, this further exemplified the importance of being cautious with the overall PS claims, when PS has been measured so differently across studies.

5.5.5 Perceptual Speed Sample

Another factor to consider that may have impacted the overall results, especially given the uncertainty in PS thresholds, concerns the number of participants studied in each sample. It

Table 5.13: The number of participants in every study which found significant (sig), non-significant, both, or unknown results, depending on the dependent variable investigated.

Category	Sig results only	Both sig and non-sig results	Non-sig results only	Results unknown
Performance	50 [9], 62 [56], 80 [13], 85 [174], 99 [64], 100 [11].	n/a	20 [43], 39 [84], 56 [4], 56 [220]	166 [139].
Experience	16 [225], 21 [46], 39 [84].	20 [71].	32 [18], 56 [4], 56 [220], 99 [64]	35 [221], 50 [9]
Time	21 [46], 56 [4] , 85 [174], 100 [11].	n/a	16 [225], 20 [43], 32 [18] , 56 [220]	n/a
Behaviour	80 [12], 85 [174].	21 [46], 32 [18]	16 [225].	50 [9].
Physiology	35 [222].	n/a	n/a	n/a
Interaction	20 [43], 30 [178], 32 [18], 35 [222], 45 [68], 62 [56], 62 [218], 80 [12], 80 [13], 97 [243]	16 [225], 35 [221], 99 [64]	20 [71], 77 [125], 80 [10], 100 [11]	21 [46]
Predicting PS	n/a	n/a	n/a	20 [103], 35 [210], 40 [205], 62 [94], 95 [140], 166 [66]
No results provided	n/a	n/a	n/a	12 [154], 40 [206], 44 [204], 46 [137], 56 [219]
Sample not provided	n/a	n/a	n/a	[119], [211], [138]

is well known in research that if more participants are used, then the results may be more powerful [49]. Subsequently, it might be assumed that if there was a discrepancy between significant and non-significant results found across studies investigating PS, then the significant results may have occurred in papers with larger sample sizes. Therefore, for each dependent variable category, the papers that found significant findings, both significant and non-significant findings, non-significant findings, and unknown, were categorised by the number of participants that contributed towards the results (*See Table* 5.13).

Although three studies did not provide details for their number of participants in their study [119, 138, 211], every other paper in the sample did. Consequently, looking across all papers in Table 5.13, the lowest number of participants used was 12 [154], and the highest was 166. There is therefore a large range in the generalisability of results from each study, which must be taken in account when considering the reliability of results.

Nonetheless, while the non-significant column did have some studies with smaller numbers of participants (such as 16 and 20), this was not a prerequisite for significance, as studies with sample sizes of 16 and 21 still occurred in the significant-only column. Additionally, for the *Performance*, *Experience*, and *Time* categories, there were always studies that had more participants in the non-significant column, in comparison to the significant column. For example, focusing on *Performance*, significant results occurred with a sample size of 50, and yet non-significant results occurred with a sample size of 56. Consequently, the sample size of each study did not appear to be a main explanation for the diverse results found.

However, a further difference was identified regarding the number of participants in the same user study, which was reported across different papers. It should be assumed that if multiple papers were using data from the same user study, then the sample of participants listed should be the same. Generally, this was the case. For example, in Toker et al.'s 2018 paper that explained there were 56 participants [219], Toker et al.'s 2019 paper also stated

56 participants [220]. Similarly, in Conati et al.'s 2017 paper [66] and Conati et al.'s 2020 paper [67], 166 participants were reported. However, in Lalle et al.'s 2015 paper [140], which used the same data as was reported in 2014, in Conati et al. [64], a slightly different number of participants were reported (Lalle et al. [140]: 95; Conati et al. [64]: 99). Yet, both reported the same age range (16-40 years old). It is unknown why Lalle et al. removed 4 participants before their secondary analysis was conducted, as no explanation was provided. Regardless, this lack of consistency, or detailed explanation, creates a sense of doubt to the reader as to whether the original results were reliable or not. Otherwise, it is not clear why some participants would have been removed.

Furthermore, irrespective of the overall sample size in each paper, a final factor to consider regarding PS sample size is the size of each group within a sample. For example, in Brauner et al.'s research [43], which claimed an interaction that women had a significantly higher PS score in comparison to men (*See Section* 5.4.2), this was based upon 20 participants. Yet, only 8 of these were female. As Brauner et al. [43] identified that PS did not influence search accuracy, or search time overall when completing a simulated business game (*See Section* 5.4), it could be hypothesised that if more participants in the sample had been used, such as men who supposedly had a lower PS, then a significant result may have occurred. Then, perhaps other studies that found more significant results occurred with a sample that consisted of more men, where a bigger difference in Low-PS and High-PS could be found. However, as no other study in the sample considered sex as a confounding variable, this hypothesis is not grounded in multiple data points. Thus, this demonstrates the importance of future research experimenting with a larger sample of participants, before conclusive results can be obtained as to whether other variables can explain the overall results found.

5.5.6 Perceptual Speed Analysis

With such varying numbers of participants involved in each study, alongside the fact that every paper investigated different search tasks and other variables, it is unsurprising that the types of method used to analyse results also varied considerably between papers. Overall, it would not be appropriate to compare like with like, such as the results from an ANOVA, in comparison to the results from multilevel linear modelling. However, a general observation could be conducted, to check whether all papers that used one type of method, such as an ANOVA, found significant results, in comparison to another method, such as Regression, always finding non-significant results.

While some papers did not report what kind of statistical analysis method was conducted [137, 139, 219], others reported that they used a 'General Linear Model' [10, 11, 56, 68, 221]. As this term is an umbrella term, this could refer to different statistical tests, such as an ANOVA *or* a Regression analysis [80]. Consequently, creating an accurate tally of what tests

were administered across different studies cannot be precise, and some studies additionally used multiple methods.

Nonetheless, as a vague idea, some kind of regression or mixed model analysis appeared to be the most common statistical method, with 13 papers using this [9, 43, 64, 71, 84, 154, 174, 205, 206, 218, 220, 221, 243]. The second most popular analysis method was then an ANOVA, which was used in 8 papers [12, 13, 43, 46, 125, 178, 204, 225]. Less common statistical analysis methods used were then Correlation [11, 119, 204, 243], Principal Component Analysis [18, 222], and the Mann-Whitney [4] and Chi-square test [211].

However, after analysing the different and same methods used across studies, this did not appear to account for the overall differences in results found. For example, when considering the *Performance* category, in one of Allen's papers that utilised an ANOVA, significant results were found that stated High-PS users were significantly more accurate in their searches [13]. In contrast, Brauner et al.'s paper that also utilised an ANOVA found non-significant results, which stated accuracy was not influenced by PS [43]. The other categories (*Experience, Time, Behaviour*, and *Physiology*) were also examined, but similarly, no trends emerged. Consequently, the analysis method also did not appear to explain the overall differences in significant and non-significant results found.

5.5.7 Comparing different variables together

It is important to note, that through evaluating all papers, regardless of the PS definitions used, or analysis methods implemented, the significant and non-significant findings were aggregated to provide a comprehensive classification of all claims found. Therefore, although one study may have claimed something, just because another study claimed the opposite, this may have been because a completely different setup was administered, such as a different visualisation. For example, in Toker et al.'s research [221] that claimed Low-PS users struggled with Radar graphs and took significantly longer completing search tasks using this interface, Conati & Maclaren [68] claimed that Low-PS users performed searches with *higher* accuracy when using *Radar graphs*. Yet, Conati & Maclaren did not compare Radar graphs to Bar graphs, as Toker had done. Thus, it is possible that if *Bar graphs* had been used in Conati & Maclaren's research, then this may have resulted in better performance using *Radar graphs* for Low-PS users. Alternatively, although Toker et al. identified that Low-PS users took longer completing tasks with the *Radar* graph, this was not explicitly compared against search performance. Therefore, it is also possible that taking longer completing a search was not a bad thing, but in fact could result in higher search accuracy, if more processing had occurred.

With many papers that measured search performance not also measuring search time (e.g. Brauner et al. [43]), and vice versa, papers that measured search time did not measure performance (e.g. Brennan et al. [46]), it is difficult to combine results together to understand the full picture of how PS affects a user. This assimilates to the other variables investigated, such

as user experience, where a significant difference between Low-PS and High-PS users may have been reported, but there was no comparison as to whether a lower user experience corresponded to a lower task accuracy (e.g. Arguello & Choi [18]), or whether a lower task accuracy resulted in no subjective awareness of this difference (e.g. Conati et al. [64]). Consequently, more research is needed to compare all different variables together simultaneously, such as search time *and* performance *and* user experience, across many different variations of tasks, in order to have conclusive understandings as to how PS truly affects a user.

5.5.8 Other potential factors impacting results

Although the factors investigated above resulted in no *clear* explanation of results, such as to why many significant and non-significant contrasting results occurred, other explanations were still possible. As identified in Section 5.5.7, every paper conducted a unique study. Whilst it may be difficult to merge together different results from different studies in order to find consistent explanations, having become familiar with the papers, an observation emerged that may explain some results: if one paper reported no differences between Low-PS and High-PS users overall, such as similar overall search times (e.g. Turpin et al. [225]), but another paper did identify differences, such as Low-PS taking longer (e.g. Brennan et al. [46]), or Low-PS being faster (E.g. [18]) then a particular interaction may have been occurring.

In particular, in Turpin et al.'s research [225], overall, no differences in search time were reported between Low-PS and High-PS users. However, although High-PS were unaffected by different interfaces, Low-PS users spent significantly longer completing tasks with the 'Blended' interface, and significantly less time completing tasks with the 'Non-blended' interface (See Section 5.4.2.3). Therefore, the two different conditions balanced out the overall effect of PS. Consequently, if other research claimed results that Low-PS took less time completing tasks overall, then perhaps this interface assimilated to the 'Non-blended' interface where less visual information was present, whereas different research that claimed that Low-PS took longer, may be a more similar setup to a 'Blended' interface, where lots of verticals had to be processed. Indeed, this did appear to be a possibility. For example, in Arguello & Choi's research [18] which claimed that Low-PS were slightly faster completing search tasks, a 'Blocked' condition had been implemented, which clearly separated visual elements into separate visual areas for the user to process, potentially similar to Turpin et al.'s 'Non-blended' interface. Then, in Brennan et al.'s [46] research, the open web was used, where lots of visual information in the form of verticals would have all been present and mixed into the same screen, which may visually have looked similar to Turpin's 'Blended' interface.

Therefore, one possible explanation for the overall contrasting results found relates to the visual setup of information a user must process. This would make sense, as the overall definition of PS is that a user with Low-PS takes longer to process information, and yet is still ultimately less accurate. Thus, if there was more visual clutter to process (as was described earlier in the

Literature Review, Chapter 2, Section 2.7), then a user with Low-PS may perform less efficient searches, through either taking longer to complete the search in say, a 'Blended' interface, versus being more efficient in a 'Non-blended' or 'Blocked' condition. Looking across the other papers in the sample, with the specific focus of identifying visual differences, the amount of visual clutter did appear to differ between interfaces investigated. For example, looking again at the images used in Conati & Maclaren's research [68] (See Figure 5.5), where Low-PS were more accurate with a radar-graph, this interface subjectively looked less cluttered and more organised, than the interface with colored boxes scattered all over the screen. Similarly, going back to the hierarchical database structures examined in the 1990s, such as Allen [12], Low-PS learned more using a *multiwindow display*, where the information was clearly divided and less visually cluttered than the word map condition (See Figure 5.7). Consequently, future research should further investigate the role of visual clutter and PS.

5.5.9 Summary of Evaluation

While there was some doubt about the reliability and validity of the overall claims and interactions reported regarding PS in Computer Science, there was nothing identified that could clearly explain the inconsistency in results found. For example, even if a paper was not grounded by a theoretical framework, did not define PS or explain how it was measured, or failed to report precise scores, significant findings still emerged that claimed PS did affect a user's ability to search for and retrieve information, across a variety of tasks relevant to Computer Science. Although a significant result does not necessarily mean the results can be trusted—especially when a smaller sample size reduces the generalisability of results—with *many* papers finding significant results, this would imply that PS was still an important concept to consider when designing future computer systems. Furthermore, having become familiar with the literature—alongside prior knowledge on visual perception more generally—a possible explanation emerged that could explain the inconsistent results: if more visual clutter was present during the search task, then Low-PS users appeared to be less efficient in their search outcomes; if the interface was more organised, then Low-PS were helped to perform similarly to, or equally with, High-PS users. However, as this is only one indication, further research would be needed to clarify this hypothesis.

5.6 Systematic Review Conclusion

Overall, this Systematic Review was motivated based on some uncertainty with how PS had been studied within the field of Computer Science. As there was some confusion as to whether PS did, or did not, affect users completing various search tasks—given that various papers had inaccurately cited previous conclusions, results relating to PS had completely disappeared, and the PS tests posed various reliability and validity concerns—it was thought necessary to

categorise every claim that existed regarding PS in the area, whilst attempting to explain the results through additionally being cautious of any reliability or validity concerns.

In response to the first research question, ***'What claims have occurred regarding PS in Computer Science?'***, all findings were categorised into how PS had been reported to affect a user's search performance, behaviour, time, user experience, physiology, and interactions with other variables. Firstly considering physiology, which was the only area where consistent significant results had occurred, it was identified that Low-PS users had a lower fixation-rate, fixated on different areas of the screen, and took longer processing certain elements. If users are fundamentally looking at and processing different areas of the screen, it is unsurprising that there were also many differences reported in user experience, behaviour, time, and search performance for Low-PS and High-PS users.

Although many differences reported would reaffirm that PS is an important aspect to consider when designing interactive systems, that can accommodate each individual user's unique abilities, many contrasting findings were also identified. For example, while different studies individually claimed that Low-PS users took longer searching, clicked on more things, and reported higher workload, other studies claimed the opposite. However, many studies did not report multiple variables together, such as whether higher workload corresponded with a longer time searching, or less time. Therefore, it remained unknown as to how the different variables related to each other, which would be an important area to explore in future research. Nonetheless, through clearly categorising everything together, future researchers will now be able to position their findings precisely in the literature, to begin to identify similarities and differences that can better understand the most optimal search environment for a user with Low-PS.

Nevertheless, given that many contrasting results were found across the literature, this highlighted the importance of considering explanations for the overall results, and answering the second research question of, ***'How can the results be explained?'***. Through this evaluation, some possible concerns of reliability or validity were also identified. For example, factors that may have reduced overall validity, such as different PS tests being administered, were explored for whether this appeared to impact the contrasting results found. However, through considering many different areas—such as the use of a theoretical framework, PS definitions, scores and measurements— no obvious patterns emerged where concerns of reliability or validity explained the overall findings. Nonetheless, these explorations have opened up further areas of future research, such as fully identifying the differences between different PS tests.

Finally, whilst it was acknowledged that every study investigated had implemented different contexts and search tasks, a possible explanation of contrasting results emerged: if there was more visual clutter visible—such as through less organised interfaces that contained many different elements, like videos and images simultaneously—then Low-PS appeared to struggle

and perform search tasks less efficiently than High-PS users. If this were the case, then intelligent systems that can accommodate individual users who struggle could be developed, such as through designing more organised and less cluttered interfaces. However, as this is only one hypothesis, more research is needed to verify whether visual clutter does, or does not, explain the overall differences in PS found, in order to provide consistent findings that can direct future designs to become the most optimal for each individual user.

5.7 Chapter 5 Summary

This Chapter presented the Systematic Review which documented all previous claims of PS in the field of Computer Science. This demonstrated that PS does have a significant impact on users' search tasks, with Low-PS users differing from High-PS users in terms of search performance, behaviour, experience, and physiology. Yet, there were also many contrasting findings between different studies, and thus the review also explored possible explanations for these differences. The chapter then concluded with a suggestion that designing more organized and less cluttered interfaces could accommodate users with Low-PS, and this highlighted the importance of further research that can direct future designs to become the most optimal for each individual user. However, in order to empirically investigate the effects of visual clutter on users with differing PS levels, how PS is tested and measured needed revision—in order to be more reliable and valid—and thus how new tests were developed is the focus of the following Chapter 6.

Table 5.14: Definitions of Perceptual Speed, by article that quoted them, and the source which was attributed.

Definition Given	Source of definition	Article in Review
None provided	N/A	[4, 43, 119, 178, 205, 221, 243]
"speed in comparing figures or symbols, scanning to find figures or symbols, or carrying out other very simple tasks involving visual perception".	Ekstrom Manual, 1976	[11, 13, 18, 67, 206, 219, 220]
	Ekstrom Kit, 1976	[154, 225]
	Not referenced	[9, 68]
"speed in comparing figures or symbols, or in searching through a visual field."	Not referenced	[125]
"speed in carrying out very simple tasks involving visual perception".	Ekstrom Manual, 1976	[10]
"a measure of speed when performing simple perceptual tasks".	Ekstrom Manual, 1976	[66, 218]
	Ekstrom Manual for Kit, 1996	[94, 137]
	Not referenced	[56, 209, 210, 211, 222] [64, 138, 139, 140]
"related to the speed with which productions can be implemented and compiled."	Ackerman, 1988	[204]
"the cognitive ability to quickly and accurately find target information in literal, digital or figural forms, make comparisons, and carry out other very simple tasks involving perception".	French et al. 1963	[103]
"a person's ability to efficiently view and identify differences and similarities, patterns, and anomalies when conducting tasks involving symbols and figures".	Carroll 1993	[46]
"It draws on the ability to scan information effectively, make choices for response and is said to be related to automatic mental processes".	Ekstrom 1973	[46]
"a measure of an individual's ability to search and compare visual symbols or patterns in rapid succession".	Horn 1991	[71]
"someone's ability to compare visual patterns or identify a visual pattern among distracting patterns".	Pawlik 1966	[18]
"related to the selection of scanning strategies".	Not referenced	[12]
"the speed in finding a known pattern in a visual field".	Not referenced	[84]
"the speed of finding the correct answer in the searching process."	Not referenced	[174]

[a] Please note, whilst there were 40 articles in the present review, there are 42 articles listed in this table. This is because two articles presented two definitions for PS: Arguello & Choi [18] and Brennan et al. [46].

Chapter 6

Updating Perceptual Speed Measurement

6.1 Chapter 6 Overview

This Chapter explains why PS measurement needed to be updated, and how two new digital PS tests were created. In addition to creating new tests, and providing the results of these tests from a sample of participants, different ways of analysing results are also discussed. The chapter then concludes with the most optimal analysis that should be conducted, and how this was implemented on the current participants is explained.

6.2 Introduction

In response to some of the points raised in the previous chapters, it became apparent that how PS was measured needed to be re-examined and updated, before the first main research question could be answered of "***How can a user with Low Perceptual Speed be helped to achieve a more positive online search experience, both subjectively, and objectively?***". Specifically, Chapter 4 (Section 4.6.1) explained that the movement towards more online experiments meant that the original paper-pen PS tests—that had been created over 50 years previously—were now outdated, and thus PS tests of the future needed to be administered online. Not only could this expand their use in more experiments—and thus more awareness of different results could be generated— but digital PS tests could also be useful for designing Human-Computer Interaction systems that can adapt to, and accommodate, each unique user, depending on their PS.

Furthermore, both Chapter 4 (Section 4.5.2) and Chapter 5 (Section 5.5.3.2) highlighted that many different PS tests had been administered in different studies, and despite the original guidelines stating that more than one test was needed for a valid measurement, many studies only administered one test. Consequently, whilst devising digital PS measurement, it is necessary to understand the difference between multiple tests, to explore how an overall measure of PS can be created. Additionally, given that Chapter 5 (Section 5.4.1.6) highlighted many

contrasting results found regarding how PS affects IIR, the creation of a clear digital PS measurement will act as a baseline test for future studies to improve overall consistency in the field. This led onto the following sub research question of "***How can a digital measurement of PS be created?***", which is discussed in the remaining chapter.

6.3 Selecting Perceptual Speed Tests

To create a digital measurement of PS, two tests were chosen, based on *Ekstrom's Kit of Factor Referenced Cognitive Tests* [82].

Ekstrom's tests motivated the current selection, given that these (*Finding A's, Number Comparison, and Identical Pictures*) were the most used tests in previous literature, as opposed to other tests, such as the Digit Symbol Substitution (See Section 5.5.3.1).

Only two tests were re-created, out of the three options provided in the original manual, because: (a) it would have been inconsistent with previous literature to administer three tests simultaneously on the same participant; and (b) an overall measure of PS was to be created, which was to be deduced based on scores from two tests.

Although *Identical Pictures* appeared to have been the most frequently used PS test in previous literature (See Table 5.11), there were many occasions where the results of this test were disregarded. For example, in both Allen (1992, [9]) and Allen (1994, [11]), although both *Identical Pictures* and *Number Comparison* had been administered, it was observed that only *Number Comparison* showed significant results compared against the dependent variables investigated, and thus results concerning *Identical Pictures* were disregarded (See Section 5.5.3.2).

Furthermore, although the results from *Number Comparison* had once been disregarded for unknown reasons (e.g. Allen [13] stated that: "*To avoid ambiguity, scores on the number comparison test were not used in the analyses reported here*", in Section 5.5.3.2), *Number Comparison* represented the PS test which had been used over the longest time frame. For example, it was administered in the first study investigated in the Systematic Review sample, published in 1992 by Seagull & Walker [204], and also in one of the last papers examined, published in 2020 by Naghib et al. [174] (See Section 5.5.3.1).

In contrast, despite *Finding A's* having been utilised less often than *Number Comparison* or *Identical Pictures* in previous literature, it had always been associated with some form of significant finding with a dependent variable investigated in a user study (e.g. [18, 84, 225], see Table 5.10). Therefore, it was believed that *Finding A's* may provide more insightful findings.

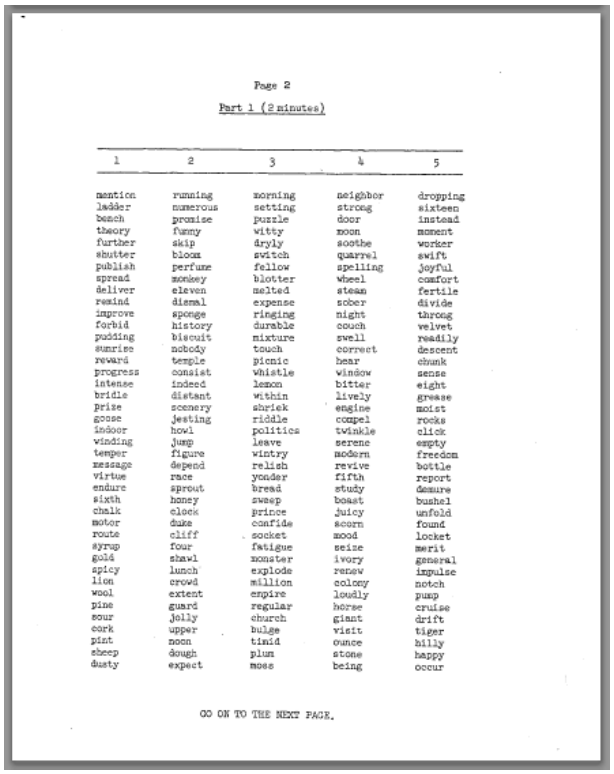
Consequently, *Identical Pictures* was not selected as a test to update in the present work, and instead *Finding A's* and *Number Comparison* were chosen.

6.4 Examining Previous Tests

Before a digital version of both *Finding A's* and *Number Comparison* was created, it was firstly necessary to examine the original paper-based versions.

6.4.1 Finding A's

For *Finding A's*, the original instructions in the Ekstrom Kit [82] stated the following: "*This is a test of your speed in finding the letter 'a' in words. Your task is to put a line through any such word. Listed below are five columns of words. Each column has five words containing the letter 'a'... You will have 2 minutes for each of the two parts of this test. Each part has four pages.*"



Page 2

Part 1 (2 minutes)

1	2	3	4	5
mention	running	morning	neighbor	dropping
ladder	numerous	setting	strong	sixteen
bench	promise	puzzle	door	instead
theory	funny	witty	moon	moment
further	skip	dryly	soothe	worker
shutter	bloom	switch	quarrel	swift
publish	perfume	fellow	spelling	joyful
spread	monkey	blotter	wheel	comfort
deliver	eleven	melted	steam	fertile
remind	dismal	expense	sober	divide
improve	sponge	ringing	night	throng
forbid	history	durable	couch	velvet
pudding	biscuit	mixture	swell	readily
sunrise	nobody	touch	correct	descent
reward	temple	picnic	bear	chunk
progress	consist	whistle	window	sense
intense	indeed	lemon	bitter	eight
bridle	distant	within	lively	grease
prize	scenery	shriek	engine	moist
goose	jesting	riddle	compel	rocks
indoor	howl	politics	twinkle	click
winding	jump	leave	serene	empty
temper	figure	wintry	modern	freedom
message	depend	relish	revive	bottle
virtue	race	yonder	fifth	report
endure	sprout	bread	study	desire
sixth	honey	sweep	boast	bushel
chalk	clock	prince	juicy	unfold
motor	duke	confide	scorn	found
route	cliff	socket	mood	locket
syrup	four	fatigue	seize	merit
gold	shawl	monster	ivory	general
spicy	lunch	explode	renew	impulse
lion	crowd	million	colony	notch
wool	extent	empire	loudly	pump
pine	guard	regular	horse	cruise
sour	jelly	church	giant	drift
cork	upper	bulge	visit	tiger
pint	noon	timid	ounce	hilly
sheep	dough	plum	stone	happy
dusty	expect	mass	being	occur

GO ON TO THE NEXT PAGE.

Figure 6.1: An original page from the paper-based *Finding A's* Perceptual Speed test. Although only one page is visible here, in the original test, there were eight pages. (Taken from: Ekstrom [82].)

Having inspected the test (See Figure 6.1), every page had 41 rows and 5 columns, equating to 205 words per page. As there were 5 words that contained the letter 'a' in every column, this meant that 25/205 words contained the letter 'a'. In other words, 12% of all stimuli were a target.

For Part 1, there were 4 pages to navigate. For Part 2, there were also 4 pages to navigate. Therefore, in the total time of four minutes, a maximum of 1,640 words (205 x 8) could have

Table 6.1: A sample of three columns in the *Finding A's* Perceptual Speed Test, showing how many words contained each number of letters and syllables.

		Frequency in column		
		1	2	3
Letters	4	7	1	1
	5	11	19	8
	6	11	12	12
	7	11	6	14
	8	1	3	6
		=41	=41	=41
Syllables	1	12	17	7
	2	28	21	30
	3	1	3	4
		=41	=41	=41

been viewed, and of these, 200 were targets. Consequently, the maximum score a user could achieve was 200.

6.4.1.1 Original Finding A's Stimuli

Out of the 40 columns overall (5 columns x 8 pages), three were randomly sampled to identify what kinds of words had been used. This showed that in every column examined, the number of letters ranged from 4 to 8, and number of syllables ranged from 1 to 3 (See Table 6.1). There did not appear to be any obvious pattern to these, such as a set number of words containing each letter, and therefore the words appeared randomly generated, in terms of letters and syllables.

The index of change for where the letter 'a' appeared was also examined. Except for never appearing as the first or last letter, this also appeared randomly dispersed throughout each word.

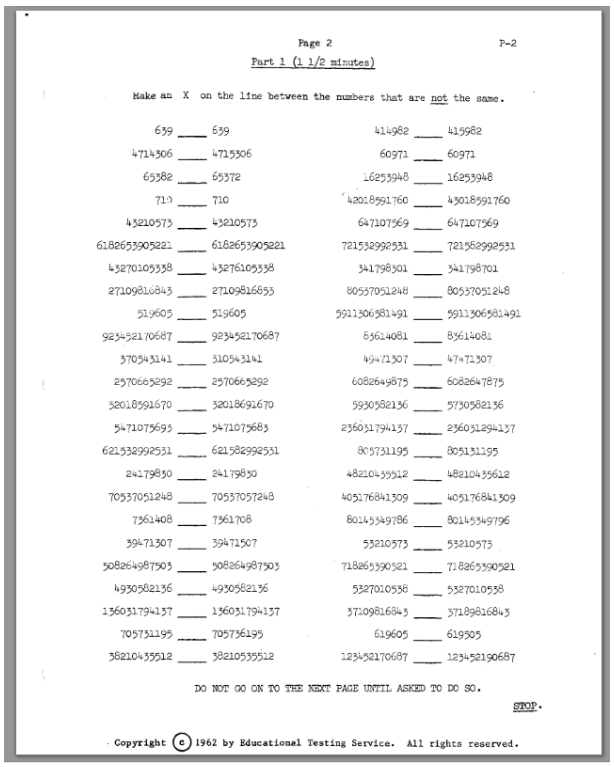
It was additionally observed that no plural or capitalised words were ever used. However, a mixture of adjectives (E.g. spicy, sour), nouns (E.g. lion, sunrise), and verbs (E.g. forbid, forgive) were used.

6.4.2 Number Comparison

For *Number Comparison*, the original instructions in the Ekstrom Kit [82] stated the following:

"*This is a test to find out how quickly you can compare two numbers and decide whether or not they are the same. If the numbers are the same, go on to the next pair, making no mark on the page. If the numbers are not the same, put an* `X` *on the line between them.... Your score will be the number marked correctly minus the number marked incorrectly. Therefore, it will not be to your advantage to guess unless you have some idea whether or not the numbers are the same. You will have 1 1/2 minutes for each of the two parts of this test. Each part has one page.*"

Having inspected the test (See Figure 6.2), every page had 24 rows of number pairs across 2 columns, equating to 48 pairs per page. For Part 1, there was 1 page to navigate. For Part



Page 2 P-2

Part 1 (1 1/2 minutes)

Make an X on the line between the numbers that are not the same.

639	____	639	414982	____	415982
4714306	____	4715306	60971	____	60971
65382	____	65372	16253948	____	16253948
710	____	710	430185917 60	____	430185917 60
43210573	____	43210573	647107569	____	647107569
61826539 05221	____	61826539 05221	721582992531	____	721582992531
43270105338	____	43276105338	341798301	____	341798701
27109816853	____	27109816853	805370512 48	____	805370512 48
519605	____	519605	591130658 1491	____	591130658 1491
923452170687	____	923452170687	83614081	____	83614081
370543141	____	310543141	49471307	____	47471307
2570665292	____	2570665292	6082649875	____	6082647875
320185916 70	____	320186916 70	5930582136	____	5730582136
5471075693	____	5471075683	2360512 94137	____	2360312 94137
621582992531	____	621582992531	8057311 95	____	805131195
24179830	____	24179830	482104 35512	____	482104 35612
705370512 48	____	705370572 48	405176841309	____	405176841309
7361408	____	7361708	801453 49786	____	801453 49796
39471307	____	39471507	53210573	____	53210573
508264987503	____	508264987503	718265390521	____	718265390521
4930582136	____	4930582136	5327010538	____	5327010538
1360317 94137	____	1360317 94137	371098168 43	____	371898168 43
705731195	____	705736195	619605	____	619505
382104 35512	____	382105 35512	123452170687	____	123452190687

DO NOT GO ON TO THE NEXT PAGE UNTIL ASKED TO DO SO.

STOP.

Figure 6.2: An original page from the paper-based *Number Comparison* Perceptual Speed test. Although only one page is visible here, in the original test, there were two pages. Source: Ekstrom [82].

2, there was also 1 page to navigate. Therefore, in the total time of three minutes (1 minute 30 seconds for each part), a maximum of 96 pairs (48 x 2) could have been viewed.

Unlike *Finding A's*, there was no indication as to how many pairs were non-identical in each column. Having manually checked this, it was apparent that 51 pairs were non-identical (See Table 6.2). Consequently, the maximum score a user could achieve was 51.

6.4.2.1 Original Number Comparison Stimuli

To gain a better understanding of the pairs of numbers used, every pair was examined for: 1) the length of numbers; and 2) the index of which number changed, and this information was also reported in Table 6.2. This shows that the shortest number length was 3 numbers, and the longest 13 numbers. There did not appear to be any obvious pattern to the length of numbers, such as having an equivalent amount of pairs with 7 numbers, or 8 numbers etc. However, having computed a histogram for every length of number possible (See Figure 6.3), there were less numbers of shorter length, in comparison to more numbers of longer length.

The index of change for where the numbers differed was also examined (See Table 6.2). Except for never appearing as the first number, this also appeared randomly dispersed throughout each pair. Except one pair which contained two different numbers, every other pair only differed by one digit.

Table 6.2: Detailed information for the stimuli present, and how the number pairs differed, in the original *Number Comparison* Test by Ekstrom [82].

	Part 1		Part 2	
	Column 1	**Column 2**	**Column 1**	**Column 2**
Number of pairs visible (total = 96)	24	24	24	24
Number of non-identical pairs (total = 51)	13	14	12	12
Range of number length possible	3 - 13	5 - 13	4 - 12	5 - 12
Index of number change	4/7 4/5 5/11 10/11 2/9 6/11 9/10 5/12 8/11 5/7 6/8 6/9 6/11	3/6 2/11 5/12 7/9 2/8 7/10 2/10 7/12 4/9 9/11 10/11 4/11 4/6 8/12	3/6 4/7 6/11 2/6 9/12 5/11 11/12 11/11 5/12 3/7 11/12 5/9	2/10 8/8 5/11 7/12 5/7 6/11 8/10 5/9 8/10 4,8/11 7/11 4/6

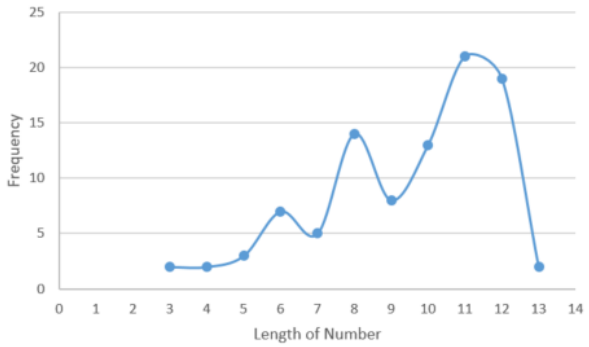


Figure 6.3: A frequency histogram showing how many times each length of number was used in the *Number Comparison* test.

6.5 Addressing Previously Identified Problems

In order to update the original PS tests and create digital versions, many factors had to be considered, which were highlighted earlier in Chapter 4 (See Section 4.6). Consequently, the below subsections will individually explain how every factor previously identified was considered—and either implemented, or disregarded—in the current PS test creation process.

6.5.1 Stimuli Content

6.5.1.1 English Language

As was described in Chapter 4 (Section 4.5.5), there appeared to be no obvious explanation for how the words that comprised the *Finding A's* test were derived. However, one thing in

particular to be aware of when creating new PS tests, was stated as the emotional sentiment that some words may contain [213]. For example, regardless of PS levels, individuals may identify words that are more common, or ones that attach a higher emotional sentimental value, more than other words.

Consequently, the recommendation of utilising The English Lexicon Project database [26] when selecting word stimuli for new PS tests, was adhered to in the digital measurement creation. This allowed all word stimuli used to be filtered and equalled for specific lexical characteristics including: an I_NMG_Mean_RT of 600-900, to ensure that the words were all fairly consistently used in the English language.

6.5.1.2 Location of target change

For both the *Finding A's* and *Number Comparison* PS tests, Chapter 4 (Section 4.5.4) had suggested that more investigation was required as to where the index of change was positioned. For example, it was unknown where in *Finding A's* the target words containing the letter 'a' should be positioned in a column containing 41 rows, or, in *Number Comparison*, where the number that differed should be located. With these unknowns, it is therefore also unknown whether different variations matter for the validity of PS, and this could act as a point for future research. For instance, perhaps having targets spaced out in certain patterns may record a more useful measure of PS. However, it was decided not to experiment with different variations of target change location in the current research, but instead to focus on creating a digital test that was as similar as possible to the original paper-based tests. If the digital tests differed greatly from the original ones, then there was a risk at changing the outcome of the whole test. Consequently, as no obvious pattern emerged for how the locations of targets were chosen in the original tests, the target locations for both of the digital PS tests were randomly generated.

6.5.1.3 Human Attention

Another point that was identified as a theme concerning uncertainties regarding PS tests in Chapter 4 was that of '*Outdated Content*' (Section 4.5.6). In particular, the notion of evolving human attention was discussed, with the aim of making researchers aware that new PS tests may require less stimuli for individuals to process. However, although this was discussed as one possible challenge of future PS testing, it was not discussed in the recommendations section as a necessity to investigate before the tests can be continually used. Furthermore, while it was described that "*it is crucial that future PS tests do not overload people's limited attentional capacities*" and "*reconsidering attention limits of participants is necessary to ensure that they are not overtired as a result of PS testing*", it was later decided that keeping the tests with a similar amount of content to the original ones was preferable. This was because research was identified which found that cognitive abilities can be affected by levels of tiredness [153]. Thus, we believed it to be important that a PS test could exceed some user's attentional limits, to

fully identify those who have a higher PS ability and are able to process more information than those who are less able.

6.5.2 Physical Layout

There have been variations of PS tests used previously, such that different numbers of stimuli, and different physical layouts of stimuli, have been present. For example, Chapter 4 (Section 4.5.4) highlighted that Ekstrom's *Number Comparison* presented stimuli in 2 columns of 24 rows, totalling 96 stimuli [83], while Zimprich & Kurtz [248] took only 60 numbers from Ekstrom's *Number Comparison*. Alternatively, a similar PS test, the Minnesota Number Comparison, presented 4 columns of 50 rows, totalling 100 stimuli [127]. As it was unknown how the layouts were devised, and whether this mattered for the validity of PS measurement, it was suggested that experiments should manipulate multiple different ways at physically viewing the PS stimuli, such as different variations of columns and rows.

However, in visual perception research more broadly, the effect of different layouts has generally been under-researched: "*another attribute that is likely to guide deployment of attention and has rarely been explored in visual search is the perceptual layout of the document, i.e. its spatial organization. Although this spatial organization is fundamental to the perception of the visual scene, it has not been incorporated within models of visual search (De Vries et al., 2013; Olds et al., 1999)*" [143]. Consequently, if models of visual search have failed to investigate different perceptual layouts, then it was decided in the current research that this would require a lot more further research that was beyond the scope of the current PhD. For example, if people scored higher PS using less columns, it would be unknown how this affected PS as a baseline. As a baseline PS test was needed, to identify how PS affects IIR, if the test was changed, then how PS affects IIR may also differ, and thus that exploration is one for future work.

Instead, the layout of the new digital PS tests was decided based upon multiple different research studies. Firstly, when completing tasks online, it is well known that a user must scroll down the page in order to view everything exhaustively [79]. Furthermore, everything visible on a web-page at any one time cannot be processed simultaneously, but instead, attention guides fixation from one point to another, where page elements are serially processed (Treisman, 1988, cited in [143]). Consequently, we believed it to be important that all targets and stimuli in the digital PS tests could be viewed without the need to scroll. This resulted in changing the number of rows visible to a user:

Finding A's

- Originally there were 5 columns of 41 rows, and 8 pages, which equated to 1640 words overall, which was to be completed in two separate 2 minute sessions.
- Whereas, the digital version contained 5 columns of 20 rows, and 5 pages, which equated to 500 words overall, and this was to be completed within one 2-minute session.

Number Comparison

- Originally there were 2 columns of 24 rows, and 2 pages, which equated to 96 pairs overall, and this was to be completed in two separate 2 minute sessions.
- Whereas, the digital version contained 1 column of 14 rows, and 5 pages, which equated to 70 pairs overall, and was to be completed within one 2-minute session.

These numbers were chosen to have a similar feel to the original paper-based test, whilst also being the maximum amount that could be comfortably viewed on a computer screen with a minimum screen resolution of 1024×768. Programmatic checks were issued to ensure these requirements were complied with. If a participant did not meet these requirements, the PS test would not run.

6.5.3 Physical Properties

Chapter 4 also proposed that many other physical properties of the PS test may be important factors to consider regarding the validity of PS testing, including: "*how participants physically select the answers on a screen such as whether selected items are scored out or change colour; whether all stimuli are presented in individual boxes, grid-lines, or blank backgrounds; if words/numbers are aligned to the left, middle, or right of the screen; what font is used; and what is the inter-letter spacing or spacing between items*" (See Section 4.6.1). However, Chapter 4 also noted that these factors may be incidental in affecting a PS score.

Nonetheless, other research in visual perception was sourced that would suggest the physical properties of the test *were* important factors to consider. For example, Leger & Chevalier [143] cited multiple studies which stated that searching for, selecting, and finding information online, may have increased difficulty when the visual salience varies (including shape, colour, and size of information) (e.g. Chevalier, Dommes, & Martins, 2013; Smith & McCombs, 1971, cited in [143]). However, Leger & Chevalier [143] also sourced other, more recent research, which would suggest that visual saliency does *not* affect visual exploration on a web page when a user is given a specific goal for their search, but rather their visual exploration is focused on the task goal instead (Borlund & Dreier, 2014; Buscher, Cutrell, & Morris, 2009; Sutcliffe & Namoun, 2012; Tabatabai & Shore, 2005, cited in [143]).

Consequently, it was believed that different physical properties—that had originally been suggested in Chapter 4—were less likely to affect PS measurement. Further inspection of the effect of different properties was thus believed to be more appropriate for future research. Instead, and given that the current digital test creation's main aim was to provide a baseline PS measurement, it was decided that the digital version should contain physical properties that were standardised and most alike to the original paper-based versions. This resulted in:

- Standardised interletter spacing.

- Helvetica font.
- Once a target was selected in *Finding A's*, a line would score out the target word.
- Once a target was selected in *Number Comparison*, a cross would appear between the numbers.

6.5.4 Test Instructions

In the original *Finding A's* PS test, Chapter 4 (Section 4.5.3) noted one possible confusion that may emerge, for both the participant completing the test and the marker scoring the test, when the instructions told participants that each column contained 5 targets. Yet, there was no guidance as to what to do if multiple columns had been attempted incompletely: whether a participant should ensure each column was complete before moving on was unknown, and this was not to be scored any differently from a participant who was able to identify the same number of targets through scanning multiple columns, but simultaneously missing many targets in their scanpath.

As there was no scoring guidance for this, such as whether a user only achieved points if the full column was correct, or whether the score differed if one column was missed (because some users may scan the words differently, such as across rows, or from right to left), it was decided not to incorporate this part of the instructions in the updated digital version. This would then allow the *Finding A's* test to be more similar to the *Number Comparison* test, where no indication of how many targets were present in each column/page were given. This adaptation was also believed to align the test as a more accurate measurement of PS, where a person's true ability to find targets, without any form of cue, would be identified.

6.5.5 Marking Tests

Another recommendation that was encouraged in the development of PS tests, in Chapter 4 (Section 4.6.2), specifically referred to the *Number Comparison* test. In particular, it was proposed that instead of just identifying the target (any pair of numbers that were *not* identical), participants should also identify any stimuli they had viewed by 'ticking' for identical number pairs, and 'crossing' for non-identical number pairs. This would allow a more precise quantification of how many number pairs a participant was able to efficiently scan through.

Indeed, on further inspection of the research area, a 2017 study by Zhang et al. was sourced with a similar implementation to this: *"Perceptual speed was measured by the Number Comparison Test (Ekstrom, French, Harman, & Dermen,1976)...The task was administered on a computer and participants responded through pressing the keyboard, with 'Q' standing for 'same', and 'P' standing for 'different'"* [243]. However, it is questionable whether this implementation is validly measuring PS, if presumably, in order to create the forced choice response, each pair of numbers were presented one-by-one. This could alter the validity because other research has

identified that when stimuli are viewed in complete isolation (with no clutter), in comparison to viewing them amongst clutter of similar composition (such as a word being read alone, in comparison to a word surrounded by other words), then both accuracy significantly improves, and response times become significantly faster, when identifying a target [87]. Therefore, if the PS test removed the clutter of other stimuli, by only allowing one to be visible at any one time in order to force a response, then this may also improve accuracy and response times in perceptual target identification, thus altering the true measurement of someone's PS.

Furthermore, in a 2019 review of evaluation approaches to functional vision for task-related ability [35], the danger of creating tasks with forced choice responses was noted. In particular, it was explained that human guesses will sometimes be correct. Consequently, having a test with a large number of choices, such as the English alphabet containing 26 letters—where the guessing rate would be 1/26—is more efficient than a small number of choices, such as the same or different, or left or right—where the guessing rate would be 1/2 [35]. For this reason, it was decided that forcing participants in the updated PS test, to either indicate that the number pairs were identical or not-identical, would allow the chance of correct guesses to negatively impact results. Therefore, the original guidelines, which was for participants to only select the targets which were not-identical, was believed to be the most optimal to reduce the likelihood that guesses could be utilised.

6.5.6 Test repetition

Finally, another point that Chapter 4 (Section 4.6.2) suggested for future PS test development referred to administrating the same PS test on multiple occasions, with the hope that participants would gain a similar score on both sessions. Otherwise, it was hypothesised that differing scores would imply the PS test was not valid, as the construct of PS being a cognitive ability was believed to be fairly stable throughout an individual's life [183]. However, on further investigation of the literature, it appeared that testing PS on multiple occasions, and expecting a consistent score to reveal a valid measurement, would not be accurate for multiple reasons.

Firstly, a movement of research involving cognitive abilities, and specifically PS, has highlighted the need for user-adaptive systems which can be tailored to help the user, based upon their unique and changeable abilities (e.g. [66, 67, 94, 210]). These different studies have incorporated eye-tracking or computer interaction data, or both, during IIR tasks, with the aim of identifying a user's dynamic cognitive abilities. It has therefore now been widely agreed that cognitive abilities are not stable, but instead can be influenced by a range of environmental factors, such as tiredness levels or depressive symptoms [32, 153, 203]. Even medications taken short-term in 'normal' volunteers can result in improved cognitive activation on simple reaction-time measurements during cognitive test batteries [5]. Consequently, a user may score a lower PS score at one moment in time, but given a different context, may have a higher PS.

Furthermore, another reason why the newly created digital PS tests were not administered on multiple occasions concerned the role of practice. As identified in the earlier subsection 3.8.2, practice effects can improve test performance, due to repeated exposure to test procedures, thus masking a user's true cognitive ability [244]. Additionally, practice effects have been found to occur within a range of test-retest intervals, from 1 day, to weeks, months, or even years [7]. Thus, regardless of when the digital PS was repeated to participants, practice effects could still negatively impact reliable results being found, thus negating the point of repeat PS testing.

6.6 Creating and Piloting the Tests

The digital versions of the PS tests were implemented using the Django framework [22]; a Python library to develop web applications using the Model-View-Controller (MVC) approach [90]. This code integrates HTML and JavaScript (jQuery) to create a functional application. It manages UI elements, tracks user interactions (including both correct and incorrect responses), and provides feedback on the user's performance in the test. All the recorded data for each user are then stored using a dedicated SQLite3 database.

Once the PS tests had been created, a pilot study was carried out to ensure that the PS tests were clear to a user. This remaining section will outline the steps involved in the piloting process, and describe what changes were made to the PS tests as a result of the pilot.

Firstly, a group of 5 volunteer PhD researchers were recruited from our local Computer & Information Science department. To ensure that the pilot study complied with ethical guidelines, informed consent was gained from every volunteer. Each volunteer completed the digital PS tests in a quiet laboratory, where only the main researcher and one volunteer were in the room at the same time. All volunteers received the same instructions on how to complete the tests, and after completion, their responses were logged. Additionally, each volunteer was asked about their experience of completing the PS tests using unstructured questioning afterwards, to document any issues or challenges that were faced during the testing process.

The result of this pilot produced one main outcome. Specifically, in *Finding A's*, it was reported that the words were overall unclear to scan as the word stimuli blended in with the instructions and rest of the page, and there were large gaps between columns. For a screenshot of what the test looked like, please refer to Figure 6.4.

As can be seen from Figure 6.4, the word stimuli were presented in black font, against a white background. However, the instructions and other information visible on the page were also presented in black font, against a white background.

Consequently, to ensure the stimuli were more clearly differentiated, research was consulted involving design guidelines for coherent online search behaviour. In particular, one study had identified the 'spillover' effect: "*the results from one source presented on a SERP can affect user engagement with the results from a different, completely independent source*" [17]. This

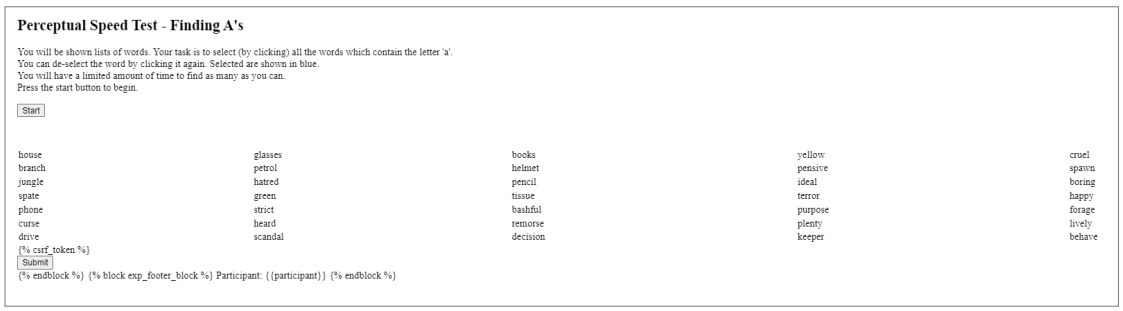


Figure 6.4: An original mock-up of the *Finding A's* test during the initial design phase.

study then identified that enclosing certain results in a border with a different-coloured background could fully eliminate the presence of spillover effects [17]. Furthermore, in Chang et al.'s [58] research, an interface was created where certain elements were segregated using a blue background, and this further enhanced the usability of the interface for participants.

The use of a border and shaded background was developed in relation to early work in Psychology, using Gestalt principles of pattern recognition, which stated that items in a display that are visually similar are perceived as a group (Koffka, 1935, cited in [17]). Consequently, the PS test was refined, so that the stimuli were presented as a group, by using a blue background bordering the black font.

Furthermore, the Gestalt principle of Proximity was also utilised, where "*items placed near each other appear to be a group...Viewers will mentally organise closer elements into a coherent object, because they assume that closely spaced elements are related and those further apart are unrelated* (Fulks 1997, Fultz 1999, Fisher and Smith-Gratto 1998–99, cited in [58]). Therefore, each column in *Finding A's* was positioned slightly closer to one another.

Once these changes were implemented on the PS tests, an additional pilot test was conducted with a new sample of 5 volunteers to confirm that the changes made to the tests were effective in addressing previous issues. This additional pilot confirmed that no volunteers reported the test to be unclear, and no further comments arose.

6.7 The Updated Perceptual Speed Tests

As a result of the pilot, and to be explicit about the digital PS tests created in order to allow repeated use, a summary of the finalised two digital PS tests have been presented below.

6.7.1 Finding As

6.7.1.1 FA Test Procedure

When participants would log onto the test, an instruction screen would firstly appear, which also contained an interactive opportunity for participants to practice the task and see whether they were correct or incorrect (See Figure 6.5).

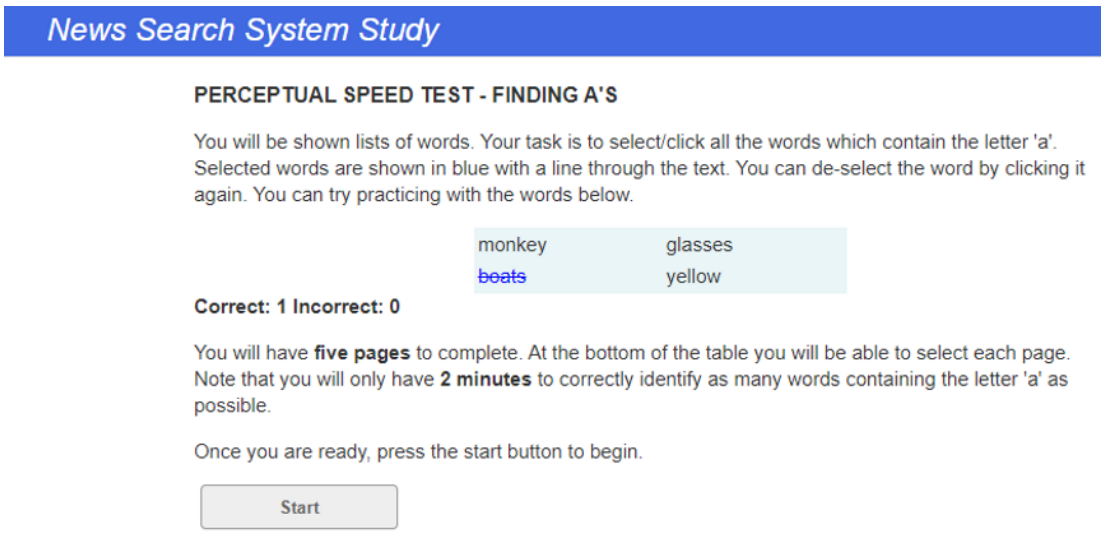


Figure 6.5: A screenshot of the instructions page for the digital *Finding A's* Perceptual Speed Test.

Once comfortable with the task, participants would begin the test by clicking 'Start'. Participants were then given two minutes to navigate five pages containing 100 words each, and select any words that contained the letter 'a' by clicking on them with their mouse cursor. When a word was clicked on, it would turn blue and contain a line scoring it out, in order to make it explicit to participants that it had been selected. If participants knowingly made a mistake, they could click on the word again and it would be deselected. Blue was chosen as the changeable colour, to be most similar to everyday websearch, where URLs most commonly turn blue after they have previously been clicked on in search engine result pages.

On each page, 12/100 words contained the letter 'a'. With 12 targets on every page, overall, out of 500 words, 60 contained the letter 'a'. A participant's score was computed by considering how many words they correctly identified, minus how many they incorrectly identified. The maximum score possible was therefore 60. Participants were unable to end the test when they wished, but instead, after two minutes elapsed, the test would time-out. Participants would be aware of how much time remained, as a countdown timer was visible at the top of the screen.

6.7.1.2 FA Stimuli Layout

As can be seen from Figure 6.6—which displays one page of the digitised *Finding A's* test—the words were presented in black font, across 5 columns and 20 rows, with left alignment. The black font stimuli were presented against a tinted blue background.

6.7.1.3 FA Selecting Stimuli

The words were generated from The English Lexicon Project database [26] and were filtered to: be English; range from 4-8 letters long; contain 1-3 syllables; consist of nouns, verbs, and adjectives; exclude plural, capitalisation, and distressing words; and be equal for specific lexical

Figure 6.6: A screenshot of the first page for the digital *Finding A's* Perceptual Speed Test.

characteristics, including an I_NMG_Mean_RT of 600-900, and Freq_HAL > 100. The last two filters were to ensure that the words contained a similar level of emotional sentimental value, and amount of usage, in the English language.

After these filters were implemented, the database exported a list of 19,935 words into a .csv file. In the adjacent empty column, a function was created to identify whether the word contained an 'a', or not: **=IF(ISNUMBER(SEARCH("a",A2)), "ContainsA","No")**.

Once this function was executed for every row, the column was re-ordered by "*Sort Ascending*" which revealed 8,934 words contained an 'a', and 11,001 did not. These were then split into two documents: one with a's; and one without a's.

In each document, another function was entered into an adjacent column: **=RAND()**.

This created a random number with up to 7 decimal places. This column was then similarly re-ordered through "*Sort Ascending*" to mix up the order of words that had initially been retrieved from the database. The first 60 rows in the document that contained a's were then selected as the target words, and the first 440 rows in the document that contained words without an 'a' were selected as the distracting words.

A new document was then opened, where the 500 word stimuli were inserted into one column.

To mix the words containing 'a' and words without an 'a' together, another random number was generated adjacent to each word. This column was then re-ordered by "*Sort Ascending*" to reveal the final order of stimuli for the PS test. This resulted in an uneven distribution of targets in every column. For example, some columns had two words containing an 'a', whereas other columns contained three or four. This variation was favoured to ensure that participants were unable to count how many targets they were finding. For example, if they knew that 3 targets were in every column, then if they found 3 targets in the first part of the column, they might skip scanning the remaining column and this could improve their overall PS score, without actually an improvement in perceiving everything.

For a full list of the stimuli used, please refer to Appendix B.A.

6.7.2 Number Comparison

6.7.2.1 NC Test Procedure

Just as was the case in the *Finding A's* test, when the *Number Comparison* test was selected, an instruction screen would firstly appear which also contained an interactive opportunity for participants to practice the task and see whether they were correct or incorrect (See Figure 6.7).

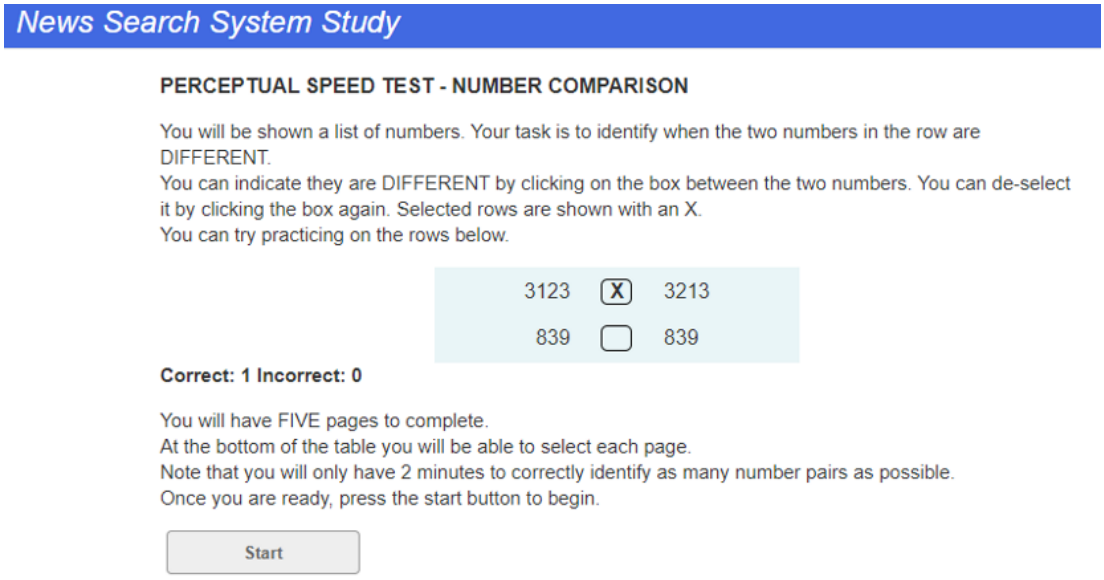


Figure 6.7: A screenshot of the instructions page for the digital *Number Comparison* Perceptual Speed Test.

Once comfortable with the task, participants would begin the test by clicking 'Start'. Participants were then given two minutes to navigate five pages containing 14 pairs of numbers each, and select any that were **not** identical by clicking in between them, in the empty box, with their mouse cursor. When a box was clicked on, a cross would appear, in order to make

it explicit to participants that it had been selected. If participants knowingly made a mistake, they could click on the cross again and it would be deselected.

On each page, between 6-8/14 pairs of numbers were **not** identical. Overall, out of 70 pairs of numbers, 34 were **not** identical. A participant's score was computed by considering how many pairs they correctly identified, minus how many they incorrectly identified. The maximum score possible was therefore 34. Participants were unable to end the test when they wished, but instead, after two minutes elapsed, the test would time-out. Participants would be aware of how much time remained, as a countdown timer was visible at the top of the screen.

6.7.2.2 NC Stimuli Layout

As can be seen from Figure 6.8—which displays one page of the digitised *Number Comparison* test—the stimuli were presented in black font, across 3 columns (a left-hand side string of numbers, middle box, and right-hand side string of numbers) and 14 rows, with middle alignment.

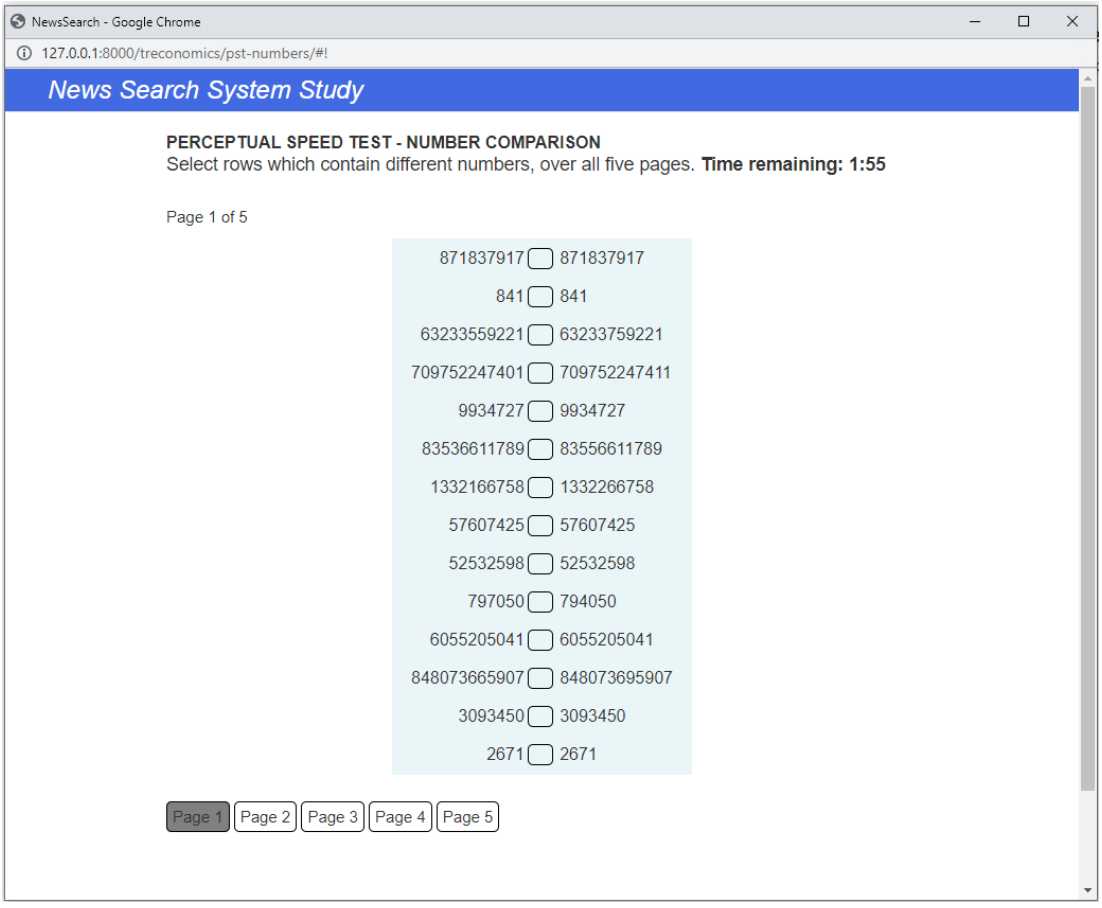


Figure 6.8: A screenshot of the first page for the digital *Number Comparison* Perceptual Speed Test.

6.7.2.3 NC Selecting Stimuli

The numbers were randomly generated from a website that produced a list of number strings for a set length [1]. Given that the original *Number Comparison* test [82] had the shortest number string of 3 numbers, 3 was also selected as the minimum number length for the digital version. Furthermore, as the original test had less numbers of shorter length, and more numbers of longer length, a similar pattern was re-produced in the digital version. Consequently, 70 strings of random numbers were generated, varying from 3 to 12 digits in length:

3x*3s*, 4x*4s*, 5x*5s*, 4x*6s*, 7x*7s*, 11x*8s*, 9x*9s*, 8x*10s*, 10x*11s*, and 9x*12s*.

In order to mix up the strings—to ensure that all strings of 3 digits, 4 digits, and so on, were not together—an adjacent column was created using the excel function `=RAND()`. This created a random number with up to 7 decimal places for every cell. This column was then re-ordered through "*Sort Ascending*" to randomly mix up the order of every row, and then deleted.

Once re-ordered, the column containing the number string was duplicated so that every string of numbers had an identical string in the adjacent column.

Roughly half of the number pairs (34) were to be **not** identical. The non-identical number pairs would only differ based on one digit, and the index of change was decided to never be the first or last digit, in accordance with the original test by Ekstrom [82]. Consequently, the index of change could range from 2 to 11 (and thus 10 different options were possible). In order to keep a varied amount of changeable indexes, it was decided that there would be the following index changes: 3x*2s*, 2x*3s*, 4x*4s*, 3x*5s*, 4x*6s*, 3x*7s*, 3x*8s*, 4x*9s*, 4x*10s*, and 4x*11s*.

Accordingly, an index change of a higher number, such as 11, would have to be positioned in a number string of longer length. Subsequently, the location of which pair of numbers would change was manually decided. This resulted in an uneven distribution of targets on every page. For example, some columns had 6 pairs of non-identical numbers, whereas other pages contained up to 8 pairs of non-identical numbers. This variation was favoured to ensure that participants were unable to count how many targets they were finding. For example, if they knew that 6 targets were on every page, then if they found 6 targets in the first part of the column, they might skip scanning the remaining column and this could improve their overall PS score, without actually an improvement in perceiving everything.

For the numbers that were to differ, the new number that the selected index would change to was calculated randomly using an online generator (See Figure 6.9)[2]. This produced a list of numbers. If ever the new number was the same as the old number, then the new number was selected as 1 less than the old number.

For a full list of the stimuli used—which also includes the number length, index of change, and what the index changed to—please refer to Appendix B.B.

[1] https://www.random.org/
[2] www.randomlists.com/random-numbers

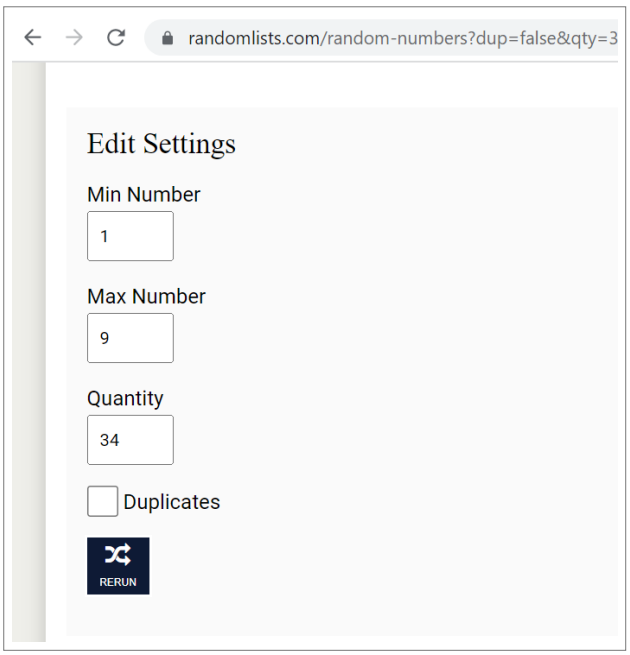


Figure 6.9: A screenshot of the website used to generate random numbers for the digital *Number Comparison* Perceptual Speed Test.

6.7.3 Summary of Updated Tests

Both digital PS tests that were developed involved participants scanning various lists of either words or numbers, and in two minutes, a participant's score was: how many correct words with the letter 'a' they had identified; and the amount of accurate non-identical pairs of numbers selected.

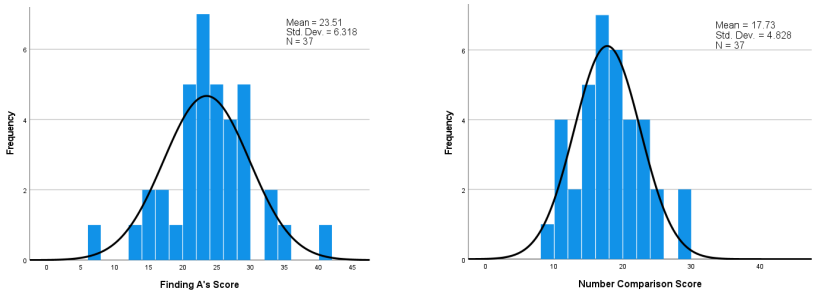
6.8 Perceptual Speed Data Collection

After the two digital PS tests were developed, they were deployed in an online experiment based upon the motivation outlined in the Methodology chapter (See Chapter 3, Section 3.5). As a reminder, this explained that participants of differing PS abilities needed to be examined completing an IIR task across different conditions of visual clutter. The detailed information about this experiment has been included in the next chapter (Chapter 7), including general demographic information about the participants who completed the study (See Section 7.5.6). However, given the current chapter aimed to describe the process of creating the PS tests, in addition to a discussion on how users should best be categorised into Low-PS and High-PS, the initial PS results have been described below.

6.8.1 Perceptual Speed Results

Overall, 38 participants completed the two PS tests. However, 1 participant was considered an extreme outlier and removed from further analysis as their logs demonstrated figures more

than 20 times larger than the other participant logs, for the experimental search tasks.



(a) Finding A's (b) Number Comparison

Figure 6.10: The distribution of scores obtained for each PS test.

Consequently, data from 37 participants who completed both PS tests were analysed. Looking at Figure 6.10, which displays the frequency histogram of scores obtained in each test, for both *Finding A's* and *Number Comparison*, the range of scores showed normal distribution. This was additionally confirmed using the Shapiro-Wilk Test, where any value greater than 0.05 demonstrated that the data was normal (See Table 6.3) .

Table 6.3: Descriptive Statistics for both the *Finding A's* and *Number Comparison* Perceptual Speed Tests.

	Finding A's	**Number Comparison**
Range	7 to 40	9 to 28
Mean	23.51	17.73
Median	23	17
Shapiro-Wilk	0.85	0.542

As can be seen from both the histogram and table:

- For *Finding A's*, scores ranged from 7-40 (the maximum score that could have been achieved was 60), with mean 23.51, and median 23.
- For *Number Comparison*, scores ranged from 9-28 (the maximum score that could have been achieved was 34), with mean 17.73, and median 17[1].

6.9 Perceptual Speed Types

6.9.1 Median Split Analysis

Previous literature has most commonly defined a user as having either Low-PS or High-PS based on a median split of the data (e.g. [4, 46, 67, 209, 218, 221]). This kind of Median Analysis represents one way of "*artificial categorization*", where categorical variables are dichotomously

[1]Please note, these figures slightly differ from the publication 'Predicting Perceptual Speed from Search Behaviour' by Foulds et al. [89], as all 38 participants were originally analysed there, before the outlier was identified.

defined from data that was originally a continuous variable [77]. This method of data categorisation has been commonly used by psychologists, when describing how a psychological trait or ability is distributed throughout a population, as it has multiple advantages.

6.9.1.1 Strengths of Median Analysis

Firstly, one strength of a Median Analysis —which can also be applied to any form of artificial categorisation— is that it enables a more simplified analysis to be undertaken: "*it is commonly easier (or at least more traditional) for researchers to analyze a variable categorically than continuously. Psychologists are typically more accustomed to using analysis of variance (ANOVA) to test influences on an outcome variable, which requires a categorical predictor variable*" [77]. Indeed, ANOVA was one of the most commonly used analysis methods in previous studies examining PS (See Section 5.5.6).

Beyond simpler analysis, the interpretations of results and thus presentation, are also simplified when using a Median Analysis through artificial categorisation. Specifically, Farrington and Loeber (2000, cited in [77]) justified the use of categorization because "*it improves communication among researchers, clinicians, and policy-makers by making results easier to understand*". This is generally because understanding differences between a limited number of groups is easier, in comparison to considering differences along a continuum.

Thirdly, DeCoster et al. [77] explained that artificial categorisation is required in certain contexts: Specifically, when researchers want to evaluate how well a diagnostic or artificially categorized measure performs in the field, dichotomization is necessary.

6.9.1.2 Limitations of Median Analysis

Whilst there are advantages to implement artificial categorisation, in Chapter 4 (Section 4.5.1), a noticeable problem of categorising data using the median split of the sample was identified: as every sample tested contained a different median, this meant that the same person could be classified differently, depending on the sample they were in. Furthermore, when a continuum is categorized, every value above the median, for example, is considered equal. This raises the question as to whether it makes sense that a value immediately above the median is considered the same as values at the higher end of the spectrum. Similarly, if a median in a sample was, for example, 41, it is questionable how much of a difference there could physically be between users with a score of say, 40, in comparison to a score of 42.

6.9.2 Extreme Group Analysis

Although Median Analysis is a common way to implement artificial categorisation, other ways of artificial categorisation are possible, which remove some of the previously identified criticism of Median Analysis. In particular, a few studies in previous literature on PS have divided users into 3 groups, based on quartiles. For example, Carenini et al. [56] explained: "*Low represents*

the bottom quartile of the values distribution (i.e. lower 25%), average represents the values within the interquartile range (i.e., middle 50%), and high represents the upper quartile (top 25%)". Conati et al. [64] similarly followed this method of three types of PS being calculated. In Extreme Group Analysis, when three groups of categories are created, the middle group—which comprises 50% of the sample—is then removed, which creates a clear separation between the two groups, at both ends of the spectrum, being compared. However, similar to Median Analysis, strengths and limitations for the use of Extreme Group Analysis also exist.

6.9.2.1 Strengths of Extreme Group Analysis

The strengths identified for Median Analysis (Section 6.9.1.1) also apply to Extreme Group Analysis. Furthermore, Extreme Group Analysis has been widely used in many domains, such as categorising IQ level, where because IQ can be distributed normally, Low and High users can be classified as those at the extreme ends of the distribution. Additionally, by conducting Extreme Group Analysis, researchers also believe that the power of statistical tests are increased: "*Restricting your focus to those at extreme ends of the distribution increases the differences between individuals on the variable, which should in turn lead to increased differences in any other variables that are related to the extremitized variable (Preacher et al., 2005)*" [77].

6.9.2.2 Limitations of Extreme Group Analysis

However, just as limitations have been identified for Median Analysis, artificial categorisation of data using Extreme Group Analysis also has limitations. In particular, some statisticians have argued that artificial categorisation, even by using Extreme Group Analysis, can distort the true research findings. Since 1974, it has been acknowledged that analysis, such as an ANOVA, on categorized data, misrepresents the relations among variables: experimental control is alluded to, even when designs lack it, subsequently reducing the size of the observed relations (Humphreys and Fleishman, 1974, and Humphreys, 1978, cited in [77]).

Furthermore, Maxwell and Delaney (1993, cited in [77]), computed that artificial categorization can lead to significant results, which are not truly genuine: "*The authors mathematically showed that there will be inflated Type I error rates for the test of the interaction between the two categorized variables if they are correlated with each other and one of them is either unrelated to or has a nonlinear relation with the outcome variable*". Taken together, these findings demonstrate that artificial categorisation may reduce the power of some statistical tests, and also produce falsely significant results.

6.9.3 Continuous Analysis

As an alternative to artificial categorisation—which regardless of which option is chosen, users are divided into categories— another option of data analysis involves analysing the whole sample of results as continuous data. Specifically, Humphreys and Fleishman (1974) and Humphreys

(1978) (cited in [77]) proposed that "*continuously measured variables should instead be left in their original form and be investigated with correlations*". This proposition was based upon the limitations of artificial categorisation previously outlined in Section 6.9.1.2, where a lack of clarity exists for how important specific numeric differences are between categories.

In the context of PS, a Continuous Analysis would mean that participants with lower scores and higher scores would be analysed by considering the slope relating the predictor variable to the outcome variable, and how it changes across the levels of other predictors.

6.9.3.1 Strengths of Continuous Analysis

Given that no standardised thresholds have previously been agreed upon for PS, it makes sense that the concept should be examined continuously, where an analysis of the whole sample is undertaken without discreet categories. Furthermore, there has been multiple instances of empirical evidence which support the use of continuous measurement for mental disorders and abilities, such as Marcus et al., and Ruscio et al. (cited in [77]). Consequently, as PS is another type of cognitive ability, a Continuous Analysis would be appropriate.

There are further advantages of undertaking Continuous Analysis. Specifically, in a review of best practice for data analysis, Decoster et al. [77] concluded that: "*The methodological literature consistently supports the superiority of continuous measures over artificially categorized measures in most circumstances....Whenever researchers are not sure whether they should work with continuous or artificially dichotomized measures, they would be best off working with the continuous measures. This suggestion is further supported by the fact that reviewers who prefer continuous measures are much more likely to criticize the use of artificial categorizations than reviewers who prefer artificial categorizations are to criticize the use of continuous measures.*"

Additionally, another advantage of Continuous Analysis refers back to a limitation of Median Analysis: given that one participant may be defined as 'low' in one sample, but 'high' in a different sample of participants, any estimates of this measure from one sample would only apply to other groups possessing similar categorisations. In contrast, Continuous Analysis considers the whole distribution, and this can be more easily generalised to other distributions.

6.9.3.2 Limitations of Continuous Analysis

There are however also limitations with conducting a Continuous Analysis. Firstly, unlike a Median or Extreme Group Analysis, it is much more difficult to statistically explain differences along a large continuum, in comparison to differences between a limited number of groups [77]. However, not undertaking an analysis due to physical difficulty, is not a justifiable reason to disregard this technique.

Yet, there have been further arguments against the use of Continuous Analysis. In particular, the general consensus agreed upon by methodologists is that certain contexts would not

be appropriate for a Continuous Analysis. As an example, in clinical decision-making contexts, these typically require the use of categorical measures [77].

6.9.4 Combining Analyses

Given that both artificial categorisation, as well as Continuous Analysis, both contain strengths and limitations, a review on best practice concluded that data should firstly be analysed continuously, and then the means from an artificial categorisation should be used to further analyse the results and help interpret the effects that were found when the variables were treated continuously [77].

Decoster et al. [77] did however also acknowledge a problem with utilising both approaches simultaneously: "*there will be times when the effects found when the variable is treated continuously will differ from the effects found when the variable is artificially categorized. It is therefore up to the researcher to examine the two models and make sure that the categorical presentation of the results accurately reflects the findings that are observed in the continuous model*". Consequently, if both analyses are undertaken, any contrasting results between the different approaches must be acknowledged.

6.9.5 Overall Perceptual Speed

Regardless of whether a participant is categorised into Low-PS or High-PS, or whether the whole spectrum from lower to higher scores is observed, Chapter 4 (Section 4.5.2) also highlighted that the original definition of PS required at least 2 tests to be administered, in order to create an overall valid measure of PS [83]. Yet, no explanation has ever been formally reported for how to create an overall measure. Instead, the few studies who have administered multiple tests believed they were measuring separate entities, and consequently analysed them separately (e.g. Allen [10]).

In order to determine the convergent validity between the two tests—the extent to which people's scores on one measure are correlated with other measures of the same construct—a participant's score in both PS tests was analysed using Pearson's correlation. This revealed a correlation coefficient of 0.19, which would imply only a weak positive correlation between the two variables. This was further observed when visually observing the correlation, where data can be seen dispersed as a random scatterplot (See Figure 6.11).

For this reason, creating an Overall measure of PS was not deemed to be appropriate through merging the results from both tests. Instead, both tests were analysed separately, and their effect on search behaviour, performance, and experience during an IIR task were compared to identify any similarities and differences.

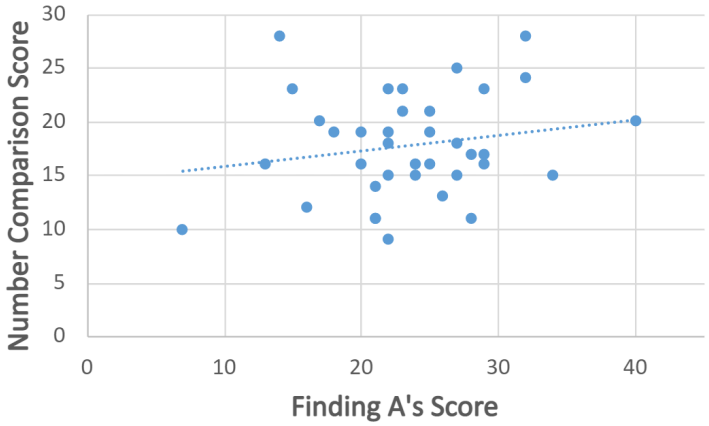


Figure 6.11: The Correlation Histogram of a participant's score in *Finding A's* compared to their score in *Number Comparison*.

6.10 Current Data Categorisation

Having considered the positives and negatives for different options of data analysis, for the current research, a mixed-approach was decided upon to be the most appropriate. Consequently, a participant's PS—as measured by the two new digital tests—was analysed as a continuous variable to observe how participants with lower scores and higher scores, were affected by different conditions during an IIR task, using correlations for each dependent variable and condition.

Then, given that certain contexts benefit from categorical measures—especially decision-making ones—participants in each PS test were also artificially categorised into Low-PS and High-PS. This was based upon previous literature, which has suggested that users with Low-PS—who may have a more negative search experience—need to be identified, with the long-term aim of developing adaptive interfaces that can help them. This would assimilate to other research irrespective of PS, but where 'Low' performers and 'High' performers of an experimental system were compared, in order to develop assistance in the future for 'Low' performers [25].

For the present research, there was also another advantage for categorising users. Given that Section 6.9.5 explained that creating a composite score of PS would not be appropriate, an important part of the current research involved comparing whether someone who scored 'Low' in one PS test, would have a similar or dissimilar experience as someone who scored 'Low' on the other PS test. Consequently, dividing users into discreet categories allowed a more explicit comparison to be undertaken between the different PS tests.

6.10.1 Extreme Low-PS and Extreme High-PS

Although two ways of artificial categorisation were described, given the variability possible from Median Analysis, the present research undertook Extreme Group Analysis.

To compute users with Extreme Low-PS and Extreme High-PS, given that the scores for both PS tests had normal distribution, any participant with a score under, or equal to, the 25% percentile was categorised as 'Low', and any participant with a score over, or equal to the 75% percentile was categorised as 'High'. Anyone in between these values represented 50% of the sample and were categorised as 'Medium'.

Based upon this analysis of the current sample, this resulted in the following number of participants in each category (See Table 6.4).

Table 6.4: The number of participants that were classified as having Low-PS (Low), 'Medium' PS, and High-PS (High) for each PS test, based on the Percentiles of all scores achieved.

		Finding A's	**Number Comparison**
Score Distribution	**25% Percentile**	20.5	15
	75% Percentile	27.5	21
Number of Participants in each category	**Low**	9	12
	Medium	19	15
	High	9	10

6.11 Chapter 6 Summary

This Chapter provided a comprehensive overview of the development of the new digital PS tests: *Finding A's* and *Number Comparison*. Furthermore, the optimal analysis to be conducted was presented, and the implementation of this approach on the current participants was outlined: PS data will be analysed both continuously, and also using artificial categorisation using Extreme Group Analysis. The following Chapter 7 then provides the results for how these two different ways of analysis will help understand how users of differing PS ability perform an IIR search task amidst different conditions of visual clutter.

Part III

Investigating the effect of different interfaces on users with differing PS ability.

Chapter 7

Investigating the effect of Clutter and Perceptual Speed during IIR

7.1 Chapter 7 Overview

This Chapter focuses on the experiment conducted with participants of differing PS levels completing an IIR task where visible clutter was either present or absent. The hypotheses are explained, before details of the method are outlined. Results are then provided for how clutter affects users of different PS ability, as measured using two separate PS tests with both a Continuous Analysis of the whole sample, in addition to Extreme Group Analysis of Low-PS and High-PS. After each dependent variable relating to search performance, behaviour, and experience are individually analysed, a summary of the main results found is presented. The chapter then concludes with a discussion of these results.

7.2 Introduction

With new digital measurements of PS having been created, the empirical component of this thesis involved administering the new tests alongside different search tasks, in order to explore how a participant with Low-PS could be helped to achieve a more positive online search experience, both subjectively, and objectively. Combined with the previous literature identified involving the field of visual clutter (See Section 2.7.6), alongside the Systematic Review conclusion which stated that visual clutter may explain the contrasting results in previous literature—for why some studies claimed that Low-PS were negatively affected during IIR, whilst others found the opposite (See Section 5.5.8)—the main research question (RQ2) explored in the current chapter was: '***What is the relationship between Perceptual Speed and visual clutter in the form of advertisements during an IIR task?***'. Furthermore, given that different PS tests were administered to participants, a further sub-question concerned: ***'Are there different effects on search outcomes, based on different PS tests?'***

7.3 Hypotheses- Research Question 2

In order to answer the main research question, many different and opposing hypotheses could be expected.

7.3.1 Low-PS will be more negatively affected by clutter during IIR in comparison to High-PS

Firstly, as previous research identified that users with Low-PS had significantly worse task performance during a webpage with higher visual complexity (Ziefle et al., 2015, cited in [44]), it was hypothesised that clutter, in the form of adverts (ads), would especially negatively affect a user with Low-PS. This led onto the first expectation which assimilated to theories on visual clutter (e.g. [2, 167, 238]) where:

- *H1) The presence of ads will cognitively overload participants and negatively distract them from completing their IIR goals, resulting in: (a) increased search time; (b) worse search performance; (c) increased difficulty in finding relevant documents; and (d) decreased visual appeal of web-pages— and all of these effects will be most pronounced for participants with Low-PS.*

Additionally, other research has found that processing too much information can increase a user's cognitive load and induce frustration and more negative emotions. However, these negative associations have resulted in a *shorter* duration of search [115]. Considering that users with Low-PS especially struggle to scan what is present [209, 211, 222], this led onto another hypothesis:

- *H2) When more information is present—in the form of more ads—Low-PS participants will have: (a) increased frustration and negative user experiences; and subsequently (b) complete their search in less time.*

7.3.2 Participants will be positively affected by clutter during IIR

Whilst some research would imply that clutter may especially negatively affect a user with Low-PS, other research has identified that images can stimulate engagement and interest in news stories [91], and with increased interaction, this results in greater accuracy at completing the search task [16]. Consequently, as ads can be presented as images, it is possible that users may benefit from this form of clutter, regardless of PS:

- *H3) The presence of ads will increase user interaction and result in higher search performance for all participants, including those with Low-PS.*

Furthermore, without any clutter present, research has found that too few elements to process can create boredom amongst users, leading them to distraction by other things around them, or causing them to abort their search all together [186]. Similarly, other studies have also identified the benefits of some clutter, where the presence of distractions can replenish mental resources, resulting in lower reported workload and stress [157]. It is therefore possible that users with Low-PS may also have positive search experiences during the presence of clutter:

- *H4) The presence of ads will reduce boredom and lower perceptions of workload, including for participants with Low-PS.*

7.3.3 Participants will be unaffected by clutter during IIR

Despite contrasting hypotheses emerging—where it is unknown whether users with Low-PS will be positively or negatively affected by the presence of advertising clutter—previous literature has also highlighted that many users ignore ads because they annoy them- a phenomenon dubbed 'banner blindness' [51] or 'ad avoidance' [147]. In a study that examined user's attitudes towards ads, banner blindness was highlighted by many users, with one explaining: "*I'm so used to seeing banner ads I tend to just ignore them*" [230]. It has been suggested that this concept is so common, that users rarely looked at ads and subsequently had very low recall of the ads that had been visible [51, 97, 128]. Thus, if the effects of banner blindness were found in the present research, a specific one-tailed hypothesis would arise:

- *H5) Webpages containing ads will be perceived as more annoying than webpages without ads.*

Then, the effects of banner blindness would be expected to transfer onto a user's general search performance and experience, leading to the final hypothesis—otherwise known as the null hypothesis—which contrasts to the previous sub-sections:

- *H6) During IIR, due to the concept of 'banner blindness', ads will not conform to the same negative effects of visual clutter, and thus a participant's search behaviour and performance will be unaffected by ad presence, including for those with Low-PS.*

7.3.4 Summary of Hypotheses

Overall, given the contrasts in theories across the literature, it is ultimately unknown how the presence versus absence of ads will affect the search behaviour, performance, and experience of a participant, depending on their PS ability.

7.4 Design

An experiment was implemented which involved both a between-participants variable and within-participants variable. The between-participants independent variable referred to a participant's PS score. This was assessed using two separate PS tests, and therefore these were analysed separately.

The within-participants independent variable concerned which condition each participant undertook their search in, during four search tasks. As previous studies have highlighted that relevant ads to the topic have elicited different attentional effects [128, 152], we created four different conditions where: 1) there were no ads visible (*No-Ads*); 2) ads were congruent with the task (*Congruent-Ads*); 3) ads were incongruent (*Incongruent-Ads*); and 4) a mixture of both congruent and incongruent ads appeared (*Mixed-Ads*).

7.5 Motivation for selecting a specific search task

To decide upon an appropriate online search task to investigate, there were many to choose from, as it has widely been acknowledged that in the field of IIR, users undertake a large variety of search tasks. Capra et al. [54] further explained that: "*search tasks vary along different dimensions, including the search task's main activity (e.g. searching vs. browsing), goal (e.g. well-defined vs. amorphous), and structure (e.g. task complexity)*".

Given that the search task chosen would need to be susceptible to a measurement of clutter, through the use of online advertising being present—in order to align to the subsequent research questions of RQ2 and RQ3—previous studies using advertising clutter were examined. However, these were not situated in the context of IIR. Instead, research has focused on the effect of advertising clutter using static websites, and not on interactive search. For example, McCay-Peet et al. used screenshots of websites, and asked participants to find a particular headline on the next page (e.g. "*Find the headline on Brad Pitt on the next page*") [164]. However, McCay-Peet et al. [164] themselves acknowledged that the static nature of their experiments was a limitation and that "*future studies should allow participants to engage more naturally with the website content*". This limitation has been shared amongst other research (e.g. [124, 195, 242]).

If interactive websearch was studied, as opposed to static images, ecological validity would be improved for the search environment, but also for the search task. This is because many visual search tasks have a pre-specified target to find [242]. Referring back to the research by McCay-Peet et al. as an example, participants knew that they were searching for the phrase "*Brad Pitt*" to appear [164]. However in real life search, a target is usually less specified in it's appearance. For example, if an information need arose such as comparing different water purification methods (as was used in Arguello & Choi [18]), the target that a user would be searching for would be unknown until the search progressed. We therefore believed it to be

important that the search task was as similar to real-life searching as possible, and thus we did not want to incorporate tasks that only involved 'Find X' (Such as [164]).

Whilst other research using advertising clutter has been interactive and more ecologically valid, the methodologies utilised have been more qualitative in nature. For example, Huang [110] investigated the effect of relevant advertising, but they did not monitor any precise search task with objective performance measurements, and instead observed the browsing behaviour of people undertaking online shopping. Although qualitative measurements allow richer detail to be identified, for the present thesis, search tasks were required that also incorporated more objective measures.

The need for objective measures was required for two main reasons. Firstly, in the field of IIR, a primary concern involves the replicability and reproducibility of experimental results. This has been deemed so severe, that it has been labelled the "reproducibility crisis", where researchers are unable to confirm previous experimental findings due to methodological flaws [47]. Therefore we aimed to create an experiment that can easily be reproduced, and in order to do this, objective measurements are ideal to act as a baseline, whereas qualitative methods are susceptible to subjectivity and thus more difficult to reproduce. Secondly, objective measurements are easier to obtain on a larger sample of participants, and this increases the generalisability of results. This would be in contrast to other research, such as Chang et al. [58], who utilised informal interviews to gauge participant responses to different types of website layout, but only interviewed 12 participants. As the present thesis aimed to understand how to help users with Low-PS achieve a better search outcome, a larger sample would be more appropriate to allow more generalisation of results to occur.

Consequently, it was identified that the search task chosen for the present research was to be as similar to real-life everyday searching as possible, using an interactive system, whilst also being able to measure both qualitative and quantitative measurements of search outcomes. This would allow a fuller understanding of how different forms of clutter impacted users with different levels of PS completing their search task.

In order to further narrow down the chosen search task, research using IIR was examined for previously used tasks that have specifically also investigated PS. Many involved users searching to find a specific answer, such as selecting a precise target from a visualisation (e.g. [103, 174, 221]). However, as was described above, we wanted a more realistic scenario to everyday search, where the target was not pre-known. Although other studies in IIR and PS did utilise such tasks, these were also not deemed to be the most ideal for the present research. For example, Kelly et al. [121] used "Create" tasks, where participants were to generate different things by answering questions such as "*What are the risks of different tanning methods?*". However, Kelly themselves recognised that these tasks involved cultural biases [121]. Instead, we wanted a task that was more culturally accessible, and that prior knowledge would not impact (For

example, somebody familiar with the area, such as a dermatologist, would be more likely to achieve better search outcomes when asked about different risks of tanning products).

The only task previously used in IIR research involving PS, which met all of the criteria outlined above —including: being interactive and ecologically valid; susceptible to different types of advertising given the use of search engine result pages; utilising a previously un-identified target; not requiring prior knowledge; and able to retrieve both objective and subjective outcome measurements, across a large sample of participants—referred to "simulated work tasks". These tasks are key to the evaluation of IIR models, through the use of realistic scenarios [38]. Here, a user is given a topic and asked to gather as many relevant and different documents that are appropriate for learning about the given topic. This task has been widely used in many studies involving IIR and PS (such as [4, 9, 11, 12, 13, 18, 125]), as Borlund [38] explained it's usefulness in the domain of IIR: "*A simulated work task situation, which is a short 'cover story', serves two main functions: 1) it triggers and develops a simulated information need by allowing for user interpretations of the situation, leading to cognitively individual information need interpretations as in real life; and 2) it is the platform against which situational relevance is judged. Further, by being the same for all test persons experimental control is provided. Hence, the concept of a simulated work task situation ensures the experiment both realism and control.*" Combining everything together, a simulated work task was thus selected as the search task under investigation for the present thesis.

7.5.1 Simulated Work Task

The context of the simulated work tasks [39] were situated within a news-based retrieval system. This was chosen to reflect a common scenario where many users retrieve news online [223]. In a simulated work task, participants are given a specific topic, and are provided with a situation that requires use of an IR system, such as imagining they were to write a short newspaper report about the topic [38]. In doing so, they must navigate a search engine to find as many different and relevant news articles[1] that they felt provided evidence for their report. The goal of the system is to help the searcher learn about a topic, and in doing so, the number of aspects that the searcher finds indicates how much they learned during the process—a process referred to as "Search as learning" (Collins-Thompson et al., 2017, cited in [161]).

As this task has been used in prior IIR studies (e.g. [122, 160, 161]), the results of the current research can be more easily compared to previous baselines. Furthermore, the use of a simulated work task offers further advantages: it is user-friendly and requires no prior knowledge; the controls are easy to learn, as they are akin to everyday searching; it can be deployed in a lab environment or exported online; and multiple search scenarios can be used, which makes it possible to explore multiple conditions.

[1] Please note, although participants were instructed to find relevant articles, throughout this thesis articles are referred to as 'documents' to be in keeping with previous IIR studies.

7.5.2 Search Topic

The *TREC Common Core 2017 (CC2017)*—which consists of over 1.8 million newspaper articles from the *New York Times (NYT)*, ranging from the period 1987 to 2006 [8]—provided the test collection as this provides topics and a subset of documents that have already been judged as relevant or non-relevant by separate assessors. This allowed a more objective and precise measure of participant performance to be quantified. Additionally, by having pre-defined topics, it was possible to ensure that ads that were visible matched the conditions needed: congruent and incongruent with the task. Otherwise, selecting relevant ads would have been difficult, as it has previously been identified that user queries are on average only about 2.5 words long, and hence difficult to interpret to retrieve relevant ads [191]. Furthermore, comparing participant performance using TREC has also been used with participants of differing visual abilities—such as those with and without dyslexia [171]—and therefore it was believed to be an appropriate measure for also assessing participants with different levels of PS.

Five topics were chosen that reportedly had similar levels of difficulties in other IIR studies [122, 160, 161, 225]: Airport Security, Wildlife Extinction, Tropical Storms, Curbing Population Growth, and Piracy. To reduce order or topic effects from occurring, all topics were randomly allocated to a condition for each participant, except Piracy, which always remained as the practice task. Although a random allocation could have resulted in a topic being linked to a specific condition more than others, this was analysed post-experiment and a fairly similar distribution was found. For example, for Interface 3—the incongruent-ad condition— 10 participants completed this under the Airport Security topic; 10 did Wildlife; 8 did Storms; and 9 did Population.

7.5.3 Search System

To run the experiment, a custom-built search system was created. This expanded upon the TREConomics framework, which had been developed over many years and has been successfully used in many other research studies such as Azzopardi et al. [21], Maxwell et al. [159], Kelly et al. [122], Edwards et al. [81], and Crescenzi et al. [73].

This framework allowed for an interface to be developed that would be familiar to anyone who had used a web-based retrieval system, which as explained by Maxwell [162], meant that the learning curve for using the interface would be low.

The interface comprised of three different main views:

i) The Search Engine Results Page (SERP): This included a box where participants could issue queries, and then 10 result snippets per page would become visible. The title, the source, and any snippet text were all provided. Given that the experiments were based on news search, the source was the name of the newswire from which the document originated.

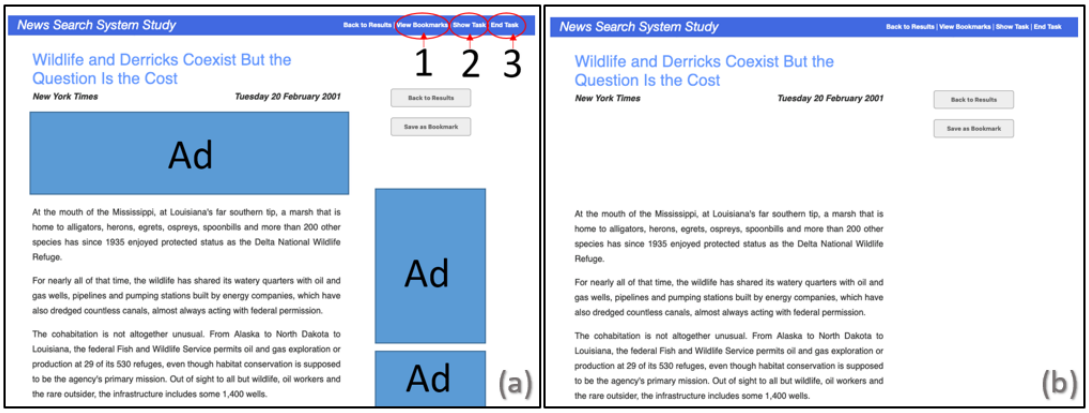


Figure 7.1: An example layout of the document view when (a) ads were present, and (b) ads were absent. Section 7.5.3 describes the annotations. Please note, in the ad conditions, additional ads were positioned on the right and bottom if a participant scrolled down.

ii) Document view: If a result snippet was clicked, then this opened up the full text of the document. A participant could then bookmark it if thought relevant, or press a button to return them to the previous SERP. An example of the document view has been presented in Figure 7.1.

iii) The Saved documents list: This provided a list of every bookmarked document from the search session. Participants could edit this list by removing any bookmarks they later deemed irrelevant.

Three additional buttons were positioned at the top of the webpage, which were always visible regardless of the main view (See Figure 7.1): **1)** 'View bookmarks' which directed participants onto the Saved documents list; **2)** 'Show task' allowed participants to remind themselves of the specific task, as it has previously been found that information workers struggle with their memory for the exact task to be completed [158]; and **3)** 'End task' allowed participants to move onto the next section when they felt they had found enough.

The Whoosh Information Retrieval (IR) toolkit[1] with the *BM25* retrieval algorithm ($\beta = 0.75$) was used as the underlying retrieval system. The P@10 values were computed for every query, which scores how many relevant results were among the top 10 results presented to a participant.

For conditions where ads were displayed, a banner ad was located at the top, bottom, and four ads in the right rail of the webpage, on both the SERP and document pages, as done in previous advertising research [15]. The ads were always randomly selected from a pool of ads on each page load depending on the condition – with the pool consisting of congruent, incongruent, or a mixture of both ads. If an ad was clicked, this would yield a popup window displaying a larger version of the ad. Note that in our study, we only observed four clicks on ads over

[1] `https://pypi.org/project/Whoosh/` – last accessed January, 2020.

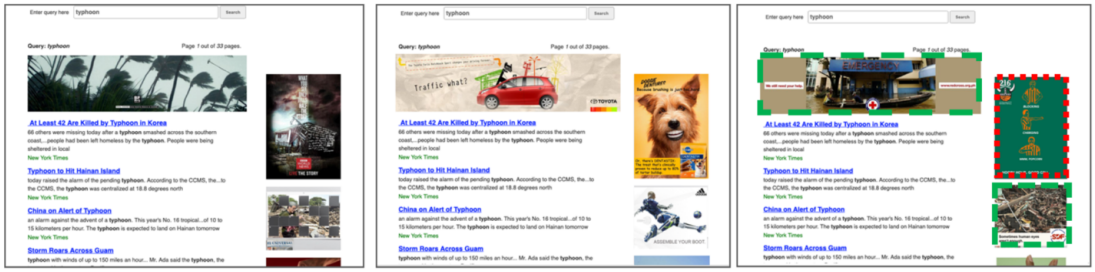


Figure 7.2: An example of the interface for every ad condition for the query 'typhoon' in response to the *Tropical Storms* topic. The far left depicts *Congruent-Ads*, the middle shows *Incongruent-Ads*, and the far right contains *Mixed-Ads*, where the visible ads are both congruent (highlighted with a green border for the purpose of demonstration) and incongruent (highlighted with a red border).

all participants, meaning that few ads were actively engaged with. For the *No-Ads* condition, blank space was left to ensure that the content information in the webpages was always in the same location and that there was no bias in presenting the information higher up.

The finalised experimental system was tested on multiple different web browsers, including Google Chrome, Apple Safari, and Mozilla Firefox. Additionally, different operating systems were also tested, including Apple macOS and Microsoft Windows. This ensured a similar search experience was occurring, regardless of individual system configurations.

7.5.4 Advertisements

As ads come in many formats, to minimise potential confounds of factors such as animations, personalised-ads, or other interactive ads affecting performance, we chose to focus on static banner ads as these appear to be the most commonly researched in the literature [152]. Static banners were sourced from the *Ads of the World* database[1] where, for each topic, a selection of congruent and incongruent ads were selected. In line with Buscher et al. [52], congruent ads were defined by their appropriateness to the search task. Three volunteers manually examined all ads for their appropriateness to each topic. All raters had to agree on an ad's appropriateness to be considered either congruent or incongruent. Inconclusive ads were discarded. This created 6 databases where each topic (Airport Security, Wildlife Extinction, Tropical Storms, Curbing Population Growth, and Piracy) had its own selection of 40 congruent ads, and then one large database comprised of 200 ads that were incongruent for all topics. An an example of *Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*, please refer to Figure 7.2.

Furthermore, as it is known that the saliency of ads can impact search outcomes [109], and more general webpage saliency can also affect search [212, 239], the present study wanted to ensure the ads used did not vary on saliency between conditions. Consequently, the ads for each topic were uploaded to QUESTIM [2], an online tool that computes various evaluation metrics,

[1] https://www.adsoftheworld.com/ – last accessed January, 2020
[2] `http://questimapp.appspot.com/` – last accessed September, 2020.

including saliency, which has been used in prior studies (e.g. [239]). For each topic, an average saliency measure was created, and a follow-up t-test between *Congruent-Ads* and *Incongruent-Ads* revealed that overall there were no significant differences ($p = .23$). Additionally, no large differences occurred between topics, reducing the likelihood that saliency was a confounding factor.

7.5.5 Experimental Procedure

The experiment was conducted on the online platform Prolific[1] where participants must have had a minimum screen resolution of 1024×768, and disabled any ad-blockers. Programmatic checks ensured these requirements were complied with. In accordance with the ethical approval sought from our University department's Ethics Committee (№ 1044), each participant was provided with instructions about what they would be required to do and then gave informed consent if they were happy to proceed. After completing a short demographics survey, the participants undertook the PS tests. Then, a practice search task using the *Piracy* topic was provided, to familiarise participants with the system so they could learn how to query, browse, and save documents. Before each main search task, participants completed a brief survey about their knowledge for the topic and then continued onto the task. As an example of the task given, for the topic *Wildlife Extinction*, participants were explicitly told: "*Find and bookmark articles that discuss EXTINCTION PREVENTION MEASURES made by countries to protect DIFFERENT WILDLIFE SPECIES*". For a full list of the information pages provided to participants, instructions, and search scenarios, please refer to Appendix C.A.

When participants felt they had saved enough relevant documents, they were to press the button labelled 'End task'. Alternatively, to ensure the overall experiment did not overrun, the system would automatically move onto the next part after eight minutes. Eight minutes was chosen as a similar experimental setup that also used the TREC Common Core 2017 collection found that users spent approximately seven minutes per task [161]. Immediately afterwards, three user search experience questionnaires were given (described later in detail in Section 7.6.3). Then, a post-task questionnaire was given to assess how many concepts participants could recall from their search on that topic. This process continued for the remaining 3 search tasks. To ensure that topic and ordering effects were minimised, a fully factorial design was implemented where the ad-type was always randomly rotated between topics, and the order of topics always varied for each participant.

The experiment was designed to last for around 45-50 minutes, which included all search tasks and surveys being completed.

[1] `https://www.prolific.co/` – last accessed July, 2020.

7.5.6 Participants

38 participants completed the study: 23 males; 14 females; and 1 did not disclose their demographic details. Ages ranged from 18 to 58, with a mean of 32 years old. One male participant, aged 22, was considered an extreme outlier and removed from analysis as their logs demonstrated figures more than 20 times larger than the other participant logs. All 37 remaining participants were native English speakers with a range of educational backgrounds, as self reported highest level of education achieved included: 4 post-graduates, 24 college graduates, and 8 high school graduates. For taking part, participants were compensated with the equivalent of US$13. Furthermore, every participant completed two PS tests. The results from these two tests were outlined in the previous Chapter 6 (Section 6.10.1).

7.6 Dependent Variables

Dependent variables for this study were split into three main categories: *Search Performance*, considering how well participants performed; *Search Behaviours*, considering participants' interactions with the system; and *Search Experiences*, considering what participants thought and felt about the task, system, and personally.

7.6.1 Search Performance

Firstly, participant performance was measured as the total number of documents a participant saved for a given topic (**Total-Saved**).

Given that the TREC CC 2017 contained pre-assessed relevance judgements [8], we were also able to estimate participant search performance through counting how many documents that participants had saved were known to be TREC-relevant for the given topic (**Relevant-Saved**) [1].

Additionally, as every participant would likely issue different queries and thus the search system would retrieve different levels of relevant results, additional performance measures were calculated for how many relevant documents had been saved in relation to how many relevant documents a participant had: *a)* saved overall (**Relevant-Saved/Total-Saved**); *b)* hovered over in the SERP (**Relevant-Saved/Relevant-Hovered**); and *c)* actually clicked on (**Relevant-Saved/Relevant-Clicked**).

Furthermore, immediately post-task, to give an indication of how much participants had learned from their search, participants had to recall as many concepts that they had previously found (**Concepts-Recalled**). For the following topics, participants were asked to recall, from what they had found in their search:

- **Wildlife extinction**: as many WILDLIFE SPECIES and their EXTINCTION PREVENTION MEASURES.

[1] Note, from here on, TREC-relevant documents will just be referred to as relevant.

- **Tropical Storms**: as many TROPICAL STORMS which caused fatal damage and their COUNTRY LOCATION.
- **Airport Security**: as many AIRPORTS and SECURITY MEASURES that were taking additional security measures.
- **Curbing Population Growth**: as many COUNTRIES and the MEASURES they use to control population growth.

To analyse whether the concepts recalled had just been learned or were already prior knowledge, two checking measures were implemented. Firstly, before each task, every participant completed a brief survey where they indicated how much they knew about the topic on a Likert-type scale ranging from *1 - Nothing* to *5 - A lot*. Secondly, each concept recalled was checked against the documents that they had identified, and only counted as correct if the content matched what a participant had interacted with in their search.

7.6.2 Search Behaviour

To provide exploratory insights into participant search behaviours, various interactions with the search system were logged for each topic, using behavioural measures that have been widely used in previous IIR studies [40, 46, 122, 225, 235] such as: the number of queries issued; average query length; and documents hovered over and clicked on (including those which had been pre-assessed as relevant). Of course, it was possible that participants may have engaged with certain aspects of the webpage without hovering or clicking on something. However, this possibility was the same in previous IIR research (e.g. Arguello & Choi [17]), and it is widely agreed that search effort should be derived from queries, clicks, and task completion time [55].

From the log, we therefore also computed a series of time-based measures, including: total time spent on SERP; total time spent examining documents; and total session time per topic. It should be noted that in this thesis, all durations of time have been reported in seconds.

7.6.3 Search Experience

Participant subjective search experiences were analysed using multiple surveys after each condition. The surveys were split into three sections, with 5-point Likert-type items adapted from various studies [18, 73, 96, 157, 164, 186]:

(1) Task-Focused Survey: Participants focused on their perception of the task over two statements with various scales. Questions included: *(a)* how difficult was it to find relevant documents for this topic? *1- Very easy* to *5- Very difficult*; and *(b)* how much did you learn about this topic? *1- Nothing* to *5- A lot*.

(2) User-Focused Survey: Using a scale of *1- Strongly disagree* to *5- Strongly agree*, participants expressed how the search made them feel, considering their: *frustration*; *confidence*; *enjoyment*; and *tiredness*.

(3) System-Focused Survey: Participants rated their perception of the system over statements with the scale of *1-Strongly disagree* to *5-Strongly agree*. Questions considered how: *aesthetically appealing*, *boring*; and *annoying* the system was.

7.7 Analysis

For RQ2 specifically, "***What is the relationship between Perceptual Speed and visual clutter in the form of advertisements during an IIR task?***", as just described, the dependent variables were split into three categories: measures of performance; behaviour; and user experience. Every dependent variables for the three different ad conditions (*Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*) were then averaged to devise an overall measure of clutter (named from now on as *All-Ads*). This operationalisation of averaging conditions aligned with the most recent literature we could find that compared the presence of one factor that had varying levels, against another condition where the main factor was absent [76, 88]. Thus, two conditions were analysed for RQ2: ***All-Ads*** versus ***No-Ads***.

To double check *All-Ads* was similar to *No-Ads*, the $P@10$ values for every query in each condition were compared using a t-test which returned a non-significant result, confirming that generally, participants in each condition saw similar levels of relevant results ($P@10$ mean in *No-Ads*: 0.33, and *All-Ads*: 0.31).

Furthermore, two different PS tests were administered on each participant. Whilst creating an overall measure of PS was considered to align with original guidelines, there was almost no correlation between a participant's scores on the *Finding A's* PS test, compared to the *Number Comparison* PS test (See Section 6.9.5). Consequently, the results from each test were analysed separately, to answer the sub-research question of: "***Are there different effects on search outcomes, based on different PS tests?***".

7.7.1 Performance and Behaviour Analysis

7.7.1.1 Correlations

As explained in Section 6.10, analysis of PS should firstly be treated as continuous data, and analysed using correlations, to identify how participants with lower and higher scores, are affected by different conditions during a simulated work task. Therefore in the current chapter, for each dependent variable measured using interval data (which was every dependent variable concerning Performance and Behaviour), individual Pearson correlations were administered that compared:

1. A participant's score in ***Finding A's***, against the dependent variable during *All-Clutter*;
2. A participant's score in ***Finding A's***, against the dependent variable during *No-Clutter*;

3. A participant's score in ***Number Comparison***, against the dependent variable during *All-Clutter*;

4. A participant's score in ***Number Comparison***, against the dependent variable during *No-Clutter*.

For every correlation, the correlation coefficient (r) is reported, where the value can range from -1 to +1. A value greater than 0 indicates a positive association; where, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases [135]. The strength of association is indicated based upon the following ranges for r: *Small* (0.1 to 0.3 *or* -0.1 to -0.3); *Medium* (0.3 to 0.5 *or* -0.3 to -0.5); and *Large* (0.5 to 1.0 *or* -0.5 to -1.0) [135].

7.7.1.2 Analysis of Variance

To further interpret any effects found, each PS test was also artificially categorised using percentiles, to identify a participant with Extreme-Low and Extreme-High PS. For the remaining chapter, Extreme-Low and Extreme-High will just be referred to as Low-PS or High-PS whenever the results correspond to participants in both PS tests. To identify participants in different PS tests, the following abbreviations have been used and colour coded to allow for easier reading:

- For ***Finding A's (FA)***, participants with Low-PS were referred to as Low-FA, and High-PS as High-FA;

- For ***Number Comparison (NC)***, Low-NC and High-NC depicted participants with Low-PS and High-PS.

As outlined in Section 6.10, the categorisation of participants into Low-PS and High-PS created the following number of participants in each Extreme Group: Low-FA: 9; High-FA: 9; Low-NC: 12; High-NC: 10.

The means of these Extreme Groups were then compared using a 2 (*No-Ads* vs *All-Ads*) x 2 (Low-PS vs High-PS) Mixed ANOVA, individually, for both *Finding A's* (Low-FA vs High-FA), and *Number Comparison* (Low-NC vs High-NC).

A Mixed-ANOVA was selected as the most appropriate method of statistical analysis because this requires one group to be 'within-participants' (which here, was whether participants performed their searches amidst *No-Ads*, or *All-Ads*), and the other to be 'between-participants' (Low-PS versus High-PS) [132]. There has previously been some criticism of using multiple comparisons after an ANOVA—instead of a MANOVA—due to an inflation of error rate [104]. However, there is now agreement that in exploratory contexts, "*the statistical significance of one finding does not conceptually (or philosophically) depend on whether another finding is determined to be statistically significant*" , and thus the "*adjustment for multiple comparisons to*

avoid inflated Type I error rates is not necessary" [104]. Furthermore, given that ANOVAs were a commonly used analysis technique in similar studies with multiple dependent variables (E.g. [12, 13, 43, 46, 88, 125, 178, 204, 225]), this verified their use and provided a means to more easily compare the current findings to previous results.

The Mixed ANOVAs were thus computed using Pingouin, and if a significant result was returned, then pairwise t-tests were used for post-hoc analysis [226]. The F-Score, p-value and the effect size η_p^2 are reported (where value/ranges of ηp^2 indicate: *small* (< 0.06); *medium* (0.06–0.14); or *large* effect sizes (> 0.14) [63].

7.7.2 Experience Analysis

Whilst measures of Performance and Behaviour had been collected as interval data and analysed using Correlations and ANOVAs, measures of User Experience were gathered as ordinal data, based upon Likert-type survey scales. Consequently, parametric statistical tests, such as an ANOVA, would not have been appropriate to perform [104].

Whilst non-parametric statistical tests were considered for analysis of the survey data, none were deemed appropriate without breaching assumptions of a specific test. For example, the Mann-Whitney U Test or Kruskal-Wallis H Test would have required the independent variable to consist of independent groups, and thus between-participants [133, 134]; and the Wilcoxon Signed Rank Test, or the Friedman Test were the opposite, requiring "matched pairs" or only within-participants [131]. Yet, the present experiment required analysis of both a between-participant *and* within-participant group.

Therefore, to gain an understanding of the differences between participants with lower and higher scores on each PS test, the medians were compared between Low-PS and High-PS, against both *No-Ads* and *All-Ads*, for both PS tests. Median comparison was selected—as opposed to the mean— as it is well known to be the most appropriate measure of central tendency for Likert type, ordinal data [104, 214].

Furthermore, to ensure the whole dataset was observed—and not just the median—graphical representation of horizontal stacked bar graphs were created to further understand results. This aligned with other research (e.g. [88]) because it allowed a wider view on the whole sample to be observed, by converting the counts for how many participants selected each option, and then displaying these as percentages of the whole sample to allow a visual comparison to occur.

7.8 Results

Given that the hypotheses for how clutter might affect participants with Low-PS varied—such as H1 hypothesising that search performance would worsen during clutter, whereas H3 hypothesising that search performance would improve amongst clutter—the following results section has firstly been split into each dependent variable, which could be grouped into three

Table 7.1: The results for every measure of performance, in both *No-Ads* (NA) and *All-Ads* (AA), for both the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test. The row shaded in grey (denoted r) represents the correlation coefficient in each condition. The other rows present the means and standard deviations observed for users with Low-PS (L) and High-PS (H). If a cell contains a *, this means an ANOVA found a significant difference between L and H. If a † is present, this represents a significant correlation.

Performance		**NA (FA)**	**AA (FA)**	**NA (NC)**	**AA (NC)**
Total documents saved	r	-0.04	-0.20	-0.01	0.19
	L	4.67 ± 4.27	6.04 ± 6.05	5.58 ± 3.06	4.61 ± 1.6
	H	4.89 ± 2.37	4.26 ± 1.12	5.8 ± 3.65	6.3 ± 5.53
Relevant-Saved / Total-Saved (%)	r	-0.09	0.39†	-0.26	0.15
	L	65 ± 29	56* ± 13	76* ± 23	55 ± 16
	H	63 ± 26	74* ± 0.07	50* ± 22	64 ± 12
Relevant-Saved / Relevant-Hovered (%)	r	-0.24	-0.36†	-0.05	0.24
	L	57 ± 36	48* ± 18	53 ± 26	26 ± 14
	H	30 ± 29	29* ± 13	43 ± 27	39 ± 14
Concepts Recalled	r	0.19	0.23	0.19	0.18
	L	3.11 ± 1.9	2.37 ± 1.01	3.58 ± 2.5	2.68 ± 1.8
	H	4.0 ± 2.06	3.02 ± 0.84	4.8 ± 1.93	2.85 ± 1.27

types: dependent variables that measured Performance, Behaviour, and User Experience. Then for each dependent variable, results from both PS tests have been presented separately. As a reminder, FA refers to the *Finding A's* PS test, and NC refers to the *Number Comparison* PS test.

To distinguish between the "No-Clutter" and "All-clutter" conditions, these were instead referred to as "*No-Ads*" and "*All-Ads*". Given that the abbreviation of "No-Clutter" would have been "NC"— which is already the abbreviation for the *Number Comparison* PS test—the use of "Ads" was selected, instead of the word "Clutter", for defining the condition names. Consequently, *No-Ads* were compared against *All-Ads*.

After the results for each variable have been described individually, a summary is then provided (in Section 7.12) that compares the different variables together, in order to address the original hypotheses.

7.9 Clutter Results: Performance

All results for measures of Performance, in both *No-Ads* and *All-Ads*, are reported in Table 7.1 and discussed in detail in the below subsections.

7.9.1 Total documents saved

Firstly considering how many documents a participant saved that they believed to be relevant to the search topic, when no clutter was visible (*No-Ads*), there was almost no correlation present for both FA (r= -0.04) and NC (r= -0.01). Similarly comparing the means of the PS Extreme Groups, the total number of documents saved were similar regardless of PS (E.g. in *No-Ads* for

users in *Finding A's*, Low-FA: 4.67, High-FA: 4.89). This would imply that regardless of PS, participants retrieved a similar amount of documents when there was no clutter visible.

However, when clutter was present in the form of ads (*All-Ads*), participants were differently affected depending on their scores on each PS test. For FA, a small negative correlation emerged (r= -0.20), implying that participants with lower PS scores saved more. This was confirmed by comparing the Extreme Group means, where Low-FA saved roughly 2 **more** documents during *All-Ads* (Low-FA: 6.04, High-FA: 4.26).

In contrast, for NC, a small positive correlation emerged (r= 0.19), and further analysis of the Extreme Groups identified that participants with Low-NC saved roughly 2 **less** documents during *All-Ads* (Low-NC : 4.61, High-NC : 6.3).

With contrasting findings present, it was immediately evident that the two different PS tests appeared to be measuring different aspects of a participant's ability.

7.9.2 Relevant-Saved / Total-Saved

When explicitly calculating the percentage—for how many relevant documents were saved, out of how many documents were saved overall—consistent patterns emerged between the two different PS tests. This showed that during *No-Ads*—for both FA and NC—a negative correlation occurred, and Low-PS participants achieved higher accuracy in comparison to High-PS participants. Whilst the difference was minimal for FA (Low-FA: 65%, High-FA: 63%), the difference for NC was significant (Low-NC : 76%, High-NC : 50%, $F(1, 20) = 10.12$, $p = 0.005$, $\eta_p^2 = 0.34$).

Yet, during *All-Ads*, the opposite pattern emerged: positive correlations were reported (of which the one for FA was significant, r= 0.39, $p = 0.02$), and Low-PS participants performed worse than High-PS participants (which was again significant for FA: Low-FA: 56%, High-FA: 74% accuracy).

Further comparing Low-PS versus High-PS, Low-PS always performed better during *No-Ads*, and High-PS always performed better during *All-Ads*. This was despite the largest number of documents being saved overall reported as Low-FA during *All-Ads*— as this was the condition where their lowest accuracy was reported, this would further emphasize the negative effects of clutter: when clutter was present, Low-FA participants were making more mistakes in their search.

Therefore, although differences were initially reported between the two PS tests—in terms of how many documents were being saved overall—when analysing the *accuracy* of these documents, the same conclusion could be found for both tests: when clutter is present, Low-PS participants are most negatively affected, and yet when no clutter is present, Low-PS participants can perform better than High-PS participants (See Figure 7.3).

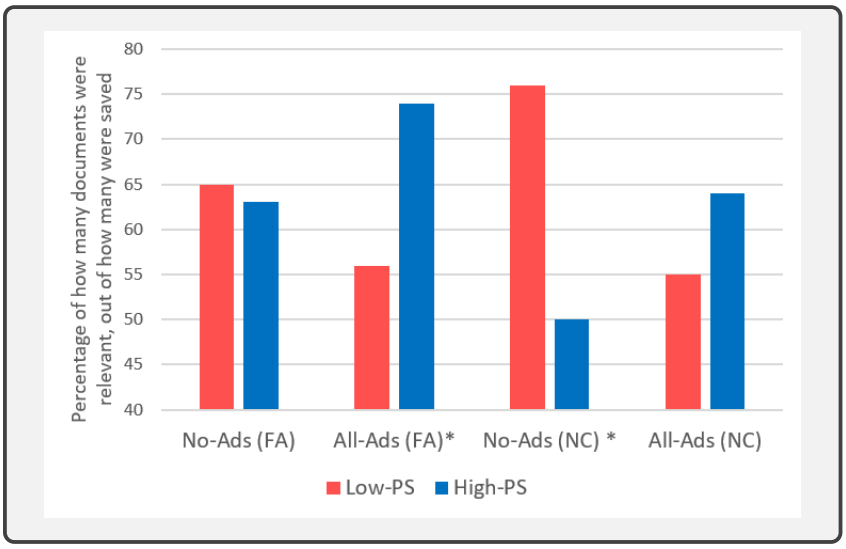


Figure 7.3: The search accuracy, as defined by the percentage of how many documents were relevant, out of how many were saved overall, for participants of different PS ability in both the ***Finding A's*** test (FA) and ***Number Comparison*** test (NC), in both the *No-Ads*, and *All-Ads* search conditions. * denotes significant differences between Low-PS vs High-PS.

7.9.3 Relevant-saved/ Relevant-Hovered

As a further measure of accuracy, the amount of retrieved relevant documents were compared against how many relevant documents a participant had physically hovered over during their search. This further showed that when clutter was present, participants with Low-PS had lower accuracy, in comparison to when no clutter was present, for both FA and NC.

Furthermore, when identifying where the highest accuracy was achieved overall (out of Low-FA, High-FA, Low-NC, High-NC, and for both *No-Ads* and *All-Ads*) this was for participants with Low-FA, during *No-Ads*, who achieved 57%. Furthermore, when analysing the correlations, negative coefficients occurred for *No-Ads*, for both FA and NC. Therefore, although previous research has implied that participants with Low-PS perform searches more poorly, the current results provide evidence that in the right condition, Low-PS participants can actually perform better than participants with High-PS (E.g. High-FA only scored 30% during *No-Ads*, in comparison to Low-FA scoring 57% there).

Whilst Low-PS participants appeared to be negatively affected by clutter, this was especially for participants of NC, where a positive correlation occurred during *All-Ads* (r= 0.24).

However, although High-PS participants generally appeared to be able to cope with clutter, in FA, High-FA participants had lower accuracy during *All-Ads* (29%) compared to Low-FA participants in the same condition (48%). This difference was significant, both in the correlation analysis (r= -0.36, $p = 0.03$) and ANOVA ($F(1, 16) = 5.59$, $p = 0.031$, $\eta_p^2 = 0.26$). Nonetheless, the score for High-FA participants during *All-Ads* (29%) was similar to their score in *No-Ads* (30%), whereas Low-FA demonstrated higher accuracy during *No-Ads* (57%) compared to

Table 7.2: The results for every measure of behaviour, in both *No-Ads* (NA) and *All-Ads* (AA), for both the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test. The row shaded in grey (denoted r) represents the correlation coefficient in each condition. The other rows present the means and standard deviations observed for users with Low-PS (L) and High-PS (H). If a cell contain a *, this means an ANOVA found a significant difference between L and H. If a † is present, this represents a significant correlation.

Behaviour		**NA (FA)**	**AA (FA)**	**NA (NC)**	**AA (NC)**
Number of queries	r	-0.04	0.03	-0.04	0.08
	L	3.0 ± 1.41	3.48 ± 1.67	2.92 ± 1.93	4.08 ± 1.83
	H	3.56 ± 3.5	3.93 ± 2.43	3.4 ± 1.51	4.4 ± 2.17
Document click count	r	0.10	0.12	0.00	0.34†
	L	6.44 ± 4.67	7.19 ± 5.97	7.92 ± 4.34	7.11 ± 2.69
	H	8.22 ± 2.64	8.59 ± 2.23	8.9 ± 4.12	10.57 ± 4.84
Relevant document click count	r	0.12	0.39†	-0.28	0.36†
	L	3.0 ± 2.06	3.19* ± 2.0	5.42 ± 3.12	3.28* ± 1.65
	H	4.33 ± 2.74	5.19* ± 0.77	3.3 ± 2.0	5.03* ± 1.88
Time session Overall	r	-0.06	0.18	-0.12	-0.01
	L	353.37 ± 158.69	361.2* ± 97.9	367.05 ± 117.6	402.86 ± 115.86
	H	396.4 ± 84.58	458.48* ± 49.67	357.09 ± 107.74	413.81 ± 95.85
Time spent on SERP	r	0.28	0.06	-0.05	-0.20
	L	73.39* ± 31.37	93.74 ± 47.21	107.96 ± 62.04	128.48 ± 44.01
	H	137.2* ± 59.84	123.62 ± 23.58	110.02 ± 32.87	112.16 ± 49.33
Time spent on documents	r	-0.09	0.12	0.01	0.08
	L	214.12 ± 109.04	224.68 ± 75.05	203.94 ± 120.25	217.24 ± 97.27
	H	203.74 ± 89.06	275.28 ± 38.98	212.38 ± 91.6	245.8 ± 68.07

All-Ads (48%). Therefore, this still suggests that High-PS participants are less affected by the presence of clutter, and that it is Low-PS participants who are negatively affected by clutter.

7.9.4 Concepts Recalled

For the concepts recalled post-task, positive correlations always occurred, in both FA and NC, and in both *No-Ads* and *All-Ads*. This demonstrated that regardless of PS test administered, participants who had lower PS scores tended to recall less than participants with higher PS scores.

Furthermore, when analysing the means of the Extreme Groups, participants with Low-PS were again most negatively affected by clutter, recalling less during *All-Ads* in comparison to *No-Ads*, for both FA and NC. However, this was not only the case for participants with Low-PS, but also High-PS, where more concepts were recalled during *No-Ads*. Consequently, regardless of PS, the negative effects of clutter appear to extend onto all participants for the amount they were able to learn as a result of their search. This corresponds with theories on visual clutter more generally, where it is well documented that memory is affected in the presence of clutter [144].

7.10 Clutter Results: Behaviour

All results for measures of behaviour, in both *No-Ads* and *All-Ads*, are reported in Table 7.2 and discussed in detail in the below subsections.

7.10.1 Number of Queries

Observing the behavioural patterns between conditions, almost no correlation was found for the number of queries a participant would issue, for scores on either PS test, or clutter being present or absent (e.g. r ranged from -0.04 to 0.08).

However, on further inspection of the Extreme Group means, it emerged that Low-PS participants always issued slightly fewer queries than High-PS participants, regardless of PS test used, or condition analysed. Furthermore, for both Low-PS and High-PS, a trend appeared where slightly more queries were issued during *All-Ads*, compared to *No-Ads*. However, these differences did not reach statistical significance and therefore may be the result of chance.

7.10.2 Total document click count

In terms of the total number of documents clicked on during a search, for both FA and NC, Low-PS participants always clicked on fewer documents than High-PS participants. Yet, this difference only reached statistical significance for NC participants during *All-Ads*, where a medium positive correlation occurred that was significant (r= 0.34, $p = 0.04$).

However, differences emerged between the two PS tests when specifically comparing clutter being present versus absent. For FA, both Low-FA and High-FA clicked on fewer documents during *No-Ads*. Yet, for NC, although the same pattern occurred for High-NC participants, Low-NC participants clicked on fewer documents during *All-Ads*. This might imply that especially for Low-NC participants, the negative effects of clutter were most prominent. However, as these differences were not significant, they can only be perceived as possible trends.

7.10.3 Total relevant document click count

Although mostly trends emerged for how many documents were clicked on, for the number of documents clicked overall *that were known to be relevant*, multiple significant findings emerged.

In *All-Ads*, for participants in both FA and NC, medium significant positive correlations occurred (FA: r= 0.39, $p = 0.02$; NC: r= 0.36, $p = 0.03$). This was further emphasized when analysing the means of the Extreme Groups, where Low-PS participants clicked on significantly fewer relevant documents in comparison to High-PS participants, for both FA and NC ($F(1, 20) = 19.56$, $p = 0.000$, $\eta_p^2 = 0.49$).

Yet in *No-Ads*, there were differences between participants of FA and NC: While Low-FA participants similarly clicked on fewer relevant documents—as had been the case when clutter was present—Low-NC participants actually clicked on *more* relevant documents amidst *No-Ads* (Low-NC: 5.42, High-NC:3.3). This pattern also appeared in the correlation analysis, where a negative coefficient occurred (r= -0.28). This further emphasizes that the two different tests of PS appear to be measuring different aspects of a participant's ability, given the differences emerging regarding their search behaviour: for Low-FA, clutter did not appear

to impact the number of relevant documents clicked on, but for Low-NC, clutter negatively impacted participants.

7.10.4 Time session overall

When analysing the total time a participant spent completing their search task overall, minimal correlations occurred. However, when the average-PS participants were removed from analysis, to only focus on the Extreme Group means, patterns did emerge.

In FA, Low-FA participants always spent *less* time completing their search task, in comparison to High-FA participants, and pairwise comparisons identified that this difference was significant during *All-Ads* (Low-FA: 361.2s, High-FA: 458.48s). Furthermore, both Low-FA and High-FA participants spent longer searching during *All-Ads*, in comparison to *No-Ads*, reaffirming previous findings on visual clutter more generally.

In NC, theories of visual clutter were also confirmed, where both Low-NC and High-NC took longer searching during *All-Ads*, in comparison to *No-Ads*. However, the only time when Low-NC participants spent longer searching, in comparison to High-PS participants, was when *No-Ads* was present—which again highlights the differences between each PS test.

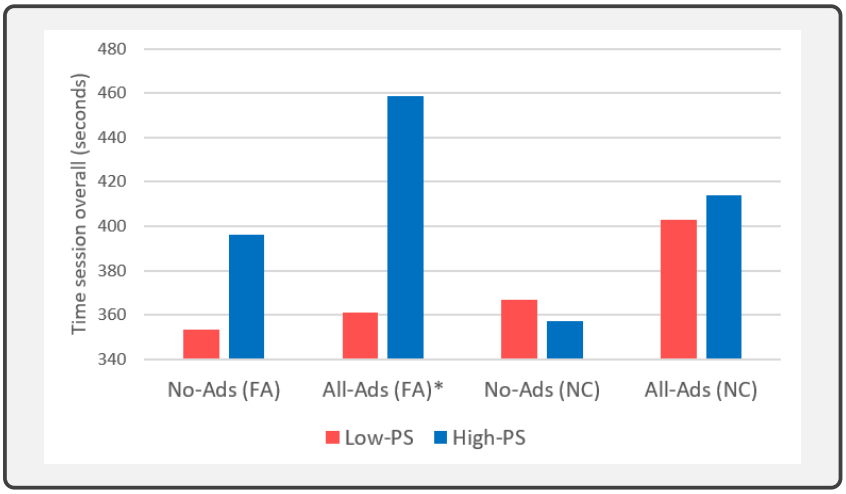


Figure 7.4: The length of time it took to complete the search task, for participants of different PS ability in both the ***Finding A's*** test (FA) and ***Number Comparison*** test (NC), in both the *No-Ads* and *All-Ads* search conditions. * denotes a significant difference between Low-PS and High-PS.

7.10.4.1 Time spent on SERP

On further inspection for where differences in the time session overall may have arisen from, a positive correlation occurred during FA and *No-Ads*, for the time spent on the SERP (r= 0.28). Additionally, the Extreme Group Analysis showed that Low-FA participants spent significantly less time on the SERP (73.39s) in comparison to High-FA participants (137.2s), who took almost double the amount of time during *No-Ads* ($F(1, 16) = 8.93$, $p = 0.009$, $\eta_p^2 = 0.36$).

The same pattern of Low-FA participants spending less time on the SERP, in comparison to High-FA participants was also observed in *All-Ads*, although this difference was not significant. However, directly comparing *No-Ads* against *All-Ads*, Low-FA participants always took longer searching amidst clutter, whereas High-FA participants took less time searching amidst clutter.

However, for NC, both Low-NC and High-NC participants took longer searching during *All-Ads*, in comparison to *No-Ads*. Furthermore, for *All-Ads*, a negative correlation occurred (r= -0.20) and Low-NC participants spent longer on the SERP (128.48s) in comparison to High-NC participants (112.16s). Yet, during *No-Ads*, Low-NC (107.96s) and High-NC (110.02s) were almost identical.

7.10.4.2 Time spent on documents

Although time spent on the SERP showed significant differences between participants depending on clutter being present or absent, the time spent examining documents did not exhibit such differences. The correlations for both PS tests were minimal (r ranged from 0.01 to 0.12), and no significant differences occurred for the Extreme Group Analysis. However, a consistent pattern did emerge where both Low-PS and High-PS participants always took longer examining documents during *All-Ads*. But combining everything together, the differences between participants regarding the time spent searching overall was mainly driven by the time a participant spent on the SERP, and not examining documents.

7.11 Clutter Results: User Experience

Measures of User Experience were categorised by three distinct surveys, all of which had a different aim: one was to gain participant perception of the task undertaken; one aimed at gathering information about how the participant felt; and one focused on participant perception of the system. Results from each category are discussed separately below.

7.11.1 Task

Firstly considering how a participant perceived the tasks undertaken, the median responses in both *No-Ads* and *All-Ads*, are reported in Table 7.3 and discussed in detail in the below subsections.

Table 7.3: The medians observed in different questions from the Task Survey, for users with Low-PS (Low) and High-PS (High), in both *No-Ads* (NA) and *All-Ads* (AA), for both the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test.

Task Survey	Perceptual Speed	NA (FA)	AA (FA)	NA (NC)	AA (NC)
Topic difficulty, *(1 = Very easy –> 5 = Very difficult)*	**Low**	1	3	2	3
	High	3	3	3	3
Topic learning *(1 = Nothing –> 5 = I know details)*	**Low**	4	3	4	3
	High	4	3	4	3

7.11.1.1 "How difficult was it to find relevant documents for this topic?"

When specifically asked how difficult it was to find relevant documents for both Low-PS and High-PS participants, when analysing the medians achieved in each group, for High-FA participants, there was general agreement that the task was neither very easy, nor very difficult, whether clutter was present or absent (median= 3 for both). However, for Low-FA this differed, as a median of 1 was reported in the *No-Ads* condition, which meant participants believed finding relevant documents was "Very easy" there. This was in comparison to *All-Ads*, where a higher median of 3 was reported by users with Low-FA.

Although Low-NC participants also reported it to be easier to find relevant documents during *No-Ads* versus *All-Ads*, the task was overall perceived as slightly more difficult, as a median of 2 was reported during *No-Ads* (in comparison to the median of 1 being reported during *No-Ads* by Low-FA).

On further analysis of the survey results through graphical representation, there was a clear pattern that the task was believed to be more difficult during *All-Ads*, compared to *No-Ads* (See Figure 7.5). Additionally, although there may have been slight differences in the medians between FA and NC, Figure 7.5 demonstrates that the same overall pattern occurs, regardless of whether participants were classed as Low-PS or High-PS in either PS test: when clutter is present, the task is perceived to be more difficult, and this difference is greatest for participants with Low-PS.

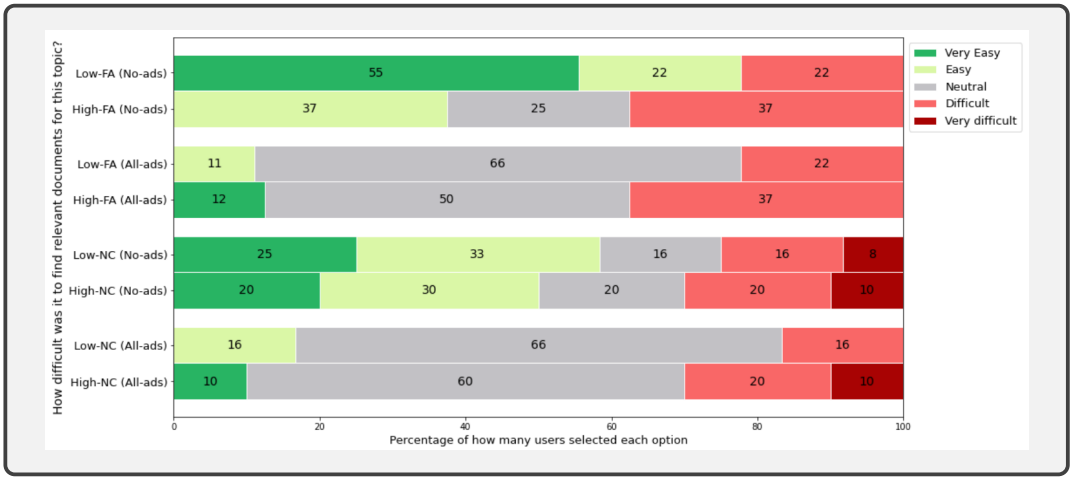


Figure 7.5: The percentages for how many participants responded with each option on the Likert-type scale for how difficult it was believed to be, to find relevant documents during the search task, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in both the *No-Ads* and *All-Ads* condition.

7.11.1.2 "How much did you learn about this topic?"

Analysing the medians, in both FA and NC, Low-PS and High-PS participants always believed they learned the most when no clutter had been present during the search (*No-Ads*: 4, *All-Ads*:

3).

Furthermore, the median analysis implied that there were minimal differences between Low-PS and High-PS, as the medians for Low-PS and High-PS were equivalent in every condition.

These patterns were also demonstrated on the graph (See Figure 7.6), where both FA and NC depict similar patterns: when clutter is present, both Low-PS and High-PS believe they learned less.

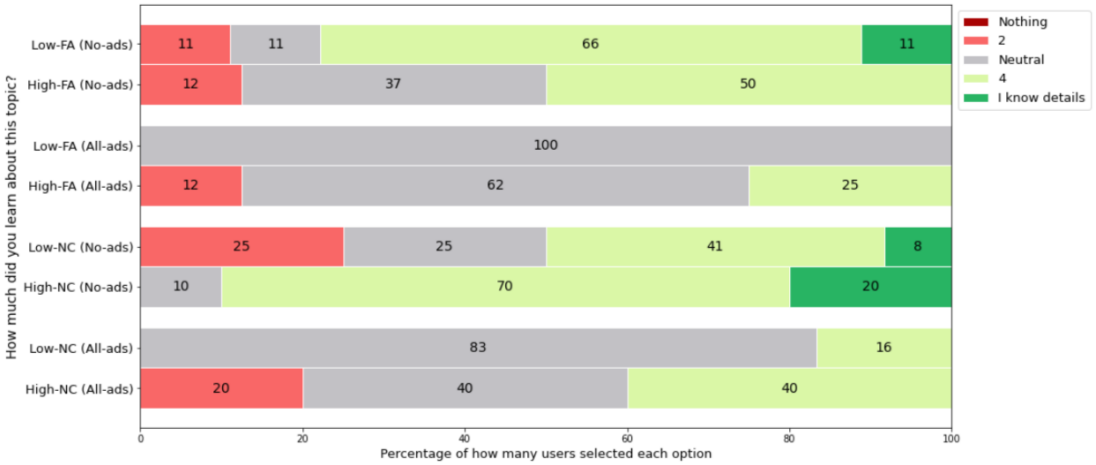


Figure 7.6: The percentages for how many participants responded with each option on the Likert-type scale for how much a participant believed they learned after their search, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in both the *No-Ads* and *All-Ads* condition.

7.11.2 User

Next considering how a participant felt, the median responses in both *No-Ads* and *All-Ads*, are reported in Table 7.4 and discussed in detail in the below subsections.

Table 7.4: The medians observed in different questions from the User Survey, for users with Low-PS (Low) and High-PS (High), in both *No-Ads* (NA) and *All-Ads* (AA), for both the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test. For all four questions, the Likert-type scale ranged from 1: 'Strongly Disagree' → 5: 'Strongly Agree'.

User Survey	**Perceptual Speed**	**NA (FA)**	**AA (FA)**	**NA (NC)**	**AA (NC)**
Perception frustration	Low	2	2	2	4
	High	3	3	2	3
Perception tiredness	Low	2	2	1	2
	High	2	2	2	2
Perception confidence	Low	4	4	5	3
	High	4	4	4	4
Perception enjoyment	Low	4	4	4	3
	High	3	3	4	3

7.11.2.1 "I felt frustrated while doing the task"

Looking at the medians reported for participants in FA, it appeared that whether clutter was present or absent did not affect the levels of frustration reported by participants. Yet, Low-FA participants consistently reported less frustration than High-FA participants, in both conditions.

However, when comparing the medians reported for participants in NC, opposing patterns emerged. Firstly, in *No-Ads*, there were no differences reported between Low-NC and High-NC. Yet in *All-Ads*, Low-NC participants reported more frustration than High-NC participants.

Secondly, differences between the conditions did emerge: more frustration was reported after *All-Ads*, in comparison to *No-Ads*. This difference was most notable for Low-NC participants because during *No-Ads*, their median was 2 which indicated that participants disagreed with the statement that they felt frustration; whereas during *All-Ads*, the median was double, at 4, showing more agreement that frustration had been felt.

However, observing the graphical representations of the results revealed more patterns—which did appear to be consistent between FA and NC. For both FA and NC, more frustration was reported during *All-Ads*, in comparison to *No-Ads*, regardless of PS. Furthermore, the most frustration—-as reported by the only times when participants indicated they 'Strongly Agreed' that they felt frustration—was by High-PS participants during *All-Ads*.

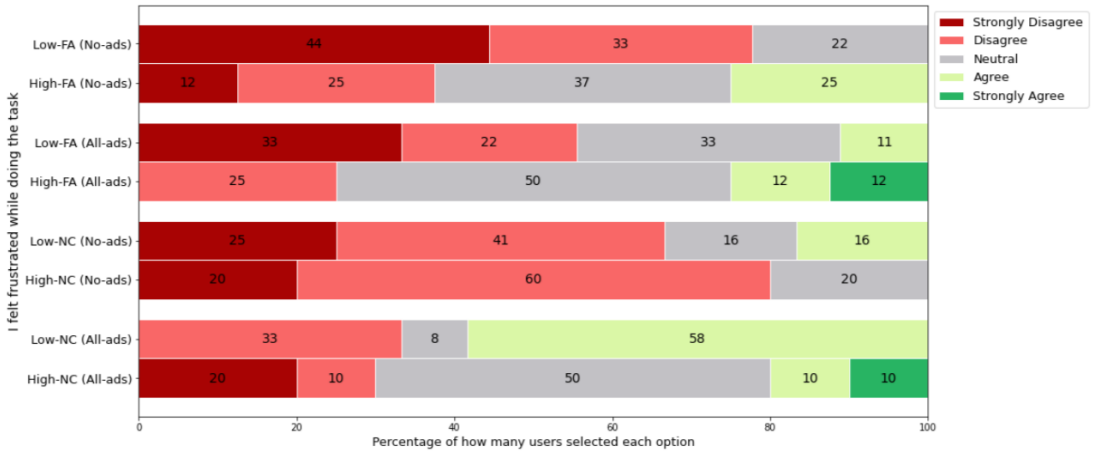


Figure 7.7: The percentages for how many participants responded with each option on the Likert-type scale for how frustrated they had been whilst doing the search task, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in both the *No-Ads* and *All-Ads* condition.

7.11.2.2 "I felt tired when completing this task"

In almost all conditions and for both Low-PS and High-PS, the median was 2, which demonstrated a fairly equivalent amount of tiredness reported, regardless of clutter or PS. The only

one that differed was for NC and *No-Ads*: participants with Low-NC reported that they 'Strongly Disagreed' (a median of 1) that they felt tired when completing the task. Therefore, Low-NC appeared to benefit the most, in terms of tiredness levels, when there had been no clutter.

Further analysing the graphical representation of all responses, it was discovered that both Low-NC *and* Low-FA benefited the most without clutter: the only time where participants never 'Agreed' or 'Strongly Agreed' that they felt tiredness, was during *No-Ads* for both Low-FA and Low-NC. Thus, this demonstrates that no clutter was eliciting the least amount of tiredness from Low-PS users.

Additionally, inspection of the percentage for how many participants reported that they 'Strongly Disagreed' also supports the conclusion that no clutter was preferable for Low-PS users. Specifically, in FA, 44% of Low-FA participants 'Strongly Disagreed' that they felt tired during *No-Ads*, and yet during *All-Ads*, this reduced to 11%. Similarly, in NC, 41% of participants 'Strongly Disagreed' during *No-Ads*, and again this reduced to only 16% during *All-Ads*. Given that strongly disagreeing with the statement meant that the participant was *less* tired, then it is clear that for both FA and NC, Low-PS participants reported less tiredness when no clutter was visible. However, for High-PS participants, there was equivalent amount of strongly disagreeing between conditions, which demonstrates that they were less affected by clutter, in terms of tiredness reported.

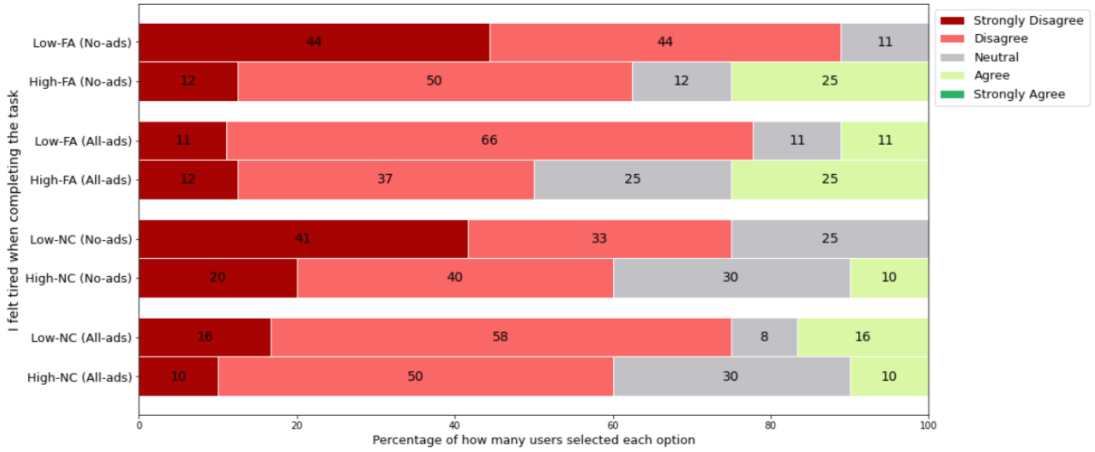


Figure 7.8: The percentages for how many participants responded with each option on the Likert-type scale for how tired they felt when completing the search task, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in both the *No-Ads* and *All-Ads* condition.

7.11.2.3 "I was confident in my decisions"

For FA, all participants reported the same confidence in their decision making in both *All-Ads* and *No-Ads* (median = 4). Whilst this was also the median for High-NC participants, for Low-NC participants, a slightly different pattern emerged: during *No-Ads*, the median was

5, highlighting that participants strongly agreed that they were confident; whereas during *All-Ads*, the median reduced to 3, indicating that participants neither agreed or disagreed in their confidence of decisions. Therefore, whilst confidence was generally similar across participants and conditions, for Low-NC participants, no clutter seemed the most optimal condition for them.

The graphical representation of all scores further emphasised the median trends: for FA, little differences emerged between conditions, and yet for NC, Low-NC participants were most confident during *No-Ads*.

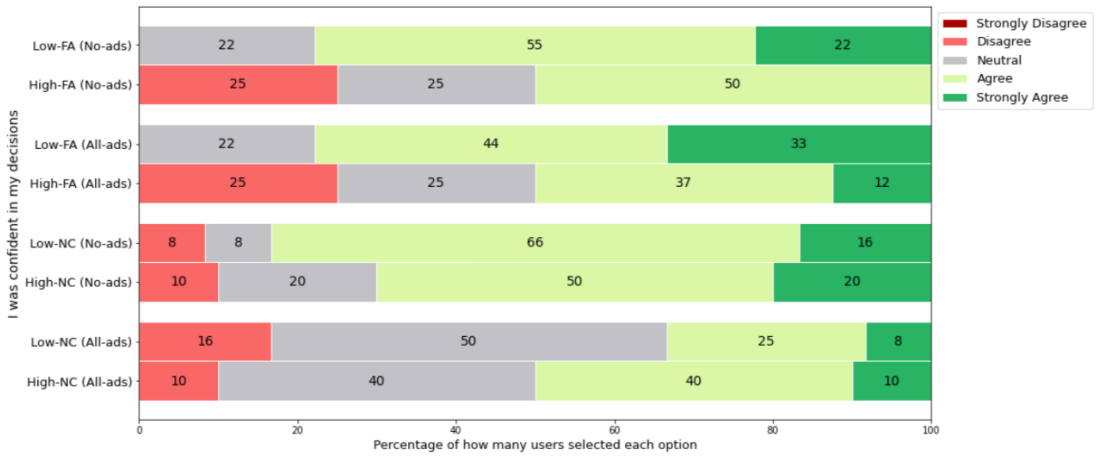


Figure 7.9: The percentages for how many participants responded with each option on the Likert-type scale for how confident they were in their decisions, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in both the *No-Ads* and *All-Ads* condition.

7.11.2.4 "I enjoyed completing this task"

In terms of enjoyment completing the task, for FA, although Low-FA reported more enjoyment than High-FA, there were no differences between *No-Ads* and *All-Ads*. However for NC, both Low-NC and High-NC had equivalent medians, which showed that more enjoyment was reported during *No-Ads*, in comparison to *All-Ads*.

Yet, when analysing the graphical representation of all scores, it was evident that for *both* FA and NC, participants reported more enjoyment during *No-Ads*, and this difference was especially prominent for participants with Low-PS.

7.11.3 System

Finally, considering how a participant perceived the system, the median responses in both *No-Ads* and *All-Ads*, are reported in Table 7.5 and discussed in detail in the below subsections.

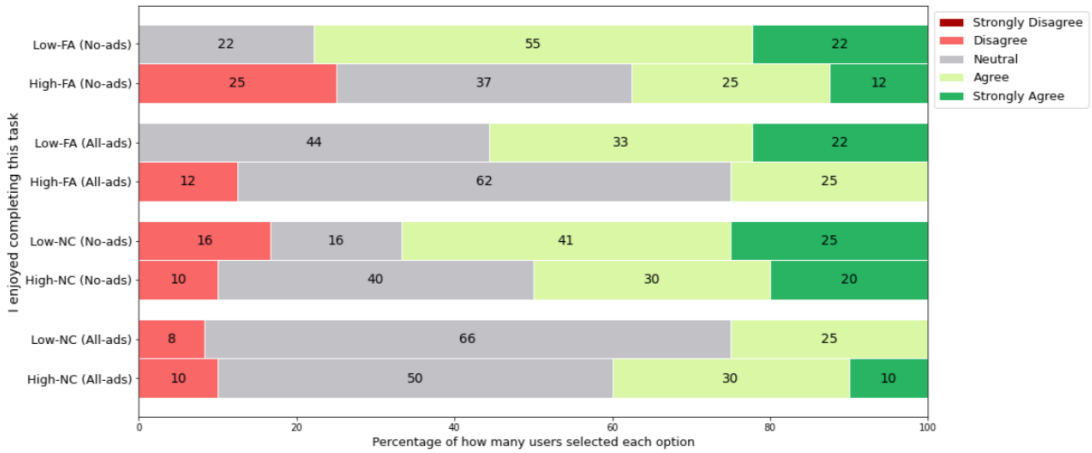


Figure 7.10: The percentages for how many participants responded with each option on the Likert-type scale for how much they enjoyed completing the task, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in both the *No-Ads* and *All-Ads* condition.

Table 7.5: The medians observed in different questions from the System Survey, for users with Low-PS (Low) and High-PS (High), in both *No-Ads* (NA) and *All-Ads* (AA), for both the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test. For all four questions, the Likert-type scale ranged from 1: 'Strongly Disagree' → 5: 'Strongly Agree'.

System Survey	Perceptual Speed	NA (FA)	AA (FA)	NA (NC)	AA (NC)
System boring	Low	4	3	3	3
	High	3	2	3	3
System annoying	Low	2	2	2	3
	High	2	2	2	3
System aesthetics	Low	3	3	3	3
	High	3	4	4	3

7.11.3.1 "The system was boring"

For FA, both Low-FA and High-FA reported that the system was more boring during *No-Ads*. Furthermore, there was a slight trend that Low-FA always perceived the system to be more boring than High-FA participants.

Yet for NC, by only examining the median, reports of how boring the system was did not differ between clutter conditions, or PS.

However, when examining the graphical representation, a trend can also be seen for NC, where both Low-NC and High-NC agreed that the system was more boring during *No-Ads*.

7.11.3.2 "The system was annoying"

Despite agreement that the system was less boring amidst the presence of clutter, for both Low-PS and High-PS in both PS tests, this did not transfer onto how annoying the different conditions were perceived. Specifically, only participants in NC (both Low-NC and High-NC) believed that the system was more *annoying* during *All-Ads*. Yet for FA, the same median of 2

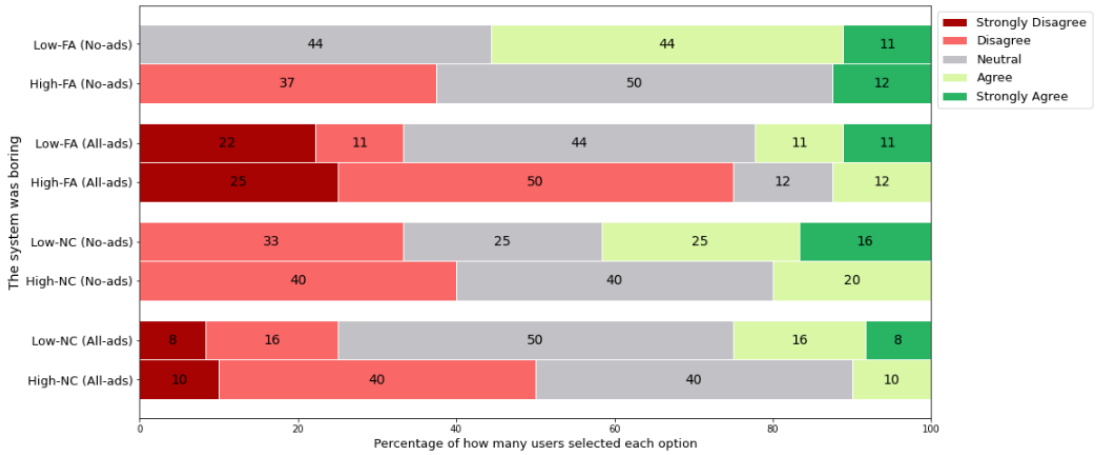


Figure 7.11: The percentages for how many participants responded with each option on the Likert-type scale for how boring the system was, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in both the *No-Ads* and *All-Ads* condition.

was reported for both Low-FA and High-FA, in both *No-Ads* and *All-Ads*, implying that clutter did not affect perceptions of annoyance there.

However, on further inspection of the percentage for how many participants agreed the system was annoying, it became evident that even for FA, more participants agreed the system was annoying in the presence of clutter. For example, 11% of Low-FA users and 0% of High-FA users agreed it was annoying in *No-Ads*, and this was compared to *All-Ads*, where 22% of Low-FA users and 12% of High-FA users reported annoyance.

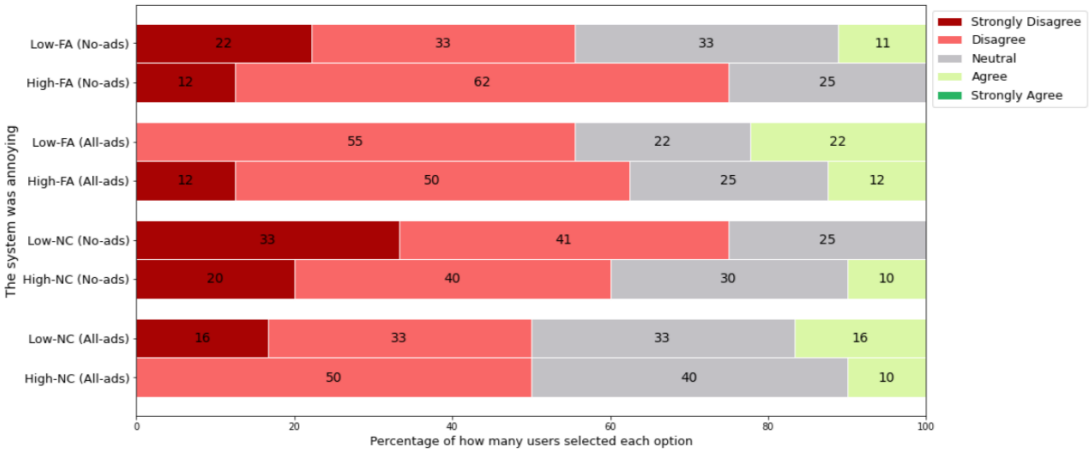


Figure 7.12: The percentages for how many participants responded with each option on the Likert-type scale for how annoying the system was, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in both the *No-Ads* and *All-Ads* condition.

7.11.3.3 "The system was aesthetically appealing"

As a final measure of system perception, participants were asked to rate how aesthetically appealing the system was. For Low-PS participants as measured by both PS tests, and in either *No-Ads* or *All-Ads*, no differences emerged.

However, for High-PS participants, opposing patterns emerged between the two PS tests: High-FA participants believed the system was more appealing during *All-Ads*, but High-NC participants believed the system was more appealing during *No-Ads*. This further demonstrates the differences between these two PS tests.

Although the medians revealed no differences in aesthetic appeal between conditions for Low-PS participants, on inspection of the graphical distribution, this confirmed that Low-FA believed the aesthetic appeal to be identical, whether clutter was present or absent. Yet, for Low-NC, differences did emerge, where more participants strongly agreed that the system was aesthetically appealing during *No-Ads*, and more participants disagreed during *All-Ads*.

For High-PS, specifically High-FA participants, the graph mirrored the medians: there was more agreement that the system was aesthetically appealing during *All-Ads*. However, for High-NC participants, although the medians implied that *No-Ads* were perceived as the most aesthetically appealing system, 30% of participants explicitly disagreed that the system was appealing during *No-Ads*, and 0% disagreed during *All-Ads*. Therefore overall, the patterns of results for High-PS participants were similar, regardless of PS test used: clutter created more aesthetic appeal. This highlights a clear divide between Low-PS and High-PS, with Low-PS participants perceiving more aesthetic appeal *without* clutter, and High-PS participants perceiving more appeal *with* clutter.

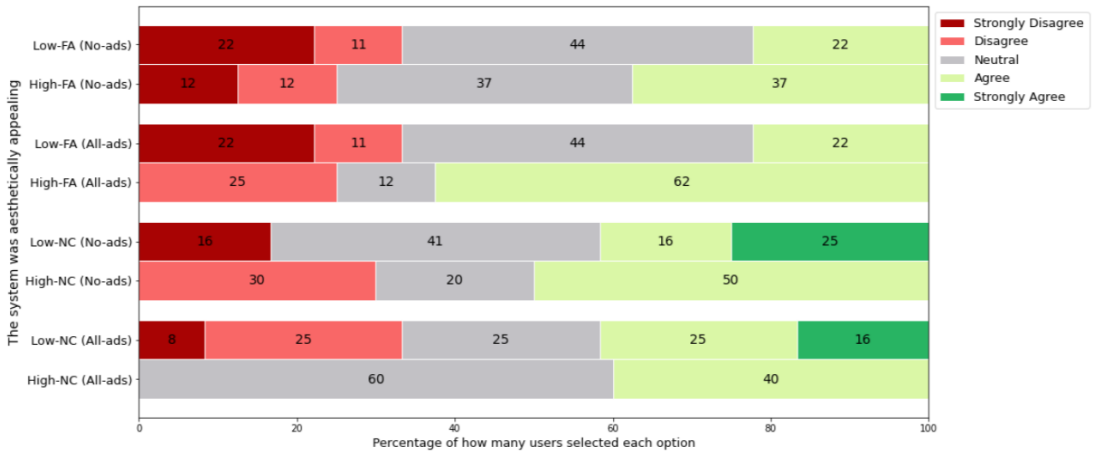


Figure 7.13: The percentages for how many participants responded with each option on the Likert-type scale for how aesthetically appealing the system was, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in both the *No-Ads* and *All-Ads* condition.

7.12 Summary of RQ2 Results

Given the many dependent variables investigated, this section provides a summary of the main results found, broken down by the two different PS tests, in order to answer the following research questions: ***(RQ2) "What is the relationship between Perceptual Speed and visual clutter in the form of advertisements during an IIR task?"***; and ***(2a) "Are there different effects on search outcomes, based on different PS tests?"***

7.12.1 Finding A's

7.12.1.1 Performance

Overall, during *No-Ads*, there were minimal differences between Low-FA and High-FA for the number of documents saved during a search task. Yet, during *All-Ads*, Low-FA participants saved more, and High-FA participants saved less. However, an increased number of saved documents did not correspond with any more documents that were known to be relevant, and thus clutter was negatively affecting Low-FA participants the most, causing them to have their lowest accuracy in the *All-Ads* condition.

When looking at other forms of accuracy, mainly Relevant-Saved / Relevant-Hovered, the negative effects of clutter were further emphasized for Low-FA participants. Additionally, this variable showed that Low-FA could actually achieve a higher accuracy compared to High-FA participants, so long as they were undertaking the *No-Ads* condition. Whereas, High-FA participants showed less differences in accuracy between clutter being present or absent, and in some measures—namely, Relevant-Saved / Total-Saved— an increase in accuracy was observed during *All-Ads*.

However, for post-task concept recall, regardless of Low-FA or High-FA, the negative effects of clutter were identified, with both Low-FA and High-FA participants retrieving their least amount of concepts in *All-Ads*. Yet, even during *No-Ads*, Low-FA participants still had worse memory in comparison to High-FA participants. This demonstrated that although Low-FA participants can perform searches more accurately during *No-Ads*, their post-task memory is still more negatively affected than High-FA participants.

7.12.1.2 Behaviour

In terms of behaviour, in order for Low-FA participants to achieve their highest accuracy during *No-Ads*, this condition was also associated with the fewest number of document clicks and spending significantly less time completing the task. Thus it appeared that when no clutter was present, Low-FA participants were able to expend less energy, in comparison to High-FA participants, to achieve the same—if not slightly better—outcome as High-FA participants.

In contrast, given that High-FA participants achieved slightly higher accuracy during *All-Ads*, this was the condition with more document clicks, and more time was spent completing

the search. Therefore, in order for High-FA participants to achieve their highest accuracy, they had to expend more energy and time.

7.12.1.3 Experience

Although it was mainly Low-FA participants who benefited from *No-Ads*, both Low-FA and High-FA participants believed that during *All-Ads*: the task was perceived to be more difficult; participants believed they learned less; more frustration was perceived; less enjoyment was reported; and the system was perceived as more annoying.

There were however also some differences in User Experience between Low-FA and High-FA. Firstly, whilst Low-FA believed the aesthetic appeal to be identical, whether clutter was present or absent, High-FA participants believed the system was more appealing during *All-Ads*. Secondly, participants with High-FA had equivalent tiredness levels in both *No-Ads* and *All-Ads*. Yet, Low-FA reported more tiredness during *All-Ads*. This reaffirms that it is Low-FA participants who are most negatively affected by clutter.

7.12.2 Number Comparison

7.12.2.1 Performance

Similar to FA, Low-NC participants achieved their highest search accuracy during *No-Ads*, and were most negatively affected by the presence of clutter. Furthermore, High-NC participants were the opposite: their highest performance occurred during *All-Ads*. However, again similar to FA, both Low-NC and High-NC achieved their worst post-task accuracy during *All-Ads*, and High-NC always obtained higher post-task accuracy compared to Low-NC.

7.12.2.2 Behaviour

Again similar to FA, in order for Low-NC participants to achieve their highest accuracy during *No-Ads*, this condition was also associated with less time spent completing the task. However, differently from FA, Low-NC participants clicked on more documents during *No-Ads*. Thus, despite taking less time, they were able to get through more documents and overall perform their best.

In contrast, but also similar to FA, given that High-NC participants achieved slightly higher accuracy during *All-Ads*, this was also the condition with more document clicks, and more time was spent completing the search.

7.12.2.3 Experience

Just as was found for FA, for NC, during *All-Ads*, both Low-NC and High-NC reported that: the task was more difficult; they learned less; more frustration was felt; they had less confidence in their decisions; less enjoyment completing the task; and the system was more annoying.

Furthermore, the differences in tiredness levels reported between Low-NC and High-NC were also the same as FA, where Low-NC participants reported more tiredness during *All-Ads*, and High-NC participants' tiredness levels were unaffected by clutter being present or absent.

The only difference in user experience reported in NC, as opposed to FA, was that Low-NC participants reported more aesthetic appeal during *No-Ads*. However, the finding that High-NC participants reported more aesthetic appeal during *All-Ads* was the same as High-FA participants.

7.13 Discussion

With contradictory results being found in previous literature with how PS affected IIR, the present experiment sought to investigate the main research question of **"What is the relationship between Perceptual Speed and visual clutter in the form of advertisements during an IIR task?"**, and then, the sub-research question of **"Are there different effects on search outcomes, based on different PS tests?"**. Summarising all results together, although there were subtle differences found between PS tests, overall, the same patterns in search performance, behaviour, and experience were observed: when no clutter was present, both Low-FA and Low-NC participants had a more positive search experience, both subjectively, and objectively, whilst taking less time to complete their task. In contrast, High-FA and High-NC participants were able to achieve their highest search accuracy during clutter. Yet, this required their longest search time and was still associated with more negative perceptions of the task, in addition to performing worse in the post-task concept recall.

Given the present results found, it became evident that the null hypothesis (H6) of "*ads will not conform to the same negative effects of visual clutter, and thus a participant's search behaviour and performance will be unaffected by ad presence, including those with Low-PS*" was rejected. This was despite the null hypothesis being created from the concept of 'banner blindness', where because users find ads annoying, they ignore them. Instead, the present research found that all participants, regardless of PS, believed the system with ads to be more annoying, which meant that H5 was supported. However, this annoyance did not translate into participants ignoring them, because the effects of ads generally negatively affected participants with Low-PS, but positively affected participants with High-PS.

Firstly considering the negative effects of clutter found for participants with Low-PS, this corresponded with the first hypothesis, where the presence of ads would conform to theories on visual clutter, and negatively distract participants with Low-PS from completing their IIR goals, resulting in: increased search time, worse search performance, and increased difficulty in finding relevant documents (H1). These results correspond with previous research on visual clutter, where increased clutter impairs efficiency of individuals performing search tasks by increasing response times and the number of errors made [2, 167, 238]. Furthermore, as the present results were most pronounced for participants with Low-PS, this also corresponds to previous research

in IIR, where although clutter was not explicitly mentioned, participants with Low-PS had significantly worse task performance during a webpage with higher visual complexity (Ziefle et al., 2015, cited in [44])—and therefore it makes sense that higher visual complexity corresponds with higher visual clutter.

Generally, the explanation for why clutter causes negative effects refers to the limits of attentional resources and short-term memory, resulting in a bottleneck in object perception, when clutter is present [144]. However, as clutter mainly negatively impacted participants with Low-PS, it is possible that users with Low-PS have more limited cognitive resources in general. Another explanation refers to research by Yeal-Lee & Cho [237], who identified that when users need to reduce their cognitive strain, they 'dump' parts of visible information, rather than trying to find a way to efficiently process it all. However, given that users with Low-PS have a lower fixation rate [209,211,222], they may be unable to identify which information can be 'dumped'. Instead, they may process the visible ads more than High-PS users, which causes an increase in search time, whilst also not helping them to achieve any more accuracy. Whilst it may be assumed that the use of eye-tracking could shed light as to the veracity of this explanation, theories on visual clutter more generally have highlighted the importance of peripheral vision cognitively overwhelming users: even without direct gaze, clutter could negatively impact a user [105]. However, it would be interesting to further investigate this, using eye-tracking, to identify whether the negative effects of clutter for participants with Low-PS are the result of *direct* or *peripheral* gaze, so that techniques could be designed to help improve perception.

For participants with High-PS, the negative effects of clutter were not entirely found. However, like Low-PS, High-PS had also spent longer searching during the presence of clutter, which rejects H2—which had hypothesised that increased clutter would result in search abandonment, and thus searches would be completed in less time. Instead, although High-PS participants also had an increased search time during *All-Ads*, and similarly reported increased difficulty in finding relevant documents, alongside other negative perceptions and user experiences—which *would* conform to theories on visual clutter— their search performance actually demonstrated their highest accuracy when clutter was present; at least, for active search performance during the task. These results would imply that for High-PS participants, their subjective experience did not coincide with their objective performance. This juxtaposition could have many possible explanations. Firstly, High-PS participants could have incorrectly believed the task was more difficult during *All-Ads*, despite actually performing their best there. This would not be the first time in IIR research where the relationship between physical effort and self-reported task difficulty was not linear. For example, Capra et al. [53] found that cognitively complex tasks took longer to complete, but were not associated with higher levels of difficulty or lower levels of satisfaction. Secondly, although High-PS participants performed their best during *All-Ads*, they also had to expend their most time during this condition, and therefore this could explain

the negative perceptions such as more difficulty and frustration being reported. However, this result would be in contrast to Jankowski et al. [115], who found that increased frustration led to abandonment of search. Alternatively, the negative subjective experience reported by High-PS participants could be due to an awareness that their memory was reduced amidst the presence of clutter, as even participants with High-PS were unable to recall as many concepts that they had learned post-task, in comparison to how many they could remember after no clutter had been present.

Regardless of the explanation for why participants with High-PS reported a negative search experience, but simultaneously had their highest active search performance during clutter, another open question remains as to why High-PS were able to combat some of the negative effects of clutter. These results corresponded with H3, where the presence of ads were hypothesised to increase user interaction and then result in higher search performance. This hypothesis was driven from research which identified that images can stimulate engagement and interest in news stories [91], and with increased interaction, this results in greater accuracy at completing the search task [16]. However, why this was the case for participants of High-PS, and not also Low-PS, remains unknown. Nonetheless, the results are similar to previous research, which identified that users with High-PS could compensate the negative influence of growing complexity better than people with Low-PS (Ziefle et al., 2015, cited in [44]). Consequently, it may just be that High-PS users are naturally better able to process visual information. Otherwise, it could also be that those with High-PS may be more goal-driven, favouring to persist or persevere with the search in order to reach a similar goal or level of performance for each task. Consequently, it would make sense why there was a difference between a negative subjective experience, but a positive search outcome.

Alternatively, the results found partially correspond with H4, which hypothesised that "*The presence of ads will reduce boredom and lower perceptions of workload*". Specifically, ads did reduce boredom, but this was for both High-PS and Low-PS participants. Yet, reduced boredom was not associated with lower workload. However, the hypothesis surrounding boredom does offer one explanation for why High-PS were negatively affected during *No-Ads*. Specifically, previous research had found that too few elements to process can create boredom amongst users, leading them to distraction by other things around them, or causing them to abort their search all together [186]. Therefore, it may be that users with High-PS, who have an increased fixation rate [209, 211, 222] require more visual stimuli in order to focus. If enough visual stimuli is not present, then they may be most affected by boredom, leading them to search abandonment. Given that High-PS participants did spend less time completing their search tasks during *No-Ads*, and yet they did not score as highly during this condition, this explanation makes sense.

Combining everything together, the present results have multiple implications for: a) better understanding the concept of PS; and b) creating a digital world that is more accessible for users, regardless of their individual abilities of perception. For the former point, the fact that low and high scorers in the two different PS tests had similar search outcomes and experiences supports the overall validity of PS measurement. Then, the fact that strikingly varied results occurred between participants and conditions really helps guide future system adaptation to support successful search. Whilst previously, there was a general consensus that users with lower levels of PS struggled with their search, in comparison to people with higher levels of PS, the present research has identified that Low-PS participants actually have the potential to achieve even better search outcomes than High-PS, given the right condition: so long as no clutter is visible—in the form of ads—which could cognitively overwhelm them. These findings thus not only support the conclusions in the Systematic Review which identified inconsistent findings in previous literature (see Chapter 5, Section 5.7) but they also create suggestions for designing future systems that can be tailored to suit the individual. For example, whilst previous research had only been able to attribute factors such as "*distraction*" and "*confusing*" as the reason for poorer performance amidst users with Low-PS [225], the present research extends this onto how interfaces can be made less confusing to a user with Low-PS: every effort must be made to ensure Low-PS users do not see clutter. This could be achieved through the use of ad-blockers or stripping back websites to read-only views. Whilst these adaptations may make the websites more boring, it would help a user with Low-PS to achieve a more positive and accurate search performance. In contrast, the elimination of clutter for users of High-PS would be detrimental, and therefore the presence of advertising for these users can be beneficial—unless the search task involved an element of memory, in which case, clutter should also be removed.

Whilst the present study found multiple interesting and significant results, it is not without limitations and many open questions do still remain. Firstly, all results in the present study were computed from either an average or sum of all queries. Yet, research is beginning to highlight the importance of dynamic search and how one query can influence another [241]. Thus, it would be worthwhile to investigate the temporal aspect of searching amongst clutter: whether clutter initially influences users on their first query, but gradually affects search outcomes less and less; or, whether clutter always has the same effect, regardless of time, remains unknown. Secondly, the present study was designed to explore whether clutter affects participants with different levels of PS. Yet, similar to previous research (e.g. Arguello & Capra [16]), the experiment was not designed to understand *why* this effect takes place. Thus, although different explanations for the current results have been discussed, without further research, these can only be speculations.

Furthermore, the current findings are based on a relatively small sample of only 37 participants, and limited to the context of clutter—as operationalised by ads—in a news search

environment. This inevitably limits the power to generalise results: with different participants and other contexts—such as entertainment or sports, or even exploratory search—different results may be found. Further research is therefore needed that examines ads of different kinds in various contexts, and with more participants. For example, perhaps when browsing social media, the presence of clutter would benefit users with Low-PS, as well as High-PS, because less boredom might be reported without the pressure of having an accurate search performance. It is also possible that given clutter can be defined in more ways than just quantity (See Section 2.7.5), that different types of clutter could impact results: some clutter may be beneficial for users with Low-PS, whilst other clutter may be negative for users with High-PS. This would be useful to know, as practically implementing no visible ads for users with Low-PS could be challenging in a digital world that is fuelled by marketing campaigns. Consequently, whilst the current study provides a promising start to developing adaptive systems that can accommodate individual users to perform positive searches with their highest performance, further research is needed that examines whether different kinds of clutter can help or hinder users with varied PS ability, to create practical systems that benefit search outcomes for all.

7.14 Chapter Summary

This Chapter provided the first experimental results—in addition to a discussion on possible explanations for these results— for how visual clutter impacted users with different levels of PS ability when completing an IIR task. Overall, it was found that users with Low-PS were most negatively affected by visual clutter: they took longer completing their search task, achieved lower accuracy, and reported more negative user experiences. Yet, when no clutter was visible, Low-PS users were able to achieve higher accuracy in comparison to users with High-PS. In contrast, High-PS users achieved their best search accuracy amidst clutter. Furthermore, although there were some differences reported between Low-PS and High-PS users (as measured by two different PS tests), overall, the main findings were consistent and thus the two PS tests did appear to be measuring the same general concept of PS. These results ultimately reinforce the need for future interfaces to be dynamic and adaptive to each individual user. However, before this can be practically implemented, the following Chapter 8 breaks the current results down further, to explore whether all clutter is bad for Low-PS users, or whether sometimes no clutter may be good for High-PS users.

Chapter 8

Evaluating how different types of clutter interact with PS during Information Retrieval

8.1 Chapter 8 Overview

This Chapter provides additional results from the experiment outlined in the previous Chapter 7. However, the specific focus of the current chapter concerns how *different types* of clutter impacted users with varying levels of PS ability when completing an IIR task. A similar format to the previous chapter is presented, where hypotheses are firstly outlined, followed by the results which are split into dependent variables which concern performance, behaviour, and experience separately. A summary of the main results is then provided, which additionally explains any differences found between PS, as measured by the two different PS tests. This is then concluded with an overall discussion, whereby possible explanations for the results found are discussed.

8.1.1 Background and Hypotheses

Given that the previous chapter identified that Low-PS participants seemed negatively affected by clutter, whereas High-PS participants were positively affected by clutter—during an IIR search task (See Chapter 7, Section 7.12)— the present chapter sought to identify whether the *type* of clutter impacted participants differently: some clutter may be beneficial for participants with Low-PS, whilst other clutter may be negative for participants with High-PS. Consequently, the third main research question, RQ3, sought to answer: '**Does clutter that is congruent with the task improve or worsen the search experience for users with Low Perceptual Speed?**', and in doing so, further hypotheses were identified.

Firstly considering that participants with Low-PS were most negatively affected by clutter—and previous research has identified that advertising clutter can be reduced by increasing the ad relevance to the website context [126]—one hypothesis was as follows:

- H1: *Congruent-Ads* will not conform to the same negative effects of clutter for participants with Low-PS, and therefore a participant's search performance and experience will only be negatively affected by the presence of *Incongruent-Ads*.

Secondly considering why High-PS participants benefited from clutter, the previous chapter provided one possible explanation: images can stimulate engagement and interest in news stories [91], and with increased interaction, this results in greater accuracy at completing the search task [16] (See Chapter 7, Section 7.13). However, other research has identified that only relevant images increase participant interaction, and subsequently improved accuracy [16]. Consequently, it is possible that participants with High-PS are also only positively affected when clutter is congruent:

- H2: High-PS participants will have positive search outcomes amidst *Congruent-Ads*, and negative search outcomes amidst *Incongruent-Ads*.

In other research that examined congruent information visible during websearch, participants were able to complete their search task faster when the surrounding images were relevant to the task [164]. This led onto the next hypothesis:

- H3: Both Low-PS and High-PS participants will complete their search task fastest during *Congruent-Ads*, in comparison to *Incongruent-Ads*.

However, whilst *Congruent-Ads* has generally been presumed the most optimal, for both Low-PS and High-PS participants, different research has highlighted that congruent ads result in greater eye fixations [110] and increased user attention [51, 52]. Then, if more attention was fixated on the congruent ad, and not the task, this may negatively affect participants:

- H4: When *Congruent-Ads* are visible, both Low-PS and High-PS will have more negative search outcomes and experiences, in comparison to when *Incongruent-Ads* are visible.

Whilst the above hypotheses have been focused on congruent or incongruent clutter visible, there was one other condition created in the present experiment: a mixture of both congruent and incongruent clutter visible simultaneously. This condition was created as previous research has highlighted the different attentional affects of perceiving ads when their relevance is unpredictable: when both congruent and incongruent ads are visible, *"users seem to get 'ad blind' so that even good ads receive less attention"* [52]. Thus, if a mixture of congruent and incongruent ads are ignored more, it is possible that their affect on a participant's search outcome will reverse the effects found in the previous chapter:

- H5: When *Mixed-Ads* are visible, a participant with Low-PS will *not* have a negative search performance and experience, and a participant with High-PS will *not* have a positive search performance and experience.

Alternatively, given the differences in eye fixation between Low-PS and High-PS users [209, 211, 222], it is possible that any clutter may affect a participant with Low-PS negatively, but positively affect a participant with High-PS:

- H6: The effect of ads will be similar for a participant, regardless of their congruence. Consequently, all ads will negatively affect a participant with Low-PS, and all ads will positively affect a participant with High-PS.

8.2 Analysis of User Study

Given that participants with varying PS levels were differently affected by the presence of clutter, the type of clutter was further examined, in order to answer RQ3, '**Does clutter that is congruent with the task improve or worsen the search experience for users with Low Perceptual Speed?**. Here, the analysis involved every search condition that a participant undertook: *No-Ads*, *Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*.

Furthermore, just as was done in Chapter 7, participants with Low-PS and High-PS in two individual PS tests were analysed separately. This was to answer the sub research question of ***(3a) Do different types of clutter impact users with different types of PS, as measured by different tests?***.

Therefore, analysis followed the same protocol as Chapter 7 (See Section 7.7).

8.2.1 Performance and Behaviour Analysis

8.2.1.1 Correlations

Firstly, individual correlations were conducted for each condition, for a participant's score in both PS tests. Considering that more conditions were analysed to answer RQ3, as opposed to RQ2, this resulted in the following 8 correlations that compared every dependent variable for a measure of Performance and Behaviour:

1. A participant's score in ***Finding A's***, against the dependent variable during *No-Ads*;
2. A participant's score in ***Finding A's***, against the dependent variable during *Congruent-Ads*;
3. A participant's score in ***Finding A's***, against the dependent variable during *Incongruent-Ads*;
4. A participant's score in ***Finding A's***, against the dependent variable during *Mixed-Ads*;
5. A participant's score in ***Number Comparison***, against the dependent variable during *No-Ads*;

6. A participant's score in ***Number Comparison***, against the dependent variable during *Congruent-Ads*;

7. A participant's score in ***Number Comparison***, against the dependent variable during *Incongruent-Ads*;

8. A participant's score in ***Number Comparison***, against the dependent variable during *Mixed-Ads*;

8.2.1.2 Analysis of Variance

Then, for every dependent variable that considered a measure of Performance or Behaviour, the means of the extreme PS groups were compared using a 2 (Low-PS vs High-PS) x 4 (*No-Ads* vs *Congruent-Ads* vs *Incongruent-Ads* vs *Mixed-Ads*) Mixed ANOVA, individually. Given that there were two PS tests, Low-PS and High-PS were analysed in different ANOVAs, based upon *Finding A's* (Low-FA vs High-FA), and *Number Comparison* (Low-NC vs High-NC).

8.2.2 Experience Analysis

Finally, analysis of User Experience involved reporting the medians for every condition (*No-Ads*, *Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*), for both Low-PS and High-PS in each PS test. Furthermore, a horizontal stacked bar graph was created for every survey response which depicted the percentage for how many participants selected each option on the survey. Each graph contained 16 rows: the first 8 responded to a participant with Low-FA and High-FA completing *No-Ads*, *Congruent-Ads*, *Incongruent-Ads*, *Mixed-Ads*, and the latter 8 referred to Low-NC and High-NC. As a reminder, for ***Finding A's (FA)***, participants with Low-PS were referred to as Low-FA, and High-PS as High-FA; and for ***Number Comparison (NC)***, Low-NC and High-NC depicted participants with Low-PS and High-PS.

8.3 Congruency Results: Performance

All results for measures of performance, in both *No-Ads*, *Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*, are reported in Table 8.1 and discussed in detail in the below subsections.

8.3.0.1 Total documents saved

Out of all conditions, for FA, the strongest correlation was during *Congruent-Ads* (r= -0.21), and this corresponded with Extreme Group Analysis, where Low-FA participants saved almost double (8.33 documents) in comparison to High-FA participants (4.78 documents). For the other clutter conditions, these also followed a negative correlation coefficient trend, and whilst Low-FA always saved more than High-FA, these differences were less apparent compared to *Congruent-Ads* (E.g. in *Mixed-Ads*, Low-FA: 4.89, versus High-FA: 4 documents). Yet whilst

Table 8.1: The results for every measure of performance for both PS tests, in the *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA) and *Mixed-Ads* (MA) condition. The row shaded in grey (denoted r) represents the correlation coefficient in each condition. The other rows present the means observed for users with Low-PS and High-PS. If a cell contain a *, this means an ANOVA found a significant difference between Low-PS and High-PS. If a † is present, this represents a significant correlation.

Performance	**Type**	**Finding A's**				**Number Comparison**			
		NA	**CA**	**IA**	**MA**	**NA**	**CA**	**IA**	**MA**
Total documents saved	r	-0.04	-0.21	-0.16	-0.14	-0.01	0.09	0.26	0.27
	Low-PS	4.67	8.33	4.89	4.89	5.58	6.0	4.08	3.75
	High-PS	4.89	4.78	4.0	4.0	5.8	8.0	5.1	5.8
Relevant-Saved / Total-Saved (%)	r	-0.09	0.01	0.14	0.43†	-0.26	-0.20	0.32	0.07
	Low-PS	65	68	49	48*	76*	64	40*	61
	High-PS	63	68	68	86*	50*	53	72*	65
Relevant-Saved / Relevant-Hovered (%)	r	-0.24	-0.23	-0.06	0.02	-0.05	-0.07	0.33†	0.18
	Low-PS	57	47	38	34	53	32	24	21
	High-PS	30	30	39	29	43	34	47	35
Concepts Recalled	r	0.19	0.19	0.22	-0.01	0.19	0.16	0.20	-0.04
	Low-PS	3.11	2.67	2.11	2.33	3.58	2.83	2.33	2.67
	High-PS	4.0	3.0	3.0	2.22	4.8	3.0	2.8	2.3

Low-FA always saved more documents in comparison to High-FA when clutter was present, this was the opposite to *No-Ads* (in *No-Ads*, Low-FA: 4.67, High-FA: 4.89).

For NC however, every condition followed the same trend as *No-Ads* during FA: Low-NC participants saved fewer documents than High-NC participants (E.g. in *Mixed-Ads*, Low-NC: 3.75, High-NC: 5.8 documents). This also corresponded with positive correlations occurring for every clutter condition, which contrasts to negative correlations occurring for every clutter condition during FA.

However, one similarity between FA and NC concerned where Low-PS saved their most amount of documents: for both, this occurred in the *Congruent-Ads* condition. For High-PS, although High-NC also saved their most during *Congruent-Ads*, High-FA saved their most during *No-Ads*.

It is important to note that although many patterns have been reported for the number of documents saved in each condition, none of these differences reached statistical significance, for either any correlation, or ANOVA. Consequently, although different patterns have emerged between the two PS tests, these could just be the result of chance.

8.3.0.2 Relevant-Saved / Total-Saved

Although Chapter 7 concluded that participants with Low-PS had a lower percentage of relevant documents saved when clutter was present, analysis of the different types of clutter revealed a slightly different picture. Specifically, for Low-FA, a slighter higher accuracy was observed during *Congruent-Ads* (68%), in comparison to *No-Ads* (65%). However, it is worth noting that these similar accuracy scores were despite Low-FA participants saving many more documents during *Congruent-Ads*, in comparison to *No-Ads*. Therefore, it appears that Low-FA partici-

pants were having to work harder, or potentially made more mistakes during *Congruent-Ads*, in comparison to *No-Ads*, and therefore overall, *No-Ads* is still preferable for participants with Low-FA.

Furthermore for Low-FA participants, whenever incongruent clutter was visible, through either *Incongruent-Ads* or *Mixed-Ads*, Low-FA participants were most negatively affected, as their accuracy dropped by 20%, to just 49% during *Incongruent-Ads*, and 48% during *Mixed-Ads*. Specifically during *Mixed-Ads*, this also revealed the largest difference between Low-FA and High-FA participants, as both a significant positive correlation occurred (r= 0.43, $p = 0.007$) and the ANOVA revealed significant differences between the means (Low-FA: 48% High-FA: 86%, $F(1, 16) = 5.05$, $p = 0.039$, $\eta_p^2 = 0.24$).

For High-FA participants, their accuracy of 86% during *Mixed-Ads* was their highest score, and their lowest was observed during *No-Ads* (63%). This reveals an opposite pattern to Low-FA, where High-FA do not benefit from *No-Ads*.

For NC, there were some similarities to FA: High-NC participants also had their lowest score during *No-Ads*, and the difference between Low-NC and High-NC here reached statistical significance ($F(3, 60) = 3.71$, $p = 0.016$, $\eta_p^2 = 0.16$). However, whilst High-NC also had a higher score during *Mixed-Ads*, their highest score was during *Incongruent-Ads*, and this also was a statistically significant difference in comparison to Low-NC participants, who gained their lowest accuracy during this condition (in *Incongruent-Ads*, Low-NC: 40%, High-NC: 72%). Thus, Low-NC participants were most negatively affected by *Incongruent-Ads*, and most benefited by *No-Ads*.

8.3.0.3 Relevant-saved/ Relevant-Hovered

In another measure of accuracy, similar patterns emerged to the previous measure: both Low-FA and Low-NC performed their best during *No-Ads*, and were most negatively affected whenever incongruent clutter was visible, either through *Incongruent-Ads*, or *Mixed-Ads*.

In contrast, High-PS participants did their best during *Incongruent-Ads*, but this had to be explicitly incongruent-only clutter. Whenever congruent clutter was visible, either directly *Congruent-Ads* or *Mixed-Ads*, then their performance lowered.

However whilst these differences emerged, the only condition which revealed significant differences was the correlation during *Incongruent-Ads* for participants in NC (r= 0.33, $p = 0.047$). This further demonstrated how Low-NC participants were much more negatively affected by *Incongruent-Ads* (achieving accuracy of only 24%) in comparison to High-NC participants (who achieved accuracy of almost double, at 47%).

8.3.0.4 Concepts Recalled

Although participants with Low-PS had shown higher search performance during *No-Ads*, in comparison to High-PS participants, for the number of concepts recalled post-task, the most

Table 8.2: The results for every measure of behaviour for both PS tests, in the *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA) and *Mixed-Ads* (MA) condition. The row shaded in grey (denoted r) represents the correlation coefficient in each condition. The other rows present the means observed for users with Low-PS and High-PS. If a cell contain a *, this means an ANOVA found a significant difference between Low-PS and High-PS. If a † is present, this represents a significant correlation.

Behaviour	**Type**	**Finding A's**				**Number Comparison**			
		NA	**CA**	**IA**	**MA**	**NA**	**CA**	**IA**	**MA**
Number of Queries	r	-0.04	0.05	0.03	-0.03	-0.04	0.16	0.00	0.03
	Low-PS	3.0	3.56	3.22	3.67	2.92	4.0	4.17	4.08
	High-PS	3.56	3.67	3.78	4.33	3.4	4.9	4.1	4.2
Document click count	r	0.10	0.04	0.21	0.10	0.00	0.31	0.17	0.32
	Low-PS	6.44	9.22	6.0	6.33	7.92	7.75	7.08	6.5
	High-PS	8.22	9.89	7.33	8.56	8.9	13.0	8.8	9.9
Relevant document click count	r	0.12	0.01	0.30	0.42†	-0.28	0.03	0.37†	0.27
	Low-PS	3.0	4.89	2.56	2.11*	5.42	4.08	2.42	3.33
	High-PS	4.33	4.89	4.67	6.0*	3.3	4.9	5.3	4.9
Time session overall	r	-0.06	0.25	0.04	0.21	-0.12	0.07	-0.17	0.09
	Low-PS	353.37	371.66*	361.96	349.98*	367.05	402.93	403.68	401.98
	High-PS	396.4	477.92*	423.26	474.25*	357.09	430.52	375.34	435.58
Time spent on SERP	r	0.28	0.15	-0.04	0.02	-0.05	-0.06	-0.19	-0.23
	Low-PS	73.39*	95.37	89.21	96.62	107.96	124.99	115.73	144.73
	High-PS	137.2*	131.17	109.36	130.31	110.02	127.24	100.62	108.61
Time spent on documents	r	-0.09	0.11	0.05	0.17	0.01	0.15	-0.05	0.14
	Low-PS	214.12	229.84	234.28	209.91	203.94	218.36	226.38	206.98
	High-PS	203.74	294.88	264.22	266.75	212.38	254.82	226.21	256.36

were always during *No-Ads*, for both Low-PS and High-PS.

Furthermore, although participants with Low-PS were most negatively affected by *Mixed-Ads* for Relevant-Saved/ Relevant-Hovered, it was *Incongruent-Ads* which produced the strongest correlation coefficients for the number of concepts recalled (FA: r= 0.22, NC: r= 0.20). When rounding to the nearest whole number, during *Incongruent-Ads*, Low-PS participants recalled roughly 1 less concept post task (2 concepts), in comparison to High-PS participants (3 concepts), in both FA and NC. Yet, when *Congruent-Ads* were present, Low-PS and High-PS participants retrieved the same number of concepts, when rounding to the nearest whole number (3 concepts).

8.3.0.5 Summary

Overall, the measures of performance have demonstrated that Low-PS are most negatively affected by incongruent clutter. When only congruent clutter is present, their performance increases, but their best performance and experience is always when no visible clutter has been present. In contrast, High-PS are most negatively affected by clutter that is only congruent, and instead do best when there is some form of incongruent clutter visible—except for post-task, their best recall always occurred after no clutter.

8.4 Congruency Results: Behaviour

All results for measures of behaviour, in *No-Ads*, *Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*, are reported in Table 8.2 and discussed in detail in the below subsections.

8.4.0.1 Number of Queries

For all conditions, minimal correlations were reported (e.g. r ranged from -0.04 to 0.05). However, looking at the Extreme Group means, some trends could be seen. Regardless of the condition, Low-PS participants always issued fewer queries in comparison to High-PS participants, and their fewest number of queries was during *No-Ads*.

Similarly, High-PS participants issued their fewest number of queries during *No-Ads*. However generally, minimal differences were observed between conditions. For example, for High-FA, the number of queries for each condition were: *No-Ads*: 3.56, *Congruent-Ads*: 3.67, *Incongruent-Ads*: 3.78, *Mixed-Ads*: 4.33. Yet rounding to the nearest whole number, 4 queries were issued in every condition.

8.4.0.2 Total document click count

Just as Low-PS participants always issued fewer queries than High-PS participants, Low-PS participants also clicked on fewer documents than High-PS participants, across every condition, for both PS tests. This also corresponded with the correlations, where positive correlations were always reported, with some being medium (e.g. r= 0.32 for NC, during *Mixed-Ads*).

Looking further at the Extreme Group means, other patterns emerged. For FA, generally Low-FA clicked on a similar amount of documents, when either *No-Ads* (6.44) *Incongruent-Ads* (6), and *Mixed-Ads* (6.33) were present. Yet during *Congruent-Ads*, roughly 3 more documents were clicked on (9.22). Similarly, High-FA also clicked on their most documents during *Congruent-Ads*.

For NC, High-NC participants similarly clicked on their most documents during *Congruent-Ads*. However, for Low-NC participants, their largest number of document clicks were for both *Congruent-Ads* and *No-Ads*.

8.4.0.3 Total relevant document click count

When looking at the different clutter conditions, for both PS tests, when *Congruent-Ads* were present, minimal correlations occurred (FA: r= 0.01, NC: r= 0.03). This was further emphasized when comparing the means of the Extreme Groups, where no differences emerged (e.g. Low-FA: 4.89, High-FA: 4.89).

However, when *Incongruent-Ads* were present, a medium positive correlation occurred in FA (r= 0.3) and Low-FA participants were clicking on roughly half the amount of relevant documents (2.56) in comparison to High-FA participants (4.67). Furthermore, when *Mixed-Ads* were present, a stronger correlation occurred (r= 0.42), and this was significant (p = 0.01). Here, Low-FA participants clicked on roughly a third fewer relevant documents (2.11) in comparison to High-FA participants (6)—and this difference also reached significance ($F(1, 16) = 6.78$, $p = 0.019$, $\eta_p^2 = 0.30$).

For participants in NC, similar patterns were observed, where there was less of a difference between Low-NC and High-NC during *Congruent-Ads*, and greater differences for *Incongruent-Ads* and *Mixed-Ads*. For *Incongruent-Ads*, this correlation was also significant (r= 0.37, $p = 0.02$).

Combining everything together, these results align with the performance measures: less relevant document clicks occurred during *Incongruent-Ads* or *Mixed-Ads* for participants with Low-PS, which corresponded with their lowest accuracy in these conditions. Yet for High-PS, less relevant documents were clicked during *No-Ads*, which again was the condition where their lowest accuracy had occurred.

8.4.0.4 Time session overall

As participants with Low-PS had a preference to perform best during *No-Ads*, this was also the condition whey they spent less time completing their search task. In contrast, High-PS participants also spent their least time searching during *No-Ads*, and yet, High-PS participants performed their worst during this condition. This shows an apparent difference in searching behaviour between Low-PS and High-PS.

Furthermore, differences emerged between the two PS groups. Firstly considering FA, Low-FA participants always spent less time completing the search task, in comparison to High-FA participants. These differences were especially prominent in *Congruent-Ads* and *Mixed-Ads*, where positive correlations occurred (r= 0.25 and 0.21, respectively). Although these correlations were not significant, significant differences were found when comparing the means of the Extreme Groups: Low-FA participants spent almost two minutes less completing their search amidst *Congruent-Ads* (Low-FA: 371.66s, High-FA: 477.92s), and High-FA participants spent more than two minutes longer completing their search amidst *Mixed-Ads* (Low-FA: 349.98s, High-FA: 474.25s, $F(1, 16) = 4.89$, $p = 0.042$, $\eta_p^2 = 0.23$). Yet, for both Low-FA and High-FA, their longest time was during *Congruent-Ads*.

For NC, different patterns occurred. For both *No-Ads* and *Incongruent-Ads*, in comparison to higher PS levels, participants with lower PS actually spent *longer* completing their search task (r = -0.12 and -0.17, respectively) and this was also the case when analysing the Extreme Groups (E.g. for *Incongruent-Ads*, Low-NC: 403.68s, High-NC: 373.34s). However, these differences did not reach significance. Furthermore, Low-NC participants spent a similar amount of time searching, regardless of clutter condition, but their least time searching was during *No-Ads*. Whereas, for High-NC, their longest searches were during *Congruent-Ads* and *Mixed-Ads*, and their least was also during *No-Ads*.

8.4.0.5 Time spent on SERP

For both Low-PS, and High-PS, an identical pattern emerged between PS groups for the condition where the least time spent on the SERP occurred. For Low-PS, this was during *No-Ads*,

and for High-PS, this was for *Incongruent-Ads*. However, whilst Low-PS spent their longest time on the SERP in *Mixed-Ads*, regardless of PS test, for High-FA participants, their longest was during *No-Ads*, and for High-NC participants, their longest search was during *Congruent-Ads*. However, as it is already known that participants with High-PS did not perform their best during either *No-Ads* or *Congruent-Ads*, this shows that spending longer on the SERP did not equate to a higher performance.

Furthermore, differences emerged between the two PS tests. Specifically, for FA, minimal correlations occurred, but analysis of the Extreme Group means showed that Low-FA participants always spent *less* time on the SERP, in comparison to High-FA participants, for all conditions. This difference was only significant during *No-Ads* ($F(1, 16) = 6.21$, $p = 0.024$, $\eta_p^2 = 0.28$).

Yet for NC, negative correlations occurred for every clutter condition, although this was greatest for *Mixed-Ads* (r= -0.23). Furthermore, although Low-NC participants similarly spent less time than High-NC participants amidst *No-Ads* and *Congruent-Ads*, there were only minimal differences between the groups (E.g. during *Congruent-Ads*, Low-NC: 124.99s, High-NC: 127.24s). Yet, for both *Incongruent-Ads* and *Mixed-Ads*, Low-NC participants spent *longer* on the SERP in comparison to High-NC participants, and this difference was greatest during *Mixed-Ads*, where Low-NC participants spent roughly 30 seconds longer (Low-NC: 144.73s, High-NC: 108.61s). Yet again, although these patterns differ between PS tests, the same overall trend is apparent: a longer time spent on the SERP did not equate to a higher search performance.

8.4.0.6 Time spent on documents

Except in the *Congruent-Ads* condition for NC, which showed a slight negative correlation of -0.05, and the Extreme Groups were almost identical (Low-NC: 226.38s, High-NC: 226.21s), every other clutter condition, for both PS groups, showed a positive correlation, where Low-PS participants always spent less time examining documents in comparison to High-PS participants. However, when *No-Ads* were present, whilst Low-NC also spent less time on documents in comparison to High-NC, Low-FA actually spent slightly longer in comparison to High-FA. However, none of these differences were significant, and only minimal correlations were observed.

8.5 Congruency Results: User Experience

Following the same order as the User Experience metrics reported in Chapter 7 (Section 7.11), for the results from the types of congruence visible, User Experience was also categorised by three distinct surveys, all of which had a different aim: one was to gain participant perception of the task undertaken; one aimed at gathering information about how the participant felt; and

Table 8.3: The medians observed in different questions from the Task Survey, for users with Low-PS and High-PS in both the ***Finding A's*** and ***Number Comparison*** PS test after *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA) and *Mixed-Ads* (MA). For Topic Difficulty, 1 = Very easy –> 5 = Very difficult; and for Topic Learn, 1 = Nothing –> 5 = I know details.

Task	**Type**	**Finding A's**				**Number Comparison**			
		NA	**CA**	**IA**	**MA**	**NA**	**CA**	**IA**	**MA**
Topic Difficulty	Low-PS	1	2	3	3	2	3	4	3
	High-PS	3	3	4	4	3	4	3	4
Topic Learn	Low-PS	4	3	2	3	4	3	3	3
	High-PS	4	4	4	2	4	4	3	3

one focused on participant perception of the system. Results from each category are discussed separately below.

8.5.1 Task

Firstly considering how a participant perceived the tasks undertaken, the median responses in *No-Ads*, *Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*, are reported in Table 8.3 and discussed in detail in the below subsections.

8.5.1.1 "How difficult was it to find relevant documents for this topic?"

Looking at the different conditions, for Low-FA, the medians for how difficult they believed the task was mirrored their actual performance: the task was believed to be more difficult during *Incongruent-Ads* and *Mixed-Ads* (medians = 3), slightly easier during *Congruent-Ads* (median = 2), but easiest during *No-Ads* (median = 1). This was also similar for Low-NC, except *Mixed-Ads* and *Congruent-Ads* were rated the same.

For High-PS participants however, the difficulty ratings did not correspond with their actual performance. Instead, High-FA rated the task easier during *No-Ads* and *Congruent-Ads* (medians = 3), and most difficult during *Incongruent-Ads* and *Mixed-Ads* (medians = 4), which contradicts their performance.

Yet for High-NC, the medians told another story: *Congruent-Ads* and *Mixed-Ads* were perceived as the most difficult conditions, and *No-Ads* and *Incongruent-Ads* easier.

Furthermore, observation of the horizontal stacked bar chart of the whole sample revealed the same patterns as the medians (See Figure 8.1).

8.5.1.2 "How much did you learn about this topic?"

Looking across the different conditions for how much a participant believed they learned, more patterns emerged. For FA, it was apparent that Low-FA participants believed they learned the most during *No-Ads*, and the least during *Incongruent-Ads*. Whereas, for Low-NC participants, whilst *No-Ads* was also the condition where most was believed to be learned, the medians

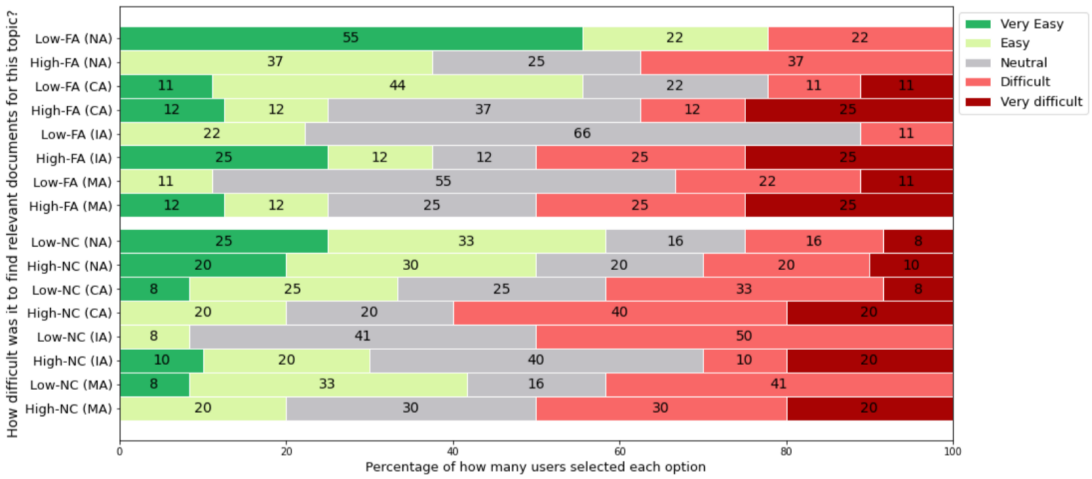


Figure 8.1: The percentages for how many participants responded with each option on the Likert-type scale for how difficult it was believed to be, to find relevant documents during the search task, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA).

were identical for *Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*. However, on further analysis using the horizontal stacked bar graph, it appeared that 41% of participants reported more negative experiences of learning less during *Congruent-Ads*, compared to only 25% of participants during *Incongruent-Ads* and *Mixed-Ads*. Thus, it appeared that *Congruent-Ads* was the worst condition in terms of learning, which did not correspond with the concepts recalled, as *Congruent-Ads* actually produced the most learning, compared to *Incongruent-Ads* and *Mixed-Ads*, for Low-FA participants.

For High-PS participants, based upon the medians, different patterns became evident. For High-FA, learning was believed to be equivalent during *No-Ads*, *Incongruent-Ads*, and *Congruent-Ads*. Only for *Mixed-Ads* did learning appear to be the least. Similarly, for High-NC, more learning was perceived during *No-Ads*, and less during *Incongruent-Ads* and *Mixed-Ads*. However, whilst the medians implied that *Congruent-Ads* were associated with the same learning as *No-Ads*, the graph (See Figure 8.2) makes clear that the most learning occurred for High-NC during *No-Ads* (90% reported a positive experience), compared to *Congruent-Ads* (where only 50% reported a positive experience).

8.5.2 User

Next considering how a participant felt, the median responses in *No-Ads*, *Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*, are reported in Table 8.4 and discussed in detail in the below subsections.

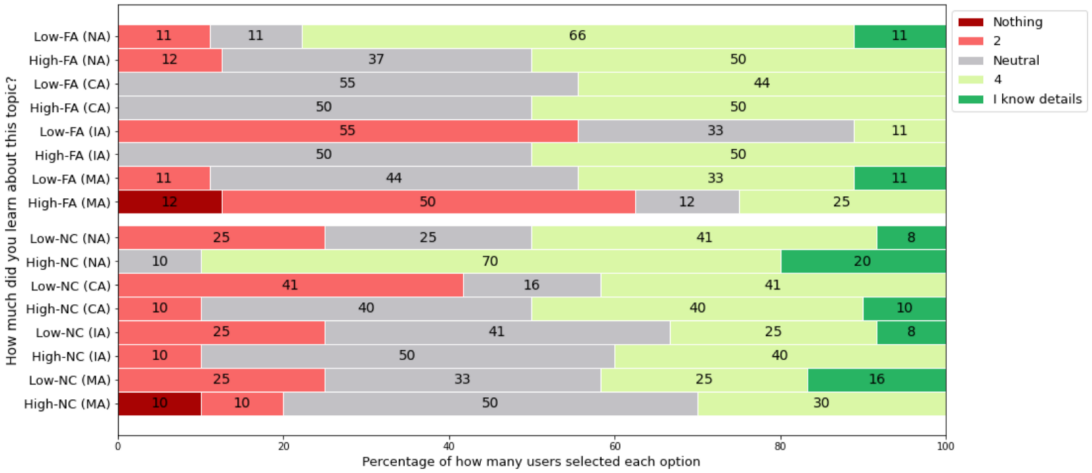


Figure 8.2: The percentages for how many participants responded with each option on the Likert-type scale for how much they believed they learned after their search, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA).

Table 8.4: The medians observed in different questions from the User Perception Survey, for users with Low-PS and High-PS in both the ***Finding A's*** and ***Number Comparison*** PS test after *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA). For all four survey questions, the Likert-type scale ranged from 1: 'Strongly Disagree' → 5: 'Strongly Agree'.

Perception	**Type**	**Finding A's**				**Number Comparison**			
		NA	**CA**	**IA**	**MA**	**NA**	**CA**	**IA**	**MA**
Frustration	Low-PS	2	2	2	2	2	4	4	3
	High-PS	3	4	3	3	2	4	2	3
Tiredness	Low-PS	2	2	2	2	1	2	2	2
	High-PS	2	2	3	2	2	2	2	2
Confidence	Low-PS	4	4	4	4	5	3	4	3
	High-PS	4	4	4	4	4	3	4	4
Enjoyment	Low-PS	4	4	4	4	4	4	4	3
	High-PS	3	4	3	3	4	3	4	3

8.5.2.1 "I felt frustrated while doing the task"

For FA, Low-FA participants appeared indifferently affected by the different conditions, where the same median was reported for all. However, for High-FA, the most frustration was reported during *Congruent-Ads*.

High-NC participants were similar, where most frustration was also reported during *Congruent-Ads*. However, for Low-NC, there was a clear pattern that *No-Ads* produced the least frustration, compared to all other clutter conditions.

Looking across the graph (See Figure 8.3), it appeared that despite subtle differences in medians between PS groups, the overall trends were comparable. For example, in Low-FA, the highest percentage of participants who 'Disagreed' or 'Strongly Disagreed' that frustration was

felt was observed during *No-Ads* (77%). This was compared to the other clutter conditions, which all had 66% disagreement. Thus, *No-Ads* elicited the least frustration for Low-FA participants, just as Low-NC participants.

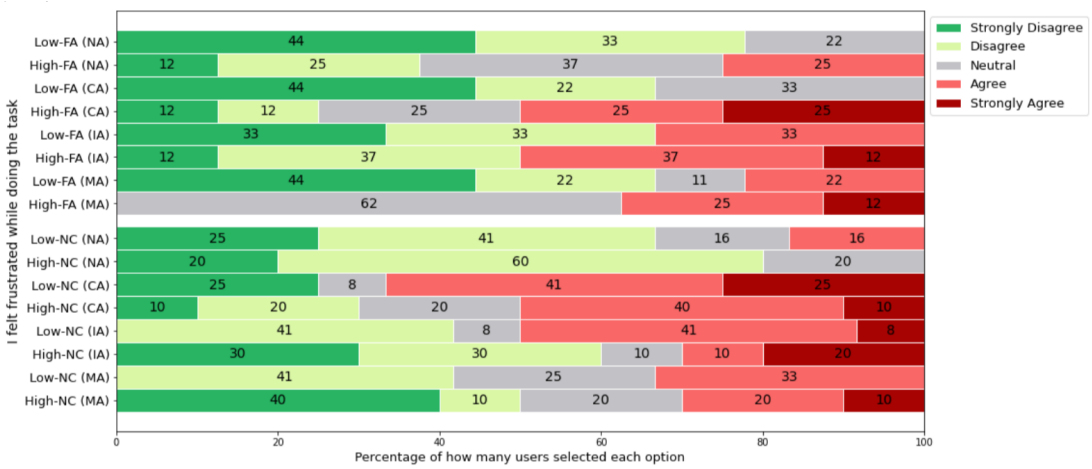


Figure 8.3: The percentages for how many participants responded with each option on the Likert-type scale for how frustrated they had been whilst doing the search task, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA).

8.5.2.2 "I felt tired when completing this task"

Firstly, Low-FA participants reported equivalent tiredness levels in all conditions (median = 2). The same median was also reported for all clutter conditions during Low-NC, however, when *No-Ads* was present, Low-NC participants reported less tiredness (median = 1).

For High-PS participants, the same levels of tiredness were also reported in almost all conditions (median = 2, in 7/8 conditions). The only condition which differed was a median of 3, for High-FA participants during *Incongruent-Ads*, demonstrating their most tiredness during this condition.

Although the medians implied no differences between conditions for Low-FA participants, the graph (See Figure 8.4) revealed that the least tiredness was during *No-Ads* (88% disagreed they were tired), compared to *Congruent-Ads* (77%), *Incongruent-Ads* (66%), and *Mixed-Ads* (66%). The same pattern can also be observed for Low-NC participants, where 41% 'Strongly Disagreed' they were tired during *No-Ads*, less during *Congruent-Ads* (25%), and even less during *Incongruent-Ads* (16%) and *Mixed-Ads* (8%).

For High-PS, the graph mirrored the medians: High-FA had their most tiredness during *Incongruent-Ads*. However, *Mixed-Ads* was their only condition where participants neither 'Agreed' or 'Strongly Agreed' that they were tired, which demonstrates that *Mixed-Ads* elicited less tiredness than all other conditions. This was also the same for High-NC.

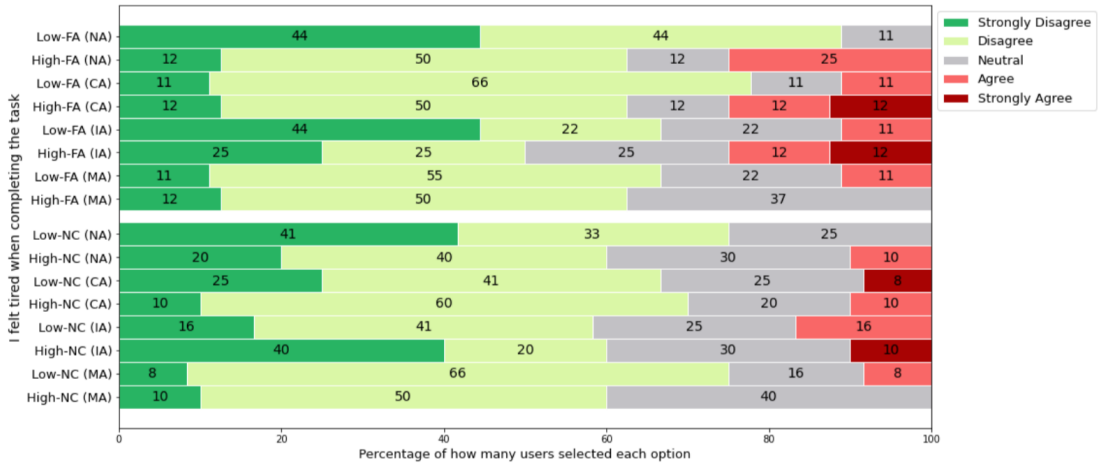


Figure 8.4: The percentages for how many participants responded with each option on the Likert-type scale for how tired they felt when completing the search task, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA).

8.5.2.3 "I was confident in my decisions"

For FA, the same median of 4 was reported in all conditions, regardless of Low-FA or High-FA.

Whereas for NC, Low-NC reported their most confidence during *No-Ads*, and their least during *Congruent-Ads* and *Mixed-Ads*. Then, for High-NC, an equivalent confidence rating of 4 was found during *No-Ads*, *Incongruent-Ads*, and *Mixed-Ads*. Yet a slightly lower median of 3 was reported for *Congruent-Ads*, implying less confidence was had during this condition.

However, looking at the graph (See Figure 8.5), more patterns can be seen. Whilst the medians for Low-FA were equivalent between conditions, it was evident that for *No-Ads* and *Congruent-Ads*, 0% of these participants 'Disagreed' or 'Strongly Disagreed' that they were confident, and thus these 2 conditions represent their most confidence.

For High-FA, it was also evident that more people 'Disagreed' or 'Strongly Disagreed' they were confident during *Congruent-Ads* (30%), compared to any other condition (*No-Ads*: 10%, *Incongruent-Ads*: 10%, *Mixed-Ads*: 10%). Consequently, *Congruent-Ads* appeared the worst condition for High-FA participants, in terms of their confidence. Therefore overall, the same patterns emerged between both PS tests.

8.5.2.4 "I enjoyed completing this task"

For participants with Low-PS, the same enjoyment levels (median = 4) were identified in all but one condition— *Mixed-Ads* for Low-NC participants (median = 3)—implying less enjoyment was experienced during *Mixed-Ads*.

For participants with High-PS, opposing patterns were observed between PS tests. For example, High-FA participants had their most enjoyment during *Congruent-Ads*, whereas for

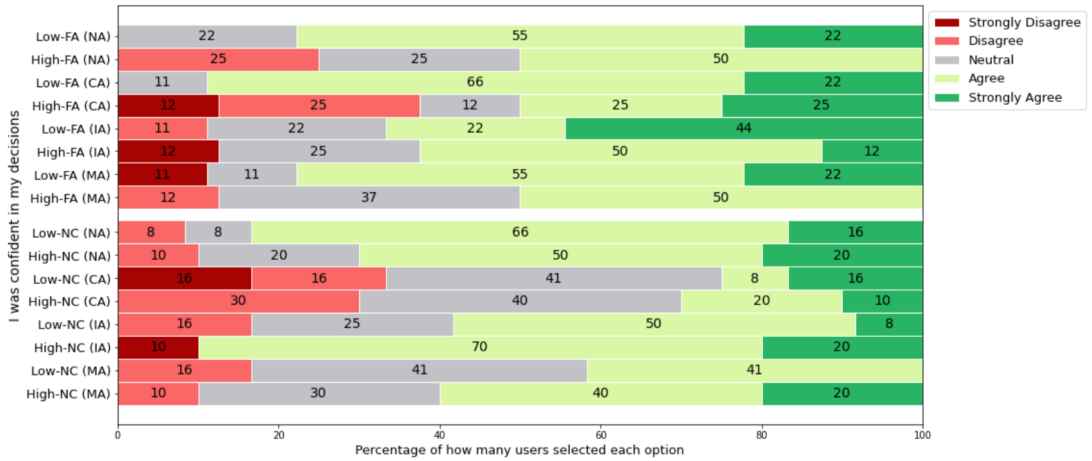


Figure 8.5: The percentages for how many participants responded with each option on the Likert-type scale for how confident they were in their decisions, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA).

High-NC participants, their joint least enjoyment was during this condition, alongside *Mixed-Ads*.

However, upon inspection of the graph, further trends can be seen. For Low-FA, 0% of participants 'Disagreed' or 'Strongly Disagreed' that they felt enjoyment, and thus *No-Ads* represented the most positive condition. Then for High-FA participants, although the median was highest during *Congruent-Ads*, this condition was actually the only time where participants 'Strongly Disagreed' that they felt enjoyment. This would then correspond with High-NC, where the highest percentage of participants also 'Strongly Disagreed' during *Congruent-Ads* (16%), compared to *No-Ads* (0%), *Incongruent-Ads* (10%), and *Mixed-Ads* (0%). Thus overall, *Congruent-Ads* appeared the least enjoyable condition for High-PS participants.

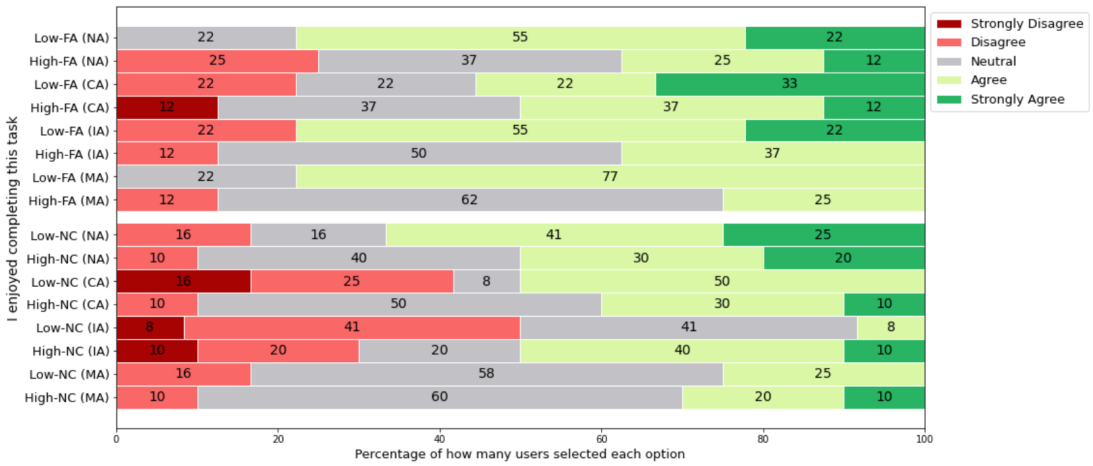


Figure 8.6: The percentages for how many participants responded with each option on the Likert-type scale for how much they enjoyed completing the task, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA).

Table 8.5: The medians observed in different questions from the System Perception Survey, for users with Low-PS and High-PS in both the ***Finding A's*** and ***Number Comparison*** PS test after *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA).. For all three survey questions, the Likert-type scale ranged from 1: 'Strongly Disagree' → 5: 'Strongly Agree'.

Perception	Type	Finding A's				Number Comparison			
		NA	CA	IA	MA	NA	CA	IA	MA
Boring	Low-PS	4	3	3	3	3	3	4	3
	High-PS	3	2	2	2	3	2	2	3
Annoying	Low-PS	2	2	3	2	2	2	3	3
	High-PS	2	2	2	3	2	3	3	3
Aesthetics	Low-PS	3	2	3	3	3	3	3	4
	High-PS	3	4	4	4	4	4	3	4

8.5.3 System

Finally, considering how a participant perceived the system, the median responses in *No-Ads*, *Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*, are reported in Table 8.5 and discussed in detail in the below subsections.

8.5.3.1 "The system was boring"

For FA, the same pattern emerged between Low-FA and High-FA participants: *No-Ads* was associated as more boring than all other clutter conditions, which were equivalent. For NC, High-NC participants also rated *No-Ads* as the most boring, but this was also rated the same as during *Mixed-Ads*. In contrast, for Low-NC participants, *Incongruent-Ads* was believed to be most boring.

Analysis of the graph (See Figure 8.7) revealed no further patterns than the medians had already identified.

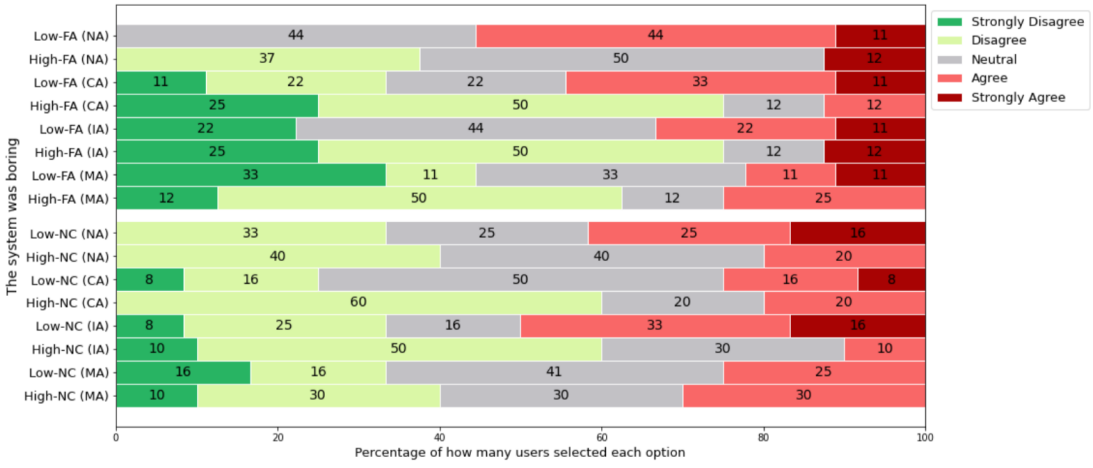


Figure 8.7: The percentages for how many participants responded with each option on the Likert-type scale for how boring the system was, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA).

8.5.3.2 "The system was annoying"

Firstly, Low-FA participants rated the most annoying system as *Incongruent-Ads*, and Low-NC believed that both *Incongruent-Ads* and *Mixed-Ads* were annoying. Yet, High-FA participants believed *Mixed-Ads* to be most annoying, whereas High-NC participants showed no difference between clutter conditions: all were rated more annoying than *No-Ads*.

Although *No-Ads*, *Congruent-Ads*, and *Mixed-Ads* had identical medians for Low-FA, the graph revealed that the least annoying system was during *No-Ads* (11% agreed it was annoying), in comparison to *Congruent-Ads* (22%) and *Mixed-Ads* (22%). *Incongruent-Ads* remained the most annoying, with 33% agreeing.

Similarly, for Low-NC, whilst the same medians were reported in *No-Ads* and *Congruent-Ads*, only *No-Ads* reported 0% of participants agreeing that the system was annoying, and thus this represented the condition of least annoyance. Then, whilst the most annoying were equally perceived as *Incongruent-Ads* and *Mixed-Ads*, the graph showed that more participants 'Agreed' or 'Strongly Agreed' that the system was annoying during *Incongruent-Ads* (33%), compared to *Mixed-Ads* (25%).

For participants with High-FA, the graph highlighted the same pattern as the medians. However, for High-NC, although the medians of all 3 clutter conditions were equivalent, *Incongruent-Ads* was the only condition where participants 'Strongly Agreed' that the system was annoying, demonstrating the most annoyance there.

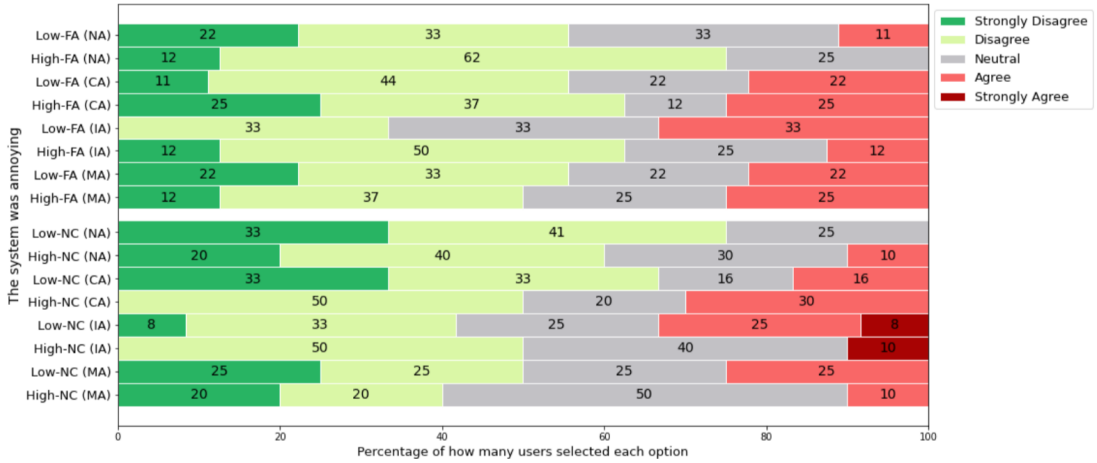


Figure 8.8: The percentages for how many participants responded with each option on the Likert-type scale for how annoying the system was, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA).

8.5.3.3 "The system was aesthetically appealing"

For how aesthetically appealing the system was believed to be, opposing results occurred between Low-PS and High-PS. The *least* aesthetically appealing was during *Congruent-Ads* for

Low-FA participants, and *No-Ads* for High-FA participants. However for NC, *No-Ads*, *Congruent-Ads*, and *Incongruent-Ads* were rated identically, for participants with Low-NC, and only *Mixed-Ads* was believed to be more appealing. Yet for High-NC, it was *No-Ads*, *Congruent-Ads*, and *Mixed-Ads* that revealed identical medians, and *Incongruent-Ads* was reported as the least appealing.

On inspection of the graph, although Low-NC participants had equivalent medians for 3 conditions, the highest percentage of participants who reported 'Strongly Agree' that the system was appealing was during *No-Ads* (25%), compared to *Congruent-Ads* (8%), *Incongruent-Ads* (16%), and *Mixed-Ads* (16%).

Furthermore, despite High-FA participants having equivalent medians for the 3 clutter conditions, it was apparent that the highest percentage of participants agreed the system was aesthetically appealing during *Congruent-Ads* (62%) compared to *Incongruent-Ads* (50%), and *Mixed-Ads* (50%).

For High-NC, *Congruent-Ads* also exhibited the highest percentage of agreement (60%), but this was also equal to *Mixed-Ads* (60%). However, this was double the amount of agreement in comparison to *Incongruent-Ads*, where only 30% agreed the system was aesthetically appealing.

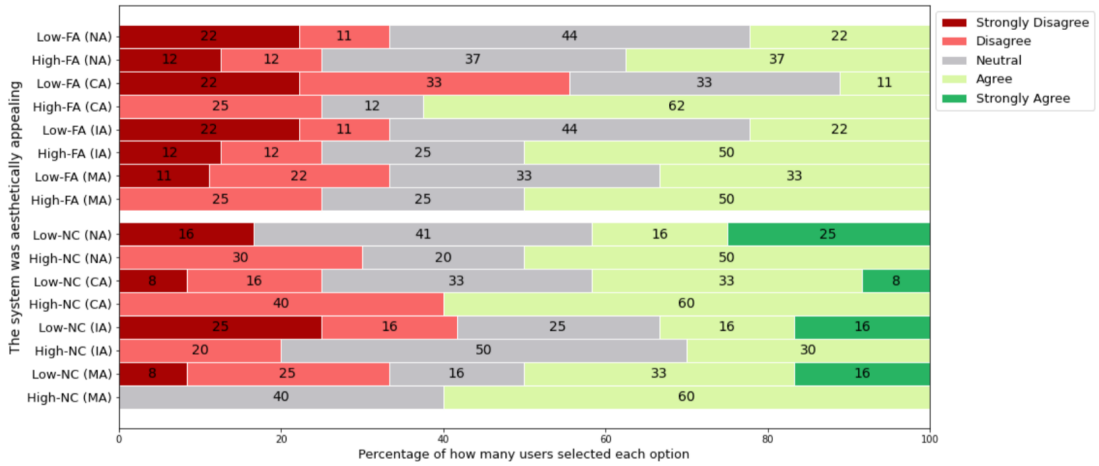


Figure 8.9: The percentages for how many participants responded with each option on the Likert-type scale for how aesthetically appealing the system was, by participants of either Low-PS or High-PS in the ***Finding A's*** (FA) and ***Number Comparison*** (NC) PS test, in *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA).

8.6 Summary of RQ3 Results

With many results for different dependent variables reported, this section provides a summary of the main results found, broken down by participants with Low-PS in each PS test, and then participants with High-PS in each PS test, in order to answer the following research questions: ***(RQ3) "Does clutter that is congruent with the task improve or worsen the search***

experience for users with Low Perceptual Speed?"; and ***(3a) "Do different types of clutter impact users with different types of PS, as measured by different tests?"***

8.6.1 Low-PS

8.6.1.1 Low-FA

For Low-FA participants, although one measure of performance (Relevant Saved / Total Saved) indicated that participants performed best during *Congruent-Ads*, their overall accuracy of how many relevant documents they had saved, out of how many they had hovered over, revealed that *No-Ads* was their most optimal condition. Similarly, *No-Ads* represented the condition where their highest post-task recall performance was also obtained. Yet whenever *Incongruent-Ads* or *Mixed-Ads* were present, their accuracy lowered in all measures of performance.

For measures of search behaviour, although Low-FA participants maintained a similar number of document clicks during *Incongruent-Ads* and *Mixed-Ads*, as they had clicked on during *No-Ads*, there were less relevant clicks during *Incongruent-Ads* and *Mixed-Ads*. This implies that whenever incongruent clutter was visible, either directly or as a mixture, Low-FA participants were somehow distracted by incorrect information. Instead, when *Congruent-Ads* were visible, significantly more documents were clicked on, and their longest time overall searching was obtained. This extra time and processing of more documents enabled them to achieve accuracy that was higher than *Incongruent-Ads* or *Mixed-Ads*, but it did not compensate for the negative effects of clutter overall.

In terms of User Experience, these results mirrored their actual performance. During *No-Ads*: the task was believed to be easiest; they believed they learned more; less frustration, tiredness, and annoyance was reported; but the system was ultimately perceived as the most boring. Then, the most difficult task, and most negative participant experiences, were found during *Incongruent-Ads* and *Mixed-Ads*. However, whilst *Congruent-Ads* was generally associated with positive experiences in comparison to *Incongruent-Ads* and *Mixed-Ads*—such as more confidence—*Congruent-Ads* was reported as the condition with least aesthetic appeal.

8.6.1.2 Low-NC

For participants with Low-NC, all measures of performance, both during and post-task, were best during *No-Ads*. Additionally, *Congruent-Ads* represented the best clutter condition, as worst performance was observed during *Incongruent-Ads* and *Mixed-Ads*. These results therefore correspond with Low-FA.

However, whilst Low-FA spent their longest time searching during *Congruent-Ads*, for Low-NC, there were minimal differences in search time between clutter conditions. Yet despite this similarity of search time, during *Congruent-Ads*, more relevant documents were able to be clicked on, again reaffirming that *Congruent-Ads* represented the most optimal clutter condition.

There were however differences in User Experience, as reported by Low-NC participants, in comparison to Low-FA participants: During *Congruent-Ads*, Low-NC participants believed they learned their least and had their least confidence, and yet this was their best clutter condition for how many concepts were recalled post-task. Furthermore, *No-Ads* did not represent the condition believed to be most boring: instead, the most boring was perceived as *Mixed-Ads*. However, for the other measures of User Experience, these corresponded with their actual performance, such as during *No-Ads*, there was least tiredness, less frustration, and most confidence reported. Similarly, both *Incongruent-Ads* and *Mixed-Ads* were overall reported as the most annoying.

8.6.2 High-PS

8.6.2.1 High-FA

In contrast to Low-FA, High-FA participants performed their best during *Incongruent-Ads* or *Mixed-Ads*, and their worst during *Congruent-Ads* or *No-Ads*. However, post-task recall remained highest during *No-Ads*, as had been the case for Low-FA.

Despite having a poor performance during *Congruent-Ads*, this was the condition where High-FA clicked on the most documents and spent their most time completing the task (almost one minute longer than during *Incongruent-Ads*). However, *Incongruent-Ads* and *Mixed-Ads* were not the conditions where they completed their task in the fastest times, as *No-Ads* always represented the least amount of time searching.

Despite a longer time searching during *Congruent-Ads*, this was actually the condition where High-FA participants rated the task as easier, and learning was perceived as equivalent to *Incongruent-Ads*. Furthermore, other measures of User Experience appeared to contrast their actual performance: *Congruent-Ads* was perceived as the most aesthetically appealing clutter condition; High-FA participants had their most enjoyment during *Congruent-Ads*; and despite *Incongruent-Ads* and *Mixed-Ads* being their best search performance conditions, it was *Incongruent-Ads* that was rated as most tiring, and *Mixed-Ads* as most annoying. However, some measures of User Experience did correspond with their overall performance: more frustration, and less confidence, was reported during *Congruent-Ads*.

8.6.2.2 High-NC

Similar to High-FA, High-NC participants also performed their worst during *Congruent-Ads* or *No-Ads*, but with their highest post-task recall during *No-Ads*. However, it was only *Incongruent-Ads* where participants performed their best overall, and not *Mixed-Ads*.

In terms of behaviour, *Congruent-Ads* was associated with the most document clicks, and *Incongruent-Ads* the least. Yet, more of these clicks were relevant during *Incongruent-Ads*, which corresponds with higher accuracy during this condition. Measures of time were then

similar to High-FA: almost a minute longer was spent searching during *Congruent-Ads*, in comparison to *Incongruent-Ads*, but *No-Ads* was the shortest search overall.

Although *Incongruent-Ads* was the best condition for High-NC participants search performance overall, and also required less time to complete the task, it was also believed to be the most annoying system to use, the least aesthetically appealing system, and greatest levels of tiredness were reported there. Furthermore, less learning was believed to have occurred during *Incongruent-Ads*, even although post-task recall was equivalent between *Congruent-Ads*, *Incongruent-Ads*, and *Mixed-Ads*. However, the other measures of User Experience did correspond with overall performance, as *Congruent-Ads* was rightly perceived as the most difficult and most frustrating task, and participants had their least confidence and enjoyment there.

8.7 Discussion of User Study

Given that the previous chapter identified that clutter has been found to affect participants with different levels of PS ability differently—with Low-PS being negatively affected, and High-PS being positively affected during IIR—the present chapter sought to investigate the next research question of '***How does the clutter congruence impact userss with different PS abilities during an IIR task?***, and then ***'(3a) Do different types of clutter impact users with different types of PS, as measured by different tests?'***.

Summarising everything together, there were *some* differences between the two PS tests for the different types of clutter visible. For example, whilst participants with Low-FA reported positive experiences during *Congruent-Ads*, Low-NC reported negative experiences there. Furthermore, although High-FA participants achieved their highest search accuracy during *Incongruent-Ads* and *Mixed-Ads*, *Mixed-Ads* was not associated with the highest accuracy for High-NC participants. However, the overall patterns remained similar between the two PS tests: the condition that participants with Low-PS performed most efficiently in was always *No-Ads*; and *Incongruent-Ads* or *Mixed-Ads* were associated with the most negative performance and experiences. Then, for High-PS, the opposite findings occurred: *Incongruent-Ads* was the condition where participants performed their best, and *No-Ads* their worst – except for post-task recall, which was always highest after *No-Ads*.

Firstly considering the results found for participants with Low-PS—with *Congruent-Ads* representing their best clutter condition, and anything containing incongruent clutter their worst (either *Incongruent-Ads* or *Mixed-Ads*)—this corresponded with the first expectation that was hypothesised, where a participant's search performance and experience would only be negatively affected by the presence of *Incongruent-Ads*, in comparison to *Congruent-Ads* (H1). These results are similar to previous research where relevant ads can reduce the perception of clutter [126], and therefore the final expectation (H6)—which stated that all ads would similarly affect a user's IIR— was rejected. However, relevant ads did not completely ameliorate

the negative effects of clutter altogether, as participants with Low-PS still performed best in the absence of clutter: they were able to expend their least amount of time, and yet gain their highest accuracy during *No-Ads*, demonstrating both efficiency and effectiveness. Instead, when *Congruent-Ads* were present, Low-PS participants had to expend a lot more time completing the task, in order to gain similar levels of accuracy. This finding contrasts with previous literature, where searches were completed faster when surrounding images were congruent [164], and thus H3 was also rejected. Furthermore, whilst *Mixed-Ads* also contained clutter that was congruent, this was not associated with any benefits for participants with Low-PS, which also rejects H5.

One explanation for understanding why Low-PS participants were able to perform better amidst *Congruent-Ads*, as opposed to when any incongruent clutter was visible (either directly during *Incongruent-Ads*, or as a mixture during *Mixed-Ads*), refers to the fact that Low-PS are known to have a lower eye fixation rate and struggle with processing what is in front of them [67]. Consequently, they may be unable to 'dump' additional information in order to reduce their cognitive strain [237], and subsequently get distracted by visible ads. Given that images have the power to convey meaning instantly whilst overcoming language barriers [129], if incongruent clutter is processed, this might provide false information to the Low-PS user that they are performing incorrectly. Additionally, users may become more frustrated, as incongruent clutter requires a greater cognitive effort to process, because a user must expend energy in working out why they have been shown something off-topic [27, 36, 128]. Indeed, in the present study, Low-PS participants did report more frustration after *Incongruent-Ads* or *Mixed-Ads*, in comparison to *Congruent-Ads*, which adds further support to this explanation. Furthermore, this explanation is akin to previous work done with users with dyslexia, who similarly have lower visual processing abilities [48], but who have noted a preference for topical images that help support their understanding of the search [171]: congruent ads ease information processing as users see more sense, or logic, in congruent associations [36]; and viewing a topical image can verify that the query issued matched the initial concept a user had in mind [208], which reassured the user that they were searching successfully [91, 171, 228].

Although *Congruent-Ads* was the most beneficial clutter condition for participants with Low-PS, this was not the case for participants with High-PS. Instead, High-PS participants performed their best searches during *Incongruent-Ads* and *Mixed-Ads*, which rejects H2, H5, and H6. This was despite the fact that *Congruent-Ads* was associated with increased interaction for High-PS participants — with more document clicks and a significantly longer time spent completing the task— and this corresponded to previous research which found that relevant images increased participant interaction [16]. Yet whilst increased interaction has previously been found to lead to better search outcomes [16, 188], this did not happen in the current experiment. Instead, High-PS demonstrated more positive performance measures

in *Incongruent-Ads* where they had interacted with the search system less. However, given that *Congruent-Ads* represented the clutter condition where worst performance occurred, this supported H4. However, *Congruent-Ads* was also the condition where High-FA participants spent their longest time searching, and for High-NC participants, searches during *Congruent-Ads* were completed almost a minute longer than *Incongruent-Ads*. Therefore, H3 was rejected, which expected searches to be completed fastest during *Congruent-Ads*. This finding was similar to participants with Low-PS, and thus also contrasts with previous literature, where searches were completed faster when surrounding images were congruent [164].

Given that users with High-PS are known to be able to scan accurately and quickly [10], it is perhaps surprising that they took so much longer examining *Congruent-Ads*, to then not achieve higher performance in terms of relevant documents saved or the accuracy of how many relevant documents saved had been hovered on. Instead, it appeared that *Congruent-Ads* were somehow negatively distracting participants, to spend more time on the task, but click on less relevant documents. Indeed, when *Congruent-Ads* had been present, a clear trend emerged that more frustration was reported, in comparison to any other condition. This contrasts with Low-PS, where more frustration was observed during *Incongruent-Ads*, presumably because a participant had to expend more energy in working out why they had been shown something off-topic [27,36, 128]. Therefore, understanding why *Incongruent-Ads* did not similarly frustrate participants with High-PS remains open to interpretation. One possible explanation is that participants with High-PS attended more to congruent clutter—which corresponds with previous research [51, 52, 110]—and thus too much attention was focused on the clutter, instead of the task. The use of eye-tracking in future studies should help to either accept or reject this theory. Alternatively, another explanation refers to High-PS participants being able to process more visual information [67], and thus they are easily able to identify that the congruent clutter is on topic. Consequently, as High-PS participants may realise that their search results do not *exactly* match the ads visible, it is possible that they continued their search until they could directly match an ad with the content of a document, even if it did not answer the task aim properly. However, this is just one hypothesis and more research would be needed to fully understand this.

Although *Congruent-Ads* were associated with more frustration for High-PS participants, for many other measures of User Experience, these appeared to contradict their actual performance. For example, more enjoyment occurred during *Congruent-Ads*, whereas *Incongruent-Ads* was perceived as most tiring. One theory for this contradiction could have referred to the User Experience being determined based on post-task recall, instead of active search performance. However, this was not the case, because similar recall was observed in *Incongruent-Ads*, as was found in *Congruent-Ads*. This therefore rejects two of the explanations provided in Chapter 7 (Section 7.13) for why more negative experiences were reported during clutter, despite an

overall positive performance there. Instead, the explanation that High-PS participants were just unaware of their search performance seems plausible, and this would correspond with previous research that has identified perception differs from actual performance [53]. This would then further confirm that High-PS and Low-PS have very difficult search experiences, with Low-PS being more aware of their performance.

Further comparing Low-PS against High-PS, it is evident that different search strategies were implemented, as well as different experiences. Specifically, when *No-Ads* was present, Low-PS participants were able to achieve their *best* search performance in the least amount of time, whereas High-PS did their *worst* search in their least time there. Why High-PS participants abandoned their search when *No-Ads* were visible, may have been due to increased boredom (See Chapter 7, Section 7.13). However, Low-PS participants also perceived most boredom during *No-Ads*, out of any other condition. This therefore indicates that boredom is responded to differently, depending on a participant's PS ability, and subsequently should be further investigated in future research.

Yet despite many differences between Low-PS and High-PS, one similarity refers to how participants perceived the *Mixed-Ads* condition. Although opposing results occurred—with Low-PS performing badly, and High-PS generally doing well there—*Mixed-Ads* was always perceived similarly to *Incongruent-Ads*; it was never associated with similar patterns to *Congruent-Ads*. For example, Low-PS participants struggled during both *Incongruent-Ads* and *Mixed-Ads*, and High-PS benefited from *Incongruent-Ads* and *Mixed-Ads*. This suggests that the presence of incongruent information was dominating the congruent information. Why this occurred remains unknown. However, *Mixed-Ads* was probably most similar to what would usually be visible in everyday search, because retrieval algorithms that attempt to provide congruent ads do not always succeed, which results in a mixture of congruent and incongruent ads visible simultaneously [192]. It is therefore possible that the *Mixed-Ads* condition was most akin to what participants have normally experienced online, and this would correspond with why previous literature believed that Low-PS perform poorer searches, and High-PS are more accurate (e.g. [9, 56]).

Bringing all results together, there are two main implications based upon the current experiment. Firstly, given that the overall pattern of results remain similar between how a Low and High participant in two different PS tests performed during an IIR task, this further validates that the two tests appear to be measuring the same concept of PS. However, as some differences were observed, this highlights that there are subtle differences between the two. Secondly, suggestions for designing future system development that can be tailored to suit the individual user are as follows. If a user with Low-PS must complete a specific search goal, then every effort must be made to remove *Incongruent-Ads* from the display. Although *Congruent-Ads* could be visible, this should only be in tasks where time is not an important factor. Instead, considering

that High-PS users actually perform better when incongruent clutter surrounds their search, either *Incongruent-Ads* or *Mixed-Ads* should be visible. Ranking algorithms that aim to retrieve relevant advertising should thus be reversed, so that irrelevant advertising appears. However, if a user with High-PS has to perform a task where they must prioritise their memory, then *No-Ads* would be the most optimal condition, and ad-blockers should be implemented.

Given the different optimal conditions for users with different PS ability, and depending on the task priority, these findings highlight the need for future systems to be dynamic and adapt to the user, based on their individual differences, alongside the task goal. This aligns with the aims of other research in the area, which has attempted to predict PS based upon eye-gaze, so that dynamic systems can be created [66, 210]. However, only the present research remains the first to have provided evidence of what an optimal search environment could look like: for Low-PS, the order of preference, from most to least optimal, would be *No-Ads*, then *Congruent-Ads*, and then *Incongruent-Ads* or *Mixed-Ads*; and for High-PS, it would be *Incongruent-Ads* or *Mixed-Ads*, followed by *Congruent-Ads*, and then *No-Ads* (unless memory was required). Whilst this implementation would be the most ideal, at first, it may be difficult to accomplish, as ranking algorithms that aim to retrieve relevant advertising do not have a 100% success rate [192]. However, further research may be able to suggest ways around this. Specifically, it may be possible that only ads in certain locations are influencing users, and thus this could guide where relevance ranking algorithms should position the most relevant, or least relevant ads.

As the location of the clutter was not presently studied, this highlights one limitation of the current experiment. Future research should therefore explore whether the location of clutter impacts a user with different PS ability. Other research has identified that right-hand side ads capture more attention in comparison to banner ads [128], and therefore this motivates that differences in search performance may be influenced by clutter location. Additionally, although every effort was made to keep ads consistent between conditions—such as by investigating their saliency—in line with Buscher et al. [52], we only considered ads to be topical or off-topic, and no other factors such as reputation of sponsor were considered. Given that clutter can be affected by various factors other than congruence—such as the colours present [124, 148] or size of visible input [168]—it is possible that different types of ads could impact users differently. For example, animations may not result in the same clutter effects as the static banners used in the present experiment. Thus with further research examining whether different kinds of ads conform to the same effects of congruent and incongruent clutter with users of varied PS ability, this will eventually lead to the most optimum systems that can accommodate and benefit search outcomes for every user.

8.8 Chapter 8 Summary

This Chapter presented the main results for how different types of clutter interacted with users of different PS ability (as measured by two different PS tests) when completing an IIR task. Overall, in order from most optimal, to least optimal, in terms of search accuracy, time, and experience, the main results identified that users with Low-PS should perform IIR tasks amidst no clutter, followed by congruent clutter; and incongruent clutter, or a mixture of congruent and incongruent clutter, should be avoided. In contrast, users with High-PS are the opposite, achieving their best search accuracy whenever incongruent, or a mixture of congruent and incongruent clutter, is present; and their worst amidst congruent only clutter, or no clutter. However, High-PS users always achieved their highest post-task recall after no clutter. Consequently, these results further emphasise the need for dynamic interfaces that adapt to users, based on not only their unique PS ability, but also the context of the task.

Part IV

Discussion

Chapter 9

Overall Discussion and Conclusion

9.1 Chapter 9 Overview

This Chapter begins by responding to the main research questions investigated in this thesis. Then, it widens onto an overall discussion of all results found, followed by the main contributions and implications of these results. A consideration of limitations in the research is then presented, which then leads onto proposals for future research. Lastly, the chapter concludes the thesis with final closing remarks.

9.2 Answers to Research Questions

In this section, a summary of answers to the overall research questions defined in Section 1.3 has been provided. As a reminder of the original main questions and sub-questions, these were:

1. RQ1: How can a user with Low Perceptual Speed be helped to achieve a more positive online search experience, both subjectively, and objectively?;
    (a) How has PS previously been measured during IIR?
    (b) What claims have occurred regarding PS in Computer Science?
    (c) How can previous results be explained?
    (d) How can a digital measurement of PS be created?
2. RQ2: What is the relationship between Perceptual Speed and visual clutter in the form of advertisements during an IIR task?;
    (a) Are there different effects on search outcomes, based on different PS tests?
3. RQ3: Does clutter that is congruent with the task improve or worsen the search experience for users with Low Perceptual Speed?

(a) Do different types of clutter impact users with different types of PS, as measured by different tests?

9.2.1 Research Question 1

For the first research question, in order to understand how a user with Low-PS could be helped to achieve a more positive online search experience and performance, this involved exploring multiple sub-research questions. Specifically, Chapter 4 presented themes which emerged from previous literature that indicated problems with how PS had previously been measured—including the test content, how it had been administered, and how results had been analysed and disseminated. All of these problems had not been considered in previous research. This ultimately made understanding how to help Low-PS users achieve a better search outcome difficult, as the validity of PS measurement was questioned, which then questioned the overall concept.

This awareness then led onto a Systematic Review, which categorised every claim that has occurred involving PS in the field of Computer Science, filtered into the categories of Search Performance, Behaviour, Time, User Experience, Physiology, and Interactions with other variables. Presented in Chapter 5, this revealed many contrasting findings in previous literature, with some studies claiming that Low-PS users *were* negatively affected, yet others claiming the opposite. Consequently, explanations for these contrasting results were explored, and it was concluded that the presence of visual clutter may have been impacting results: Low-PS appeared to struggle and perform search tasks less efficiently than High-PS users whenever interfaces were less organised, or contained many different elements, like videos and images simultaneously. Furthermore, it was speculated whether there might have been differences in search experiences as a result of different PS tests having been utilised.

Combining the findings together from both Chapter 4 and Chapter 5, it became evident that a new measurement of PS was needed. This was created and described in Chapter 6, where two separate tests were developed: a digital version of *Finding A's* and *Number Comparison*.

With the two new tests created, it was then possible to run an experiment to investigate how participants with Low-PS—as measured by different PS tests—could be helped during IIR, by considering two different factors discussed in the subsequent research questions, RQ2 and RQ3.

9.2.2 Research Question 2

Given the knowledge about how visual perception works in general, alongside the Systematic Review conclusions about how clutter may be an important factor that could explain differing results in previous literature, the second main research question focused on understanding the relationship between visual clutter and PS during an IIR task.

With the user-based experiment presented in Chapter 7—where clutter was operationalised as the presence or absence of ads—it was demonstrated that clutter *is* an important factor that can influence how participants of different PS ability completed an IIR search task. Specifically, when clutter was present, participants with Low-PS reported more negative user experiences, whilst simultaneously took longer to complete their search, despite obtaining poorer task performance and post-task memory. Additionally, the results revealed that Low-PS participants could achieve accuracy measures that were greater than High-PS participants, so long as no clutter was present. This therefore reaffirmed the conclusions from the Systematic Review, and offered support for why some previous studies have identified poor performance by users of Low-PS, and yet others have found the opposite.

Furthermore, although the phenomenon of clutter has previously been identified as causing negative effects for a user, the experiment revealed that for participants of High-PS, when clutter was present, positive effects were observed: this represented the condition where their highest search accuracy was achieved. Therefore, not only were participants with High-PS better able to cope with visual clutter, it appeared that they actually benefited from this in terms of performance. However, to achieve a higher performance, they also had to sacrifice more search time. This therefore indicated that given the task required for a user, different manipulations of clutter may be necessary: if time is of the essence, clutter should be minimised; otherwise, it should be visible.

Finally, although subtle differences in performance, behaviour, and experience were identified between Low-PS and High-PS participants in each separate PS test, overall, the same trends emerged. This demonstrated that both tests appeared to measure the same overall cognitive ability of PS, despite minimal correlations having been identified between the two.

9.2.3 Research Question 3

To further enhance understanding of what the most optimal search environment would be for users of differing PS ability, the *type* of clutter was explored for how it affected a participant's performance, behaviour, and experience—and again, for Low-PS and High-PS participants in two separate PS tests. There were three types of clutter examined: ads that were congruent with the task; incongruent with the task; and a mixture of both. The results were presented in Chapter 8, which highlighted that for Low-PS participants, whenever incongruent clutter was present, their most negative performance and experiences occurred. In contrast, for High-PS, incongruent clutter provided their most optimal condition— except for post-task recall, which was always highest after no clutter.

However, whilst the overall trends were similar between Low-PS and High-PS participants, as measured by different PS tests, there were also some differences identified. For example, when congruent clutter was present, Low-PS participants in one test reported positive user experiences, whereas Low-PS participants in the other test reported negative experiences there.

This indicated that whilst the two different PS tests may be measuring the same overall concept of PS, there are also differences in the type of PS being measured between the two.

9.3 Overall Discussion

In the Discussion sections for both Chapter 7 and Chapter 8, explanations were proposed for the various main results found. In particular, Chapter 7 attempted to explain why Low-PS participants were negatively affected by clutter, whereas High-PS were overall positively affected by clutter. Then, Chapter 8 considered explanations for why incongruent clutter was the most negative condition for Low-PS participants, in comparison to congruent clutter being the least optimal for High-PS participants. Furthermore, both chapters explored the juxtaposition in results between user experience and performance, for participants with High-PS. However beyond individual discussions, it is also possible to provide an overall discussion which merges various results found from across different chapters together, to especially focus on how understanding has improved, as a result of all the present research.

9.3.1 Correlations versus ANOVAs

Whilst many variables were investigated in the present research, all experimental results found for measures of search performance and behaviour originated from either an ANOVA or correlation analysis. Chapter 6, Section 6.10, justified the use for both methods of analyses. However, although the results sections of the corresponding chapters (Chapter 7 and Chapter 8) presented these raw results, it is also possible to investigate the extent to which the correlations align with or diverge from the ANOVA findings. This is necessary to comprehend the implications for the interpretation of PS holistically, and determining the appropriate statistical test for subsequent PS analyses.

Throughout these chapters, there were 124 correlations conducted on the continuous data of the whole sample, and 124 ANOVAs on the Extreme Group categorical data of Low-PS and High-PS. It must be acknowledged that a correlation and an ANOVA are ultimately two different statistical methods used to analyze different types of data. While both methods can be used to examine relationships between variables, they differ in their focus and assumptions– mainly, that correlations analyze continous data, and ANOVAs analyze categorical data. Therefore, it is possible for a correlation and an ANOVA of the same dataset to produce different patterns of results, especially when the middle group of data is removed from one set of analyses, but not the other.

Nonetheless, in the present research, the trends observed in the correlations mostly echoed the trends in the ANOVAs: if there was a positive correlation—which indicated that participants with lower PS were obtaining a lower number than participants with higher PS scores—the Extreme Group Analysis also demonstrated that Low-PS participants obtained a lower number

than High-PS participants. For example, for the dependent variable '*Relevant-Saved Documents / Total-Saved Documents %*', in the *Mixed-Ads* condition, the *Finding A's* correlation was positive and significant, at 0.43. Furthermore, the means of the ANOVA on the Extreme Groups revealed that Low-PS participants were achieving 48% accuracy, which was significantly lower than High-PS participants who achieved 86% accuracy. This would imply that overall, removing the 'average' PS participants did not interfere with the overall results found.

There were however 19/124 instances where the results in the ANOVA did not correspond with the results in the correlation. However, these differences were minimal. For example, for the variable '*Total documents saved*' in the search task for the *No-Ads* condition, the correlation was -0.01, but the Low-PS participants saved a smaller number of 5.58 documents compared to High-PS saving 5.8 documents. With a negative correlation, it would have been expected that Low-PS participants obtained a mean number that was larger than High-PS participants. However, as can be seen from the figures, the correlation was almost non-existent, and thus caution should be exercised when interpreting non-significant and negligible correlations, as over-interpreting such correlations could lead to erroneous conclusions.

Although opposing patterns may not in fact be truly opposing given the small differences observed, there is another possible explanation for opposing trends. In particular, when the middle group of data is removed—which is the largest group of data— the findings may be distorted if the distribution of the dataset does not follow a linear pattern, but rather a curve or some other shape. Consequently, future research should also look at reporting the effects found for users of 'average' PS, to identify whether they obtain figures that are in-between Low-PS and High-PS, or whether a different pattern of results occurs.

Furthermore, given that ANOVAs and correlations can exhibit different trends due to the removal of the middle group of data, the removal of this data can also result in only one test reporting a significant result, even if the trends are similar between both tests. For example, there were many times when the ANOVA reported a significant difference between Low-PS and High-PS, but the correlation of the whole sample was not significant. For instance, in the '*Time session overall*' for the *All-Ads* conditions, the *Finding A's* correlation was small, at only 0.18. Yet, Low-PS completed their search in significantly less time (361 seconds) compared to High-PS (458 seconds). Therefore, removing the noise of the 'average' participants revealed more conclusive differences. Similarly, there were also times where the correlation was significant, but the ANOVA comparing Low-PS against High-PS was not significant. This ultimately demonstrates that in future PS analysis, both statistical tests can provide useful insight.

9.3.2 Finding A's versus Number Comparison

Regardless of whether an ANOVA or correlation was undertaken, results were provided for participants in two separate PS tests: *Finding A's* and *Number Comparison*. Consequently, in addition to the main research questions investigated—for how clutter, and the type of clutter,

would impact participants with different levels of PS—both chapters explored a sub-research question. In Chapter 7, this was to understand: "*Are there different effects on search outcomes, based on different PS tests?*". Then, Chapter 8 investigated: "*Do different types of clutter impact users with different types of PS, as measured by different tests?*"

In the overall answers to these research questions outlined in Section 9.2, it can be seen that for both chapters, the same main patterns in search performance, behaviour, and experience were observed, regardless of whether a participant was 'Low' in one test or the other. This reaffirmed the overall validity of the two different PS tests, that they both did appear to be measuring the same general concept of ability.

There were however also some instances where the results between the two PS tests were opposing. For example, whilst participants with Low-FA reported positive experiences during *Congruent-Ads*, Low-NC reported negative experiences there; and although High-FA participants believed the system was more appealing in the *presence* of clutter, High-NC participants believed the system was more appealing in the *absence* of clutter. This ultimately indicates that whilst the two different PS tests may be measuring the same overall concept of PS, there are also differences in the type of PS being measured between the two. This helps to understand why being 'Low' in one test, does not correspond with being 'Low' in the other test, and accordingly, why 'High' in *Finding A's*, does not mean the participant will also be 'High' in *Number Comparison*.

As the different PS tests have slight variations between them, it could be argued that they might be measuring entirely different constructs. However, cognitive ability tests are designed to measure various mental abilities, such as memory, reasoning, attention, and problem-solving. Just as is the case for any cognitive ability test, even if multiple tests exist that are designed to measure the same cognitive ability, different tests are inevitably different, which ultimately affects the overall outcomes. One explanation for these differences concerns the test design. In the context of *Finding A's* and *Number Comparison*, the former test involved words, whereas the latter involved only numbers. Therefore, although they both involved identifying a target within a specific time period, different participants may respond to different visual prompts differently. This is especially important when considering how an interaction may exist between the test design and participant characteristics. For example, participants from different cultural backgrounds may respond differently to the PS test which involved words that were different to their native language, in comparison to the number test which involved more familiar, universal stimuli. Consequently, although a few differences were identified in some search outcomes between participants in *Finding A's* and *Number Comparison*, these do not reduce the overall validity of the tests; especially when the same overall patterns were found.

However, the few differences in search outcomes that were reported between the different PS tests cannot be ignored. Instead, they should be seen as furthering understanding of a

complex concept. Indeed, there has always been variability with how PS has been measured and conceptualised in previous literature, and this was highlighted in the earlier Systematic Review Chapter (Chapter 5, Section 5.5.2.4). Similarly, other researchers such as Hoermann & Damos [107] have concluded that different PS tests assess different attributes (See Chapter 5, Section 5.5.3.2). Consequently, although multiple PS tests may measure an overall perceptual ability, there may also be different factors that contribute to an overall PS measurement, which are indicated through different versions of tests. This would correspond with the very original guidelines of PS measurement from the 1970s, where in order to fully deduce a cognitive factor, at least two tests should be administered [83] (See Chapter 4, Section 4.3).

Bringing everything together, the results of the present research highlight that both PS tests should continue to be administered in future experiments, and further research should aim to further understand how different attributes interact with one another to form the overall construct of PS.

9.3.3 Other Types of Perceptual Speed

It is important to note that whilst two different PS tests were used in the present research, these both concerned measuring *visual* perceptual speed. There are however also other PS tests available. Specifically, whilst *Finding A's* and *Number Comparison* were based on the original test set from Ekstrom [82], Ekstrom also created *Identical Pictures*, which corresponded with a more spatial PS test (this was described earlier in Chapter 6, Section 6.3).

While there can be some overlap between spatial and visual concepts, they are fundamentally different. 'Spatial' is about the physical or abstract characteristics of space and the relationships between objects in that space, while 'visual' is about the sense of sight, visual perception, and the representation of information through visual means. Consequently, by only administering *Finding A's* and *Number Comparison*, only visual PS has been explored in the present research. This implementation was motivated from previous research finding insignificant results with the Identical Pictures test. Furthermore, the main research question sought to explore the relationship between Perceptual Speed and *visual clutter* in the form of advertisements during an IIR task. Consequently, it made most sense to administer only the PS tests which specifically involved visual perception, and not spatial.

However, in many other areas of research literature which have investigated cognitive factors during IIR, these have considered both a verbal processing channel in addition to visual processing. For example, Gwizdka [99] explored both the notion of cognitive load, as well as Multiple Resource Theory, in their experimentation with IIR tasks using the following explanation: "*When demands of one task are high, the resources committed to that task become unavailable to a second task if it requires the same type of mental resources (e.g., visual vs. auditory) and at the same stage of processing (e.g., cognitive vs. response-related). The mental limitation in perception and cognition are a result of limited capacity of working memory.*

Working memory is conceptualized as containing three subsystems, one responsible for verbal information processing (phonological loop), one for visual information processing (visuo-spatial sketchpad) and one for controlling and coordinating the processing machinery (central executive)(Baddley Hitch, 1980)". This corresponds with other research that has demonstrated that a visual task and verbal task can both be performed simultaneously [14], whereas it's harder to perform 2 visual tasks, or 2 verbal tasks simultaneously because they occupy and overload the same cognitive space [62]. Therefore, whilst the present research has focused on the concept of *visual* PS, it is possible that an interaction may exist with verbal processing, and this should be explored in future research, especially in the context of advertising, where visual *and* verbal cues are frequently used to convey messages and persuade audiences.

9.3.4 Combining Variables Together

Regardless of analysis method undertaken, or specific PS test investigated, another area of knowledge that has improved, as a result of the overall research, concerns a greater understanding of how multiple variables relate with one another. Referring back to the Systematic Review presented in Chapter 5, Section 5.5.7 highlighted that many previous studies on PS only focused on one type of variable, such as search time *or* search performance (E.g. [43, 46]). Similarly, other studies that focused on user experience neglected search performance, or vice versa (E.g. [18, 64]. This ultimately made it difficult to gain a comprehensive understanding of how PS truly affected a user, and questions such as whether a lower task accuracy corresponded with subjective awareness of this difference, or whether a longer time spent searching ultimately led to an increase, or decrease in overall search performance, remained unknown. Whereas now, multiple dependent variables were investigated for the same participants who all undertook the same IIR search tasks. Thus, the questions that originally remained unknown have begun to be answered. For example, in terms of time, participants with Low-PS and High-PS use this differently: when Low-PS participants spent their least amount of time searching, they also achieved their highest search accuracy, and thus they were not only being more efficient, but also more effective. In other words, more search time for participants of Low-PS indicated a negative outcome. In contrast, when High-PS participants completed their search in the least amount of time, this was not effective, as their search accuracy reduced. Thus, when High-PS participants spend longer completing a search, more time appears to be a good thing, in terms of search outcome. These findings therefore help to gain a clearer overall picture of how different variables interact with each other, in specific relation to PS, which overall helps improve understanding of the concept as a whole.

9.3.5 Final Recommendations

Integrating the previous sections together, various recommendations have been proposed for the future use of PS testing and implementation. As a summary, these are: that multiple forms

of analysis can be beneficial, to gain a more comprehensive understanding of the concept of PS; both *Finding A's* and *Number Comparison* should be used to measure PS; further research should use these two PS tests to understand how different attributes contribute to the overall construct of PS; and multiple dependent variables should always be investigated simultaneously, to understand any interactions that may occur between performance, behaviour, and experience, during IIR.

9.4 Overall Contributions

Combining everything learned throughout this thesis together, multiple contributions have arisen. These mainly relate to the original and novel findings generated by the answers to the research questions outlined in Section 9.2. As a summary, these are:

- That many problems have been identified with how PS has previously been measured, which questions the reliability of what was known about the overall concept;
- All claims regarding PS in the field of Computer Science have been quantified, which allows researchers to better understand previous findings in a succinct manner;
- Different explanations for contrasting results in previous literature have been explored, leading to new insights on why people with Low-PS may perform well in one search task, but less well in another;
- Whilst previously PS has been measured using paper/pen testing, a refined digital version has been created for two different PS tests, allowing a faster and more precise, standardised way to quantify and assess a user's PS ability.
- Through empirical investigation, it has been identified that visual clutter—as operationalised through the presence or absence of advertisements—significantly impacts users with different PS ability when completing an IIR task: users with Low-PS are negatively affected by clutter, whereas High-PS are positively affected by clutter.
- It has further been identified that the type of clutter interacts with individual differences. Specifically, users with Low-PS are most negatively affected when the clutter is incongruent with the task, whereas users with High-PS achieve their best search performance amongst incongruent clutter, in comparison to congruent clutter.

Ultimately, all of these contributions have shed light on how individuals process information differently, and this has many implications which are discussed below.

9.5 Research Implications

As a result of the research contributions from the answers to the main research questions, in addition to the overall discussion, multiple implications from this research have also been identified. These implications can be broadly divided into implications in IIR, and implications in Computer Science, and other domains, more generally.

9.5.1 Implications for IIR

Firstly considering IIR independently, through clearly categorising all previous research in the area of PS during IIR together in the Systematic Review, this has provided a framework where future researchers will now be able to position their findings precisely in the literature. This will enable similarities and differences to be identified, to further understand what, and why, certain environments are more, or less optimal for users with differing PS levels during IIR.

Secondly, although only the cognitive ability of PS has been investigated, the methods utilised to better understand the concept—in particular, how problems in previous measurement were identified—can be replicated for other types of cognitive abilities using the methodological diagram provided in Section 1.3 (Figure 1.1). There are many cognitive abilities—ranging from various types of memory, to verbal reasoning—and their effect on a user's IIR ability could also provide valuable insight, to create the most optimal systems that are designed to focus on the user's unique strengths and limitations.

Thirdly, the interaction between PS and visual clutter—as demonstrated in the experiment—has reaffirmed the importance of considering individual differences in cognitive abilities in the system design during IIR. Specifically, for users of Low-PS, it is important to simplify what is visible on a search system by reducing visual clutter in the form of ads. This could be achieved through tools such as ad-blockers or read-only views. By minimizing visual distractions and providing a clear and organized visual environment, this should help make it easier for users with Low-PS to attend to and process important information, and thus user experience and performance would improve. Alternatively, if clutter must be present, then ranking algorithms should prioritise relevant retrieval algorithms, as users with Low-PS performed better amidst congruent clutter, in comparison to incongruent clutter. In contrast, users with High-PS benefit from the presence of clutter during IIR. More specifically, they achieved their most efficient task performance during the IIR tasks when the clutter was incongruent with the task. Consequently, system design for High-PS users should involve *irrelevant* advertising, and thus retrieval algorithms should be reversed,

9.5.2 Implications for other domains

In addition to implications for IIR, other domains can also benefit from the present research.

Firstly, having identified problems with previously used measurements of PS, and having created new digital versions—this has provided a tool which can be used by anyone in the field of PS: whether that be researchers in Computer Science; or other fields such as Psychology. Additionally, as PS is also used in different areas—such as civilian pilot selection [107]—the current tests may also provide a more reliable and valid PS measurement in many different domains.

Secondly, to ensure the search system is the most efficient for a user's individual abilities when completing an IIR task, this inspires the design of digital interfaces in general. Therefore, if using an ad-blocker may benefit users with Low-PS, then advertisers and marketers may wish to further investigate how they can reach target audiences more effectively. Alternatively, they may also want to reconsider the default algorithm being set to retrieve relevant advertising to a user [192], and instead, ensure the type of advert retrieved is specific to a user's PS ability.

Thirdly, as the present research demonstrated the effects of clutter and perceptual speed during IIR, it is possible that the results may also extend onto other human-computer interaction tasks. For example, as Low-PS also highlighted that a system with no clutter is boring, if they were completing a task that was not time-sensitive, or did not require a measure of performance—such as browsing social media—then clutter could potentially be visible.

Moving beyond academia, there are many industries which incorporate PS testing, such as civilian pilot section as described above [107]. However to date, the aim of these testing programmes has been to employ higher PS scorers for jobs that require complex perceptual abilities [107]. Yet, the present research has demonstrated that users with Low-PS have the ability to achieve performance ratings that are higher than users with High-PS, given the right search context. This therefore highlights that users with Low-PS should not necessarily be dismissed from a certain job role, but rather, that the search environment used in the job may be able to adapt it's visual presentation to the user's individual needs. Consequently, understanding the relationship between PS and other types of visual clutter, in different domains, remains an open research question to explore in industrial contexts.

Ultimately, integrating knowledge from all of these different implications will lead to more comprehensive and insightful research that can inform the design of interactive systems that are user-centered, context-sensitive, and adaptable.

9.6 Limitations

Although the research presented in this thesis provided much needed insight into the concept of PS, it is not without limitations. In addition to the limitations specified in the discussion sections of previous chapters (including Section 7.13 and Section 8.7), other factors need to be noted.

Firstly, a participant's PS was only measured at one point in time: potential changes in PS over time were not examined. However, it is possible that PS may change in response to various factors. For example, research has found that cognitive abilities are not stable, and instead can be affected by environmental factors such as tiredness levels or depressive symptoms [32, 153]. This would imply that PS is a changeable ability, requiring regular re-testing. Accordingly, it would be preferable if PS levels could be inferred dynamically from people's interactions—rather than at one point in time— which would enable systems to adapt according to the user's current PS levels.

In addition to PS only being measured once, a participant's searching activity was also only measured at one point in time, as all four search conditions were completed in the same sitting. The results may therefore not accurately capture the long-term effects of clutter on PS. This is a limitation because it is possible that the effects of clutter may change over time, and it would be valuable to explore this in future studies. For example, perhaps as users become more familiar with the visual setup of the interface they are searching, the effects of clutter may be ameliorated.

A further limitation of this research relates to how the results can be used in the future. Specifically, although discussion of the experimental results has focused on driving adaptive systems, an important factor to consider is that not all queries have a high potential for personalisation. For example, a query of 'New york times' is quite generic, and most users would click on the same main website. Thus, whilst the present research has highlighted the need for clutter to be congruent or incongruent—depending on whether a participant has Low-PS or High-PS, respectively—before this can be implemented, systems will need to learn *when* it can personalise the interface.

One final limitation to consider is that PS is just one type of cognitive ability. As there are many other cognitive abilities and individual differences, these must also be considered when designing systems that can benefit a user. For example, it is unknown whether providing incongruent clutter to users with High-PS would be positive or negative, depending on whether the specific user also had low or high levels of other cognitive abilities, such as working memory or specific personality traits. Future research should therefore explore the potentially complex interactions of these additional factors to gain a more comprehensive understanding of PS.

9.7 Future Work

Whilst various limitations of the research have been noted—both in the individual discussion chapters and overall—in addition to considering these when interpreting the results of this research, at the same time, the limitations also provide various avenues to explore for future work. For example, whilst the present experimentation only provided knowledge on the short-term effects of different types of clutter on participants with differing PS levels, future research

could implement experimental methodology that is able to quantify the long-term effects, such as the implementation by researchers in Google (e.g. [108]).

Furthermore, as a summary of the suggestions for future work highlighted in Chapter 7 and Chapter 8, these include: repeating the experiment with a larger sample, in different contexts, with different types of ads and location of clutter, and exploring the long-term effects of clutter. Then, to summarise the suggestion for future work from the overall discussion in Chapter 9 (Section 9.3.3), this involved an exploration of possible other types of PS. Specifically, understanding whether there are differences in PS for verbal and spatial processing, in addition to just visual processing, is vital for future research, as it allows for a deeper exploration of how these differences impact various cognitive and behavioural processes. With greater understanding of these factors, the results will be able to be applied to many different contexts.

There are also many other factors that are under-explored that could potentially provide greater understanding of the relationship between PS and visual clutter during IIR. For example, previous research has highlighted the difference in online browsing behaviour based upon gender: "*Female consumers generally focus more strongly on the text, while male consumers concentrate more on photos and images*" [110]. Yet, little is known about the interaction of PS and gender, such as whether one gender is more or less likely to have Low-PS. Additionally, little is known about the interaction of gender and different types of visual clutter. Yet given that gender impacts eye fixation, it makes sense that it would also interact with PS and visual clutter. This translates into future research that could explore whether there is a three-way interaction between PS, visual clutter, *and* gender, during IIR. For example, perhaps only one type of gender with Low-PS may be negatively affected by clutter. Yet without further research and analysis, these questions remain unknown.

In addition to gender, another variable which would be worthwhile exploring for any possible further interactions concerns the role of cultural and language differences between users. Specifically, the present research was conducted using an English search system with participants who confirmed they spoke English. Whilst participants were not required to be native English speakers, every participant recorded that they were native. It is therefore possible that differences in perception may occur between users who are not native, but instead, bilingual, or who have limited proficiency. For example, if users were instead bilingual, specifically with languages that comprise of different structures—such as a bilingual Arabic speaker who can read both right-left, and left-right—then the role of clutter and PS may create different results; perhaps even in the initial PS measurement, there may be an advantage for higher PS levels with greater adaptability to different visual presentations. Alternatively, whether the same effects found in the present research would also be found across different languages, remains unknown. With the long-term aim of creating dynamic systems that can accommodate each unique individual, these variables must at least be investigated for whether any differences in

search performance or experience do occur across different languages. This would allow the results to be more representative of the entire population of online searchers, and not just limited to native English speakers.

Furthermore, as well as exploring potential interactions between PS, visual clutter, and other variables, future research would benefit from further understanding the *cause* of the current results found. Whilst it has been identified that certain types of clutter benefit users with High-PS to complete a positive and effective IIR task—and other types of clutter negatively affect users with Low-PS—the current experiment was not designed to understand *why* these effects occurred. Although different explanations have been proposed in the Discussion sections of earlier chapters, without further research, these explanations can only remain as speculation, such as whether the use of eye-tracking could confirm whether there are differences between direct gaze or peripheral vision for users of different PS ability. With greater understanding, then the results could have implications beyond creating adaptive interfaces, and move towards developing strategies that could further help a user. For example, if it was found that the effects of clutter were due to direct fixation on clutter by users of Low-PS, then perhaps techniques could be taught to re-train eye fixation. Consequently, understanding why clutter interacts with PS during IIR remains an open, yet important, area to explore.

Finally, regardless of what causes different types of clutter to be responded to differently, depending on a user's PS ability, the present research ultimately identified that clutter and PS do significantly affect search performance, behaviour, and user experience when completing an IIR task. This ultimately inspires the development of interactive systems that can adapt to each user's unique ability, to ensure both an efficient, effective, and enjoyable search experience for the user. Future research thus needs to identify how such an adaptive system can be created. There have been some attempts to use eye gaze data from users undertaking information visualisation search tasks, and various Machine Learning models have been used to predict whether the user had Low or High levels based on one PS test. Having achieved the best accuracy scores of 57.1% in [210], and 60.6% in [66], this suggests that predicting PS levels from eye-gaze is difficult. Furthermore, given the cost and obtrusiveness of eye-tracking, this does not represent an efficient solution. Consequently, detecting a user's PS level from other forms of interaction may be preferable.

As other research has found that cognitive abilities are not stable, and instead can be affected by environmental factors such as tiredness levels or depressive symptoms [32, 153], this would imply that PS is a changeable ability, requiring regular re-testing. Accordingly, it would be preferable if PS could be inferred automatically, from people's interactions, which would enable systems to adapt to the user's current PS level. Given that the present research identified that participants with Low-PS and High-PS exhibited different search behaviours, this could represent one way of developing adaptive systems: several machine learning models

could be trained on behavioural features extracted from search tasks to automatically classify a user's PS. A preliminary experiment was conducted using this motivation, and the results were encouraging: given a user's interactions from one query, a Decision Tree was able to predict a user's as Low or High with 86% accuracy [89]. Therefore, future research should focus on improving this accuracy, so that interfaces can adapt to the right type of clutter, based upon a user's current PS ability and task.

9.8 Closing Remarks

Overall, this thesis has advanced understanding of the complexities of Perceptual Speed, and been the first to identify the interaction between Perceptual Speed and visual clutter, in the form of advertisements, during Interactive Information Retrieval. Whilst previously having a 'Low' Perceptual Speed may have been considered a disadvantage, this research has highlighted that with the use of systems that are personalised to the unique ability of the user, a user with Low-PS can achieve higher search performance than a user with High-PS; specifically, through the manipulation of visible clutter that surrounds the interface. This has provided insights into the design of interactive systems that are both effective and enjoyable for users, based on their individual differences in Perceptual Speed. It is hoped that the findings of this research will ultimately lead to the development of adaptive interfaces, as well as inform future research, theory, and practice in the field of Interactive Information Retrieval.

Part V

References

Bibliography

[1] Ackerman, P. L., and Beier, M. E. Further Explorations of Perceptual Speed Abilities in the Context of Assessment Methods, Cognitive Abilities, and Individual Differences During Skill Acquisition. *Journal of Experimental Psychology: Applied 13*, 4 (2007), 249–272.

[2] Adamo, S. H., Cain, M. S., and Mitroff, S. R. Targets need their own personal space: Effects of clutter on multiple-target search accuracy. *Perception 44*, 10 (2015), 1203–1214.

[3] AdBlock. How to make the most of adblock, Jun 2017. `https://blog.getadblock.com/how-to-make-the-most-of-adblock-304f0e9dc1bc`.

[4] Al-Maskari, A., and Sanderson, M. The effect of user characteristics on search effectiveness in information retrieval. *Information Processing & Management 47*, 5 (sep 2011), 719–729.

[5] Aldenkamp, A., Arends, J., Bootsma, H., Diepman, L., Hulsman, J., Lambrechts, D., Leenen, L., Majoie, M., Schellekens, A., and De Vocht, J. Randomized double-blind parallel-group study comparing cognitive effects of a low-dose lamotrigine with valproate and placebo in healthy volunteers. *Epilepsia 43*, 1 (2002), 19–26.

[6] Alexander, A. L., Kaber, D. B., Kim, S.-H., Stelzer, E. M., Kaufmann, K., and Prinzel III, L. J. Measurement and modeling of display clutter in advanced flight deck technologies. *The International Journal of Aviation Psychology 22*, 4 (2012), 299–318.

[7] Alexander, C., Paul, M., Michael, M., et al. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals. *Journal of the International Neuropsychological Society 9*, 3 (2003), 419–428.

[8] Allan, J., Harman, D., Kanoulas, E., Li, D., Van Gysel, C., and Voorhees, E. M. Trec 2017 common core track overview. In *TREC* (2017).

[9] Allen, B. Cognitive differences in end user searching of a CD-ROM index. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '92* (New York, New York, USA, 1992), ACM Press, pp. 298–309.

[10] Allen, B. Cognitive abilities and information system usability. *Information Processing & Management 30*, 2 (mar 1994), 177–191.

[11] Allen, B. Perceptual speed, learning and information retrieval performance. In *SIGIR'94* (1994), Springer, pp. 71–80.

[12] Allen, B. Information space representation in interactive systems: relationship to spatial abilities. In *Proceedings of the ACM International Conference on Digital Libraries* (1998), pp. 1–10.

[13] Allen, B. Individual differences and the conundrums of user-centered design: Two experiments. *Journal of the American Society for Information Science and Technology 51*, 6 (2000), 508–520.

[14] Allport, D. A., Antonis, B., and Reynolds, P. On the division of attention: A disproof of the single channel hypothesis. *Quarterly journal of experimental psychology 24*, 2 (1972), 225–235.

[15] Arapakis, I., and Leiva, L. A. Learning efficient representations of mouse movements to predict user attention. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 1309–1318.

[16] Arguello, J., and Capra, R. The Effect of Aggregated Search Coherence on Search Behavior. *Proc. 21st ACM CIKM* (2012), 1293–1302.

[17] Arguello, J., and Capra, R. The effects of aggregated search coherence on search behavior. *ACM Trans. Inf. Syst. 35*, 1 (sep 2016).

[18] Arguello, J., and Choi, B. The Effects of Working Memory, Perceptual Speed, and Inhibition in Aggregated Search. *ACM Transactions on Information Systems 37*, 3 (jul 2019), 1–34.

[19] Arora, M., Kanjilal, U., and Varshney, D. Evaluation of information retrieval: precision and recall. *International Journal of Indian Culture and Business Management 12* (01 2016), 224.

[20] Aslam, W., Butt, W. H., and Anwar, M. W. A systematic review on social network analysis: Tools, algorithms and frameworks. In *Proceedings of the 2018 International Conference on Computing and Big Data* (2018), pp. 92–97.

[21] Azzopardi, L., Kelly, D., and Brennan, K. How query cost affects search behavior. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (2013), pp. 23–32.

[22] Azzopardi, L., and Maxwell, D. Tango with django: a beginners guide to web development with django 1.5. 4, 2013.

[23] Azzopardi, L., and Zuccon, G. Economics models of interaction: a tutorial on modeling interaction using economics.

[24] Baeza-Yates, R., Ribeiro-Neto, B., et al. *Modern information retrieval*, vol. 463. ACM press New York, 1999.

[25] Bailey, E., and Kelly, D. Is amount of effort a better predictor of search success than use of specific search tactics? *Proceedings of the American Society for Information Science and Technology 48*, 1 (2011), 1–10.

[26] Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. The english lexicon project, 2007.

[27] Bang, H., Kim, J., and Choi, D. Exploring the effects of ad-task relevance and ad salience on ad avoidance: The moderating role of internet use motivation. *Computers in Human Behavior 89*, April (2018), 70–78.

[28] Barchard, K. A., and Williams, J. Practical advice for conducting ethical online experiments and questionnaires for united states psychologists. *Behavior research methods 40*, 4 (2008), 1111–1128.

[29] Barral, O., Lallé, S., Guz, G., Iranpour, A., and Conati, C. Eye-tracking to predict user cognitive abilities and performance for user-adaptive narrative visualizations. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (2020), pp. 163–173.

[30] Battelle, J. *The search: How Google and its rivals rewrote the rules of business and transformed our culture*. Hachette UK, 2011.

[31] Bauerly, M., and Liu, Y. Computational modeling and experimental investigation of effects of compositional elements on interface and design aesthetics. *International journal of human-computer studies 64*, 8 (2006), 670–682.

[32] Beaujean, A. A., Parker, S., and Qiu, X. The relationship between cognitive ability and depression: a longitudinal data analysis. *Social psychiatry and psychiatric epidemiology 48*, 12 (2013), 1983–1992.

[33] Beck, M. R., Lohrenz, M. C., and Trafton, J. G. Measuring search efficiency in complex visual search tasks: Global and local clutter. *Journal of experimental psychology: applied 16*, 3 (2010), 238.

[34] Benes, R. People believe ads are becoming more intrusive. *Diakses dari https://www. emarketer. com/content/people-believe-ads- are-becoming-moreintrusive (pada tanggal 26 Agustus 2018)* (2018).

[35] Bennett, C. R., Bex, P. J., Bauer, C. M., and Merabet, L. B. The assessment of visual function and functional vision. In *Seminars in pediatric neurology* (2019), vol. 31, Elsevier, pp. 30–40.

[36] Boeuf, D., Carrillat, F., and D'Astous, A. Interference effects in competitive sponsorship clutter. *Psychology & Marketing 35*, 12 (2018), 968–979.

[37] Borkowski, K., Taha, A. Y., Pedersen, T. L., De Jager, P. L., Bennett, D. A., Arnold, M., Kaddurah-Daouk, R., and Newman, J. W. Serum metabolomic biomarkers of perceptual speed in cognitively normal and mildly impaired subjects with fasting state stratification. *Scientific reports 11*, 1 (2021), 1–12.

[38] Borlund, P. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research 8*, 3 (2003), 8–3.

[39] Borlund, P., and Ingwersen, P. The development of a method for the evaluation of interactive information retrieval systems. *Journal of documentation* (1997).

[40] Bota, H., Zhou, K., and Jose, J. M. Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of the 2016 acm on conference on human information interaction and retrieval* (2016), pp. 131–140.

[41] Bouhadjar, Y., Diesmann, M., Waser, R., Wouters, D. J., and Tetzlaff, T. Constraints on sequence processing speed in biological neuronal networks. In *Proceedings of the International Conference on Neuromorphic Systems* (2019), pp. 1–9.

[42] Bourguet, M.-L. Metrics-based evaluation of graphical user interface aesthetics: The segmentation problem. In *Proceedings of the 2018 ACM Companion International Conference on Interactive Surfaces and Spaces* (2018), pp. 31–38.

[43] Brauner, P., Calero Valdez, A., Philipsen, R., and Ziefle, M. Defective Still Deflective – How Correctness of Decision Support Systems Influences User's Performance in Production Environments. In *HCI in Business, Government, and Organizations: Information Systems*. Springer International Publishing, 2016, pp. 16–27.

[44] Brauner, P., Philipsen, R., Calero Valdez, A., and Ziefle, M. What happens when decision support systems fail?—the importance of usability on performance in erroneous systems. *Behaviour & Information Technology 38*, 12 (2019), 1225–1242.

[45] Bravo, M. J., and Farid, H. The depth of distractor processing in search with clutter. *Perception 36*, 6 (2007), 821–829.

[46] Brennan, K., Kelly, D., and Arguello, J. The effect of cognitive abilities on information search for tasks of varying levels of complexity. In *Proceedings of the 5th Information Interaction in Context Symposium* (New York, NY, USA, aug 2014), ACM, pp. 165–174.

[47] Breuer, T., Ferro, N., Fuhr, N., Maistro, M., Sakai, T., Schaer, P., and Soboroff, I. How to measure the reproducibility of system-oriented ir experiments. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 349–358.

[48] Breznitz, Z., and Meyler, A. Speed of lower-level auditory and visual processing as a basic factor in dyslexia: Electrophysiological evidence. *Brain and Language 85*, 2 (2003), 166–184.

[49] Brysbaert, M. How many participants do we have to include in properly powered experiments? a tutorial of power analysis with reference tables. *Journal of cognition* (2019).

[50] Bullers, K., Howard, A. M., Hanson, A., Kearns, W. D., Orriola, J. J., Polo, R. L., and Sakmar, K. A. It takes longer than you think: librarian time spent on systematic review tasks. *Journal of the Medical Library Association: JMLA 106*, 2 (2018), 198.

[51] Burke, M., Hornof, A., Nilsen, E., and Gorman, N. High-Cost Banner Blindness: Ads Increase Perceived Workload, Hinder Visual Search, and Are Forgotten. *ACM Trans. on HCI 12*, 4 (2005), 423–445.

[52] Buscher, G., Dumais, S., and Cutrell, E. The good, the bad, and the random: An eye-tracking study of ad quality in web search. In *Proc. 33rd ACM SIGIR* (2010), pp. 42–49.

[53] Capra, R., Arguello, J., Crescenzi, A., and Vardell, E. Differences in the use of search assistance for tasks of varying complexity. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (2015), pp. 23–32.

[54] Capra, R., Arguello, J., O'Brien, H., Li, Y., and Choi, B. The effects of manipulating task determinability on search behaviors and outcomes. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (2018), pp. 445–454.

[55] Capra, R., Arguello, J., and Zhang, Y. The effects of search task determinability on search behavior. In *European Conference on Information Retrieval* (2017), Springer, pp. 108–121.

[56] Carenini, G., Conati, C., Hoque, E., Steichen, B., Toker, D., and Enns, J. Highlighting interventions and user differences: Informing adaptive information visualization support. In *Conference on Human Factors in Computing Systems - Proceedings* (New York, New York, USA, 2014), ACM Press, pp. 1835–1844.

[57] Carr, N. Is Google Making Us Stupid? *Yearbook of the National Society for the Study of Education 107*, 2 (oct 2008), 89–94.

[58] Chang, D., Dooley, L., and Tuovinen, J. Gestalt theory in visual screen design: A new look at an old subject. In *Proc. 7th ACM CRPIT* (2002), pp. 5–12.

[59] Cheung, L. Y. T., and Cheung, S.-H. Chinese-character crowding—i. effects of structural similarity. *Journal of Vision 17*, 11 (2017), 14–14.

[60] Choi, B., Arguello, J., Capra, R., and Ward, A. R. Orgbox: A knowledge representation tool to support complex search tasks. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (2021), pp. 219–228.

[61] Christidi, F., Kararizou, E., Triantafyllou, N., Anagnostouli, M., and Zalonis, I. Derived trail making test indices: demographics and cognitive background variables across the adult life span. *Aging, Neuropsychology, and Cognition 22*, 6 (2015), 667–678.

[62] Cocchini, G., Logie, R. H., Della Sala, S., MacPherson, S. E., and Baddeley, A. D. Concurrent performance of two memory tasks: Evidence for domain-specific working memory systems. *Memory & Cognition 30*, 7 (2002), 1086–1095.

[63] Cohen, J. Eta-squared and partial eta-squared in fixed factor anova designs. *Educational and psychological measurement 33*, 1 (1973), 107–112.

[64] Conati, C., Carenini, G., Hoque, E., Steichen, B., and Toker, D. Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. *Computer Graphics Forum 33*, 3 (jun 2014), 371–380.

[65] Conati, C., Carenini, G., Toker, D., and Lallé, S. Towards user-adaptive information visualization. *Proceedings of the National Conference on Artificial Intelligence 6*, 2006 (2015), 4100–4106.

[66] Conati, C., Lallé, S., Rahman, A., and Toker, D. Further results on predicting cognitive abilities for adaptive visualizations. In *IJCAI International Joint Conference on Artificial Intelligence* (2017), pp. 1568–1574.

[67] Conati, C., Lallé, S., Rahman, M. A., and Toker, D. Comparing and Combining Interaction Data and Eye-tracking Data for the Real-time Prediction of User Cognitive Abilities in Visualization Tasks. *ACM Transactions on Interactive Intelligent Systems 10*, 2 (2020).

[68] Conati, C., and Maclaren, H. Exploring the role of individual differences in information visualization. In *Proceedings of the working conference on Advanced visual interfaces* (2008), pp. 199–206.

[69] Cool, C., and Belkin, N. J. Interactive information retrieval: history and background., 2011.

[70] Crabb, M., and Hanson, V. L. Age, technology usage, and cognitive characteristics in relation to perceived disorientation and reported website ease of use. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility - ASSETS '14* (New York, New York, USA, 2014), ACM Press, pp. 193–200.

[71] Crabb, M., and Hanson, V. L. An analysis of age, technology usage, and cognitive characteristics within information retrieval tasks. *ACM Transactions on Accessible Computing 8*, 3 (2016).

[72] Cramer-Flood, E. Global digital ad spending update q2 2020, Jul 2020.

[73] Crescenzi, A., Kelly, D., and Azzopardi, L. Impacts of time constraints and system delays on user experience. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (2016), pp. 141–150.

[74] Crisp, V. Exploring the relationship between validity and comparability in assessment. *London Review of Education* (2017).

[75] Damos, D. L., Bittner, A. C., and Niemczyk, M. Predicting flight training performance from the tabular speed test. *Aviation Psychology and Applied Human Factors* (2022).

[76] Daume, J., Graetz, S., Gruber, T., Engel, A. K., and Friese, U. Cognitive control during audiovisual working memory engages frontotemporal theta-band interactions. *Scientific reports 7*, 1 (2017), 1–13.

[77] DeCoster, J., Gallucci, M., and Iselin, A.-M. R. Best practices for using median splits, artificial categorization, and their continuous alternatives. *Journal of experimental psychopathology 2*, 2 (2011), 197–209.

[78] Dubiel, M., Halvey, M., and Azzopardi, L. *Exploring the Impact of Conversational Strategies on User Search Experience in Goal-Oriented Tasks in a Voice-Only Domain.* PhD thesis, University of Strathclyde, 2021.

[79] Dumais, S., Buscher, G., and Cutrell, E. Individual differences in gaze patterns for Web search. In *Proc. 3rd IIiX* (2010), pp. 185–194.

[80] Dunteman, G. H., Ho, M.-H. R., and Ho, M.-H. R. *An introduction to generalized linear models*, vol. 145. Sage, 2006.

[81] Edwards, A., Kelly, D., and Azzopardi, L. The impact of query interface design on stress, workload and performance. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29-April 2, 2015. Proceedings 37* (2015), Springer, pp. 691–702.

[82] Ekstrom, R. B. *Kit of factor-referenced cognitive tests.* Educational Testing Service, 1976.

[83] Ekstrom, R. B., Dermen, D., and Harman, H. H. *Manual for kit of factor-referenced cognitive tests*, vol. 102. Educational testing service Princeton, NJ, 1976.

[84] Fincannon, T., Jentsch, F., Sellers, B., and Keebler, J. R. Beyond "Spatial ability": Examining the impact of multiple individual differences in a perception by proxy framework. In *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction* (2012), pp. 127–128.

[85] Finkel, D., Reynolds, C. A., McArdle, J. J., and Pedersen, N. L. Age changes in processing speed as a leading indicator of cognitive aging. *Psychology and aging 22*, 3 (2007), 558.

[86] Fisk, J. E., and Warr, P. Age and working memory: The role of perceptual speed, the central executive, and the phonological loop. *Psychology and Aging 11*, 2 (jun 1996), 316–323.

[87] Foulds, O. Investigating how word clutter and colour impact upon learning. In *Orchestration of Learning Environments in the Digital World.* Springer, 2022, pp. 135–151.

[88] Foulds, O., Azzopardi, L., and Halvey, M. Investigating the influence of ads on user search performance, behaviour, and experience during information seeking. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (2021), pp. 107–117.

[89] Foulds, O., Suglia, A., Azzopardi, L., and Halvey, M. Predicting perceptual speed from search behaviour. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 1989–1992.

[90] Fowler, M. *Patterns of Enterprise Application Architecture: Pattern Enterpr Applica Arch.* Addison-Wesley, 2012.

[91] Frankowska-Takhari, S., MacFarlane, A., Goker, A., and Stumpf, S. Selecting and tailoring of images for visual impact in online journalism. *Information Research 22*, 1 (2017), 1619.

[92] French, J. W., Ekstrom, R. B., and Price, L. A. Manual for kit of reference tests for cognitive factors (revised 1963). Tech. rep., Educational Testing Service Princeton NJ, 1963.

[93] Gagné, N., and Franzen, L. How to run behavioural experiments online: best practice suggestions for cognitive psychology and neuroscience.

[94] Gingerich, M., and Conati, C. Constructing models of user and task characteristics from eye gaze data for user-adaptive information highlighting. In *Proceedings of the National Conference on Artificial Intelligence* (2015), vol. 3, pp. 1728–1734.

[95] Goldhammer, F., Rauch, W. A., Schweizer, K., and Moosbrugger, H. Differential effects of intelligence, perceptual speed and age on growth in attentional speed and accuracy. *Intelligence 38*, 1 (2010), 83–92.

[96] Goldstein, D. G., McAfee, R. P., and Suri, S. The cost of annoying ads. In *Proceedings of the 22nd international conference on World Wide Web* (2013), pp. 459–470.

[97] Güner, H., and İnal, Y. The effect of banner location on banner recognition in a turkish government website: an eye tracking study. In *International Conference on Human-Computer Interaction* (2015), Springer, pp. 65–72.

[98] Gwizdka, J. Assessing cognitive load on web search tasks. *arXiv preprint arXiv:1001.1685* (2010).

[99] Gwizdka, J. Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology 61*, 11 (2010), 2167–2187.

[100] Gwizdka, J. Effects of working memory capacity on users' search effort. In *Proceedings of the International Conference on Multimedia, Interaction, Design and Innovation* (2013), pp. 1–8.

[101] Gwizdka, J., and Cole, M. J. Inferring cognitive states from multimodal measures in information science. In *ICMI 2011 Workshop on Inferring Cognitive and Emotional States from Multimodal Measures (ICMI'2011 MMCogEmS)(Alicante:)* (2011).

[102] Ha, L., and McCann, K. An integrated model of advertising clutter in offline and online media. *International Journal of Advertising 27*, 4 (2008), 569–592.

[103] Haapalainen, E., Kim, S., Forlizzi, J. F., and Dey, A. K. Psycho-physiological measures for assessing cognitive load. In *UbiComp'10 - Proceedings of the 2010 ACM Conference on Ubiquitous Computing* (2010), pp. 301–310.

[104] Harpe, S. E. How to analyze likert and other rating scale data. *Currents in pharmacy teaching and learning 7*, 6 (2015), 836–850.

[105] Harrison, W. J., and Bex, P. J. A unifying model of orientation crowding in peripheral vision. *Current Biology 25*, 24 (2015), 3213–3219.

[106] Henry, S., and McInnes, B. T. Literature based discovery: models, methods, and trends. *Journal of biomedical informatics 74* (2017), 20–32.

[107] Hoermann, H.-J., and Damos, D. L. The use of a perceptual speed test in civilian pilot selection. In *20th International Symposium on Aviation Psychology* (2019), p. 391.

[108] Hohnhold, H., O'Brien, D., and Tang, D. Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), pp. 1849–1858.

[109] Holmberg, N., Sandberg, H., and Holmqvist, K. Advert saliency distracts children's visual attention during task-oriented internet use. *Frontiers in Psychology 5* (2014), 51.

[110] Huang, Y. T. *The female gaze: Content composition and slot position in personalized banner ads, and how they influence visual attention in online shoppers*, vol. 82. Elsevier B.V., 2018.

[111] Huestegge, L., and Radach, R. Visual and memory search in complex environments: determinants of eye movements and search performance. *Ergonomics 55*, 9 (2012), 1009–1027.

[112] Jaeger, J. Digit symbol substitution test: the case for sensitivity over specificity in neuropsychological testing. *Journal of clinical psychopharmacology 38*, 5 (2018), 513.

[113] Jahanian, A., Keshvari, S., and Rosenholtz, R. Web pages: What can you see in a single fixation? *Cognitive Research: Principles and Implications 3*, 1 (2018).

[114] James, W. *The principles of psychology*, vol. 1. Cosimo, Inc., 1890.

[115] Jankowski, J., Hamari, J., and Wątróbski, J. A gradual approach for maximising user conversion without compromising experience with high visual intensity website elements. *Internet Research 29*, 1 (2019), 194–217.

[116] Kaber, D. B., Alexander, A. L., Stelzer, E. M., Kim, S.-H., Kaufmann, K., and Hsiang, S. Perceived clutter in advanced cockpit displays: measurement and modeling with experienced pilots. *Aviation, space, and environmental medicine 79*, 11 (2008), 1007–1018.

[117] Kahneman, D. *Attention and effort*, vol. 1063. Citeseer, 1973.

[118] Kaur, H., Pannu, H. S., and Malhi, A. K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR) 52*, 4 (2019), 1–36.

[119] Kazmi, S. H. A., Hai, L. C., and Abid, M. M. Online Purchase Intentions in E-Commerce. In *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)* (aug 2016), vol. 2, IEEE, pp. 570–573.

[120] Kelly, D. *Methods for evaluating interactive information retrieval systems with users.* Now Publishers Inc, 2009.

[121] Kelly, D., Arguello, J., Edwards, A., and Wu, W. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proc. 5th ACM ICTIR* (2015), pp. 101–110.

[122] Kelly, D., and Azzopardi, L. How many results per page? a study of serp size, search behavior and user experience. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (2015), pp. 183–192.

[123] Kelly, D., and Sugimoto, C. R. A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology 64*, 4 (2013), 745–770.

[124] Keshvari, S., and Rosenholtz, R. Pooling of continuous features provides a unifying account of crowding. *Journal of Vision 16*, 3 (2016), 39–39.

[125] Kim, K. S., and Allen, B. Cognitive and task influences on Web searching behavior. *Journal of the American Society for Information Science and Technology 53*, 2 (2002), 109–119.

[126] Kim, N. Y., and Sundar, S. S. Relevance to the rescue: Can "smart ads" reduce negative response to online ad clutter? *Journalism & Mass Communication Quarterly 87*, 2 (2010), 346–362.

[127] Kirkpatrick, D. The Minnesota Clerical Test. *The British Journal of Psychiatry 111*, 479 (1965), 1009–1010.

[128] Kuisma, J. *Consumer Perception of Online Advertising.* PhD thesis, Aalto University, Markkinoinnin laitos, 2015.

[129] Kumar, S., Baijal, S., Chourey, L., Ramamurthy, A., and Sasikumar, M. Conceptualizing a desktop environment for cognitively challenged people. In *Proc. CUBE Conference* (2012), p. 366–370.

[130] Kurosu, M. *Human-Computer Interaction: Users and Contexts: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015. Proceedings, Part III*, vol. 9171. Springer, 2015.

[131] Laerd, S. Friedman test in spss, 2013. `https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php`.

[132] Laerd, S. Mixed anova using spss, 2013. `https://statistics.laerd.com/spss-tutorials/mixed-anova-using-spss-statistics.php`.

[133] Laerd, S. Kruskal-wallis h test using spss statistics. *Statistical tutorials and software guides* (2015). `https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php`.

[134] Laerd, S. Mann-whitney u test using spss statistics. *Statistical tutorials and software guides* (2015). `https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php`.

[135] Laerd, S. Pearson product-moment correlation, 2020. `https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php`.

[136] Lafleur, C., and Rummel, B. Predicting perceived screen clutter by feature congestion. *Mensch & Computer* (2011).

[137] Lallé, S., and Conati, C. The role of user differences in customization: A case study in personalization for infovis-based content. In *International Conference on Intelligent User Interfaces, Proceedings IUI* (2019), vol. Part F1476, pp. 329–339.

[138] Lallé, S., Conati, C., and Carenini, G. Prediction of individual learning curves across information visualizations. *User Modeling and User-Adapted Interaction 26*, 4 (oct 2016), 307–345.

[139] Lallé, S., Conati, C., and Carenini, G. Impact of Individual Differences on User Experience with a Visualization Interface for Public Engagement. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (New York, NY, USA, jul 2017), ACM, pp. 247–252.

[140] Lallé, S., Toker, D., Conati, C., and Carenini, G. Prediction of users' learning curves for adaptation while using an information visualization. In *International Conference on Intelligent User Interfaces, Proceedings IUI* (2015), vol. 2015-Janua, pp. 357–368.

[141] Landau, A. N., Aziz-Zadeh, L., and Ivry, R. B. The influence of language on perception: listening to sentences about faces affects the perception of faces. *Journal of Neuroscience 30*, 45 (2010), 15254–15261.

[142] LaPiere, R. T. Attitudes vs. actions. *Social forces 13*, 2 (1934), 230–237.

[143] Léger, L., and Chevalier, A. Location and orientation of panel on the screen as a structural visual element to highlight text displayed. *New Review of Hypermedia and Multimedia 23*, 3 (2017), 207–227.

[144] Levi, D. M. Crowding—an essential bottleneck for object recognition: A mini-review. *Vision research 48*, 5 (2008), 635–654.

[145] Li, H., Scells, H., and Zuccon, G. Systematic review automation tools for end-to-end query formulation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 2141–2144.

[146] Li, Q., Morris, M. R., Fourney, A., Larson, K., and Reinecke, K. The impact of web browser reader views on reading speed and user experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12.

[147] Li, Y. User Perception Affects Search Engine Advertising Avoidance: Moderating Role of User Characteristics. *Social Behavior and Personality: an international journal 47*, 4 (2019), 1–12.

[148] Liao, M.-J., Wu, Y., and Sheu, C.-F. Effects of perceptual complexity on older and younger adults' target acquisition performance. *Behaviour & Information Technology 33*, 6 (2014), 591–605.

[149] Lindsay, G. W. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience 14* (2020), 29.

[150] Liu, J., and Han, F. Investigating reference dependence effects on user search interaction and satisfaction: A behavioral economics perspective. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 1141–1150.

[151] Liu, Z., Crouser, R. J., and Ottley, A. Survey on Individual Differences in Visualization. *Computer Graphics Forum 39*, 3 (2020), 693–712.

[152] Liu-Thompkins, Y. A decade of online advertising research: What we learned and what we need to know. *Journal of advertising 48*, 1 (2019), 1–13.

[153] Lyons, M. J., York, T. P., Franz, C. E., Grant, M. D., Eaves, L. J., Jacobson, K. C., Schaie, K. W., Panizzon, M. S., Boake, C., Xian, H., et al. Genes determine stability and the environment determines change in cognitive ability during 35 years of adulthood. *Psychological Science 20*, 9 (2009), 1146–1152.

[154] Mahmud, Murni, M., Kurniawan, and Hastuti, H. Involving psychometric tests for input device evaluation with older people. *OZCHI '05: Proceedings of the 17th Australia conference on Computer-Human Interaction* (2005), 1–10.

[155] Manser, M. P., and Hancock, P. A. The influence of perceptual speed regulation on speed perception, choice, and control: Tunnel wall characteristics and influences. *Accident Analysis & Prevention 39*, 1 (2007), 69–78.

[156] Mark, G., Czerwinski, M., and Iqbal, S. T. Effects of individual differences in blocking workplace distractions. *Conference on Human Factors in Computing Systems - Proceedings 2018-April* (2018).

[157] Mark, G., Iqbal, S., and Czerwinski, M. How blocking distractions affects workplace focus and productivity. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (2017), pp. 928–934.

[158] Matthews, T., Czerwinski, M., Robertson, G., and Tan, D. Clipping lists and change borders: Improving multitasking efficiency with peripheral information design. In *Conference on Human Factors in Computing Systems - Proceedings* (2006), vol. 2, pp. 989–998.

[159] Maxwell, D., and Azzopardi, L. Stuck in traffic: How temporal delays affect search behaviour. In *Proceedings of the 5th information interaction in context symposium* (2014), pp. 155–164.

[160] Maxwell, D., Azzopardi, L., and Moshfeghi, Y. A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017), pp. 135–144.

[161] Maxwell, D., Azzopardi, L., and Moshfeghi, Y. The impact of result diversification on search behaviour and performance. *Information Retrieval Journal 22*, 5 (2019), 422–446.

[162] Maxwell, D. M. *Modelling search and stopping in interactive information retrieval*. PhD thesis, University of Glasgow, 2019.

[163] McAvoy, J., and Brace, N. *Investigating Methods*. The Open University, 2014.

[164] McCay-Peet, L., Lalmas, M., and Navalpakkam, V. On saliency, affect and focused attention. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), pp. 541–550.

[165] McGregor, M., Azzopardi, L., and Halvey, M. A systematic review of cost, effort, and load research in information search and retrieval, 1972-2020. *ACM Transactions on Information Systems* (2023).

[166] Mikolajewicz, N., and Komarova, S. V. Meta-analytic methodology for basic research: a practical guide. *Frontiers in physiology 10* (2019), 203.

[167] Moacdieh, N., and Sarter, N. Clutter in electronic medical records: examining its performance and attentional costs using eye tracking. *Human factors 57*, 4 (2015), 591–606.

[168] Moacdieh, N., and Sarter, N. Display clutter: A review of definitions and measurement techniques. *Human factors 57*, 1 (2015), 61–100.

[169] Moacdieh, N. M., and Sarter, N. Using eye tracking to detect the effects of clutter on visual search in real time. *IEEE Transactions on Human-Machine Systems 47*, 6 (2017), 896–902.

[170] Montani, V., Facoetti, A., and Zorzi, M. The effect of decreased interletter spacing on orthographic processing. *Psychonomic Bulletin and Review 22*, 3 (2015), 824–832.

[171] Morris, M., Fourney, A., Ali, A., and Vonessen, L. Understanding the needs of searchers with dyslexia. In *Proc. 36th ACM CHI* (2018).

[172] Moura, R., and Casais, B. Search engine marketing to attract international digital traffic. In *Digital Marketing Strategies and Models for Competitive Business*. IGI Global, 2020, pp. 74–102.

[173] Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., and Aromataris, E. Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology 18*, 1 (2018), 1–7.

[174] Naghib, F., Mirzabeigi, M., and Alborzi, M. The role of spatial intelligence in predicting web information searching behavior and performance of high school students. *Library Hi Tech 39*, 1 (jan 2020), 48–63.

[175] Nisbett, R. E., and Miyamoto, Y. The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences 9*, 10 (2005), 467–473.

[176] O'Brien, H. L., Dickinson, R., and Askin, N. A scoping review of individual differences in information seeking behavior and retrieval research between 2000 and 2015. *Library and Information Science Research 39*, 3 (2017), 244–254.

[177] O'Brien, H. L., and McCay-Peet, L. Asking "good" questions: Questionnaire design and analysis in interactive information retrieval research. *CHIIR 2017 - Proceedings of the 2017 Conference Human Information Interaction and Retrieval* (2017), 27–36.

[178] O'Brien, M. A., Rogers, W. A., and Fisk, A. D. Understanding age and technology experience differences in use of prior knowledge for everyday technology interactions. *ACM Transactions on Accessible Computing 4*, 2 (2012).

[179] Ojanpää, H., Näsänen, R., and Kojo, I. Eye movements in the visual search of word lists. *Vision Research 42*, 12 (2002), 1499–1512.

[180] Oulasvirta, A., Hukkinen, J. P., and Schwartz, B. When more is less: the paradox of choice in search engine use. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (2009), pp. 516–523.

[181] PageFair, and Adobe. The cost of ad blocking. 2015 Ad Blocking Report. Tech. rep. `https://www.gwern.net/docs/advertising/2015-pagefair.pdf`.

[182] Palan, S., and Schitter, C. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance 17* (2018), 22–27.

[183] Palmquist, R. A., and Kim, K. S. Cognitive style and on-line database search experience as predictors of Web search performance. *Journal of the American Society for Information Science and Technology 51*, 6 (2000), 558–566.

[184] Pankok Jr, C., and Kaber, D. B. An integrated measure of display clutter based on feature content, user knowledge and attention allocation factors. *Ergonomics 61*, 5 (2018), 682–696.

[185] Pastushenko, O., Hynek, J., and Hruška, T. Generation of test samples for construction of dashboard design guidelines: Impact of color on layout balance. In *World conference on information systems and technologies* (2018), Springer, pp. 980–990.

[186] Pengnate, S. F., Sarathy, R., and Lee, J. K. The Engagement of Website Initial Aesthetic Impressions: An Experimental Investigation. *International Journal of Human-Computer Interaction 35*, 16 (2018), 1517–1531.

[187] Pham, M. T., Rajić, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., and McEwen, S. A. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Research synthesis methods 5*, 4 (2014), 371–385.

[188] Pogacar, F., Ghenai, A., Smucker, M., and Clarke, C. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In *Proc. 7th ACM ICTIR* (2017), pp. 209–216.

[189] Prater, M. 25 google search statistics to bookmark asap. Tech. rep., 2021. `https://rb.gy/zo2cz`.

[190] Puljak, L., and Sapunar, D. Acceptance of a systematic review as a thesis: survey of biomedical doctoral programs in europe. *Systematic reviews 6*, 1 (2017), 1–8.

[191] Radlinski, F., Broder, A., Ciccolot, P., Gabrilovich, E., Josifovski, V., and Riedel, L. Optimizing relevance and revenue in ad search: A query substitution approach. In *Proc. 31st ACM SIGIR* (2008), pp. 403–410.

[192] Raghavan, H., and Hillard, D. A relevance model based filter for improving ad quality. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (2009), pp. 762–763.

[193] Raptis, G. E., Katsini, C., Belk, M., Fidas, C., Samaras, G., and Avouris, N. Using eye gaze data and visual activities to infer human cognitive styles: method and feasibility studies. In *proceedings of the 25th conference on user modeling, Adaptation and Personalization* (2017), pp. 164–173.

[194] Rayner, K. Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition 5*, 4 (1977), 443–448.

[195] Rea, D. J., Seo, S. H., Bruce, N., and Young, J. E. Movers, Shakers, and Those Who Stand Still: Visual Attention-grabbing Techniques in Robot Teleoperation. *ACM/IEEE International Conference on Human-Robot Interaction Part F1271* (2017), 398–407.

[196] Recker, J. *Scientific research in information systems: a beginner's guide*. Springer Science & Business Media, 2012.

[197] Resnick, M., and Albert, W. The impact of advertising location and user task on the emergence of banner ad blindness: An eye-tracking study. *International Journal of Human-Computer Interaction 30*, 3 (2014), 206–219.

[198] Rosenholtz, R., Li, Y., and Nakano, L. Measuring visual clutter. *Journal of vision 7*, 2 (2007), 17–17.

[199] Sakai, T. Statistical significance, power, and sample sizes: A systematic review of sigir and tois, 2006-2015. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (2016), pp. 5–14.

[200] Salimun, C. *The relationship between visual interface aesthetics, task performance, and preference*. PhD thesis, University of Glasgow, 2013.

[201] Salthouse, T. A., and Coon, V. E. Interpretation of Differential Deficits: The Case of Aging and Mental Arithmetic. *Journal of Experimental Psychology: Learning, Memory, and Cognition 20*, 5 (1994), 1172–1182.

[202] Schnabel, T., Amershi, S., Bennett, P. N., Bailey, P., and Joachims, T. The impact of more transparent interfaces on behavior in personalized recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 991–1000.

[203] Schneider, C., Fulda, S., and Schulz, H. Daytime variation in performance and tiredness/sleepiness ratings in patients with insomnia, narcolepsy, sleep apnea and normal controls. *Journal of sleep research 13*, 4 (2004), 373–383.

[204] Seagull, F. J., and Walker, N. The effects of hierarchical structure and visualization ability on computerized information retrieval. *International Journal of Human-Computer Interaction 4*, 4 (oct 1992), 369–385.

[205] Sharit, J., Hernández, M. A., Czaja, S. J., and Pirolli, P. Investigating the roles of knowledge and cognitive abilities in older adult information seeking on the Web. *ACM Transactions on Computer-Human Interaction 15*, 1 (2008).

[206] Sheidin, J., Lanir, J., Conati, C., Toker, D., and Kuflik, T. The effect of user characteristics in time series visualizations. In *International Conference on Intelligent User Interfaces, Proceedings IUI* (2020), pp. 380–389.

[207] Silver, E. M., and Bennett, C. Modification of the minnesota clerical test to predict performance on video display terminals. *Journal of applied psychology 72*, 1 (1987), 153.

[208] Speier, C., and Morris, M. The influence of query interface design on decision-making performance. *MIS Quarterly: Management Information Systems 27*, 3 (2003), 397–423.

[209] Steichen, B., Carenini, G., and Conati, C. User-adaptive information visualization. In *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI '13* (New York, New York, USA, 2013), ACM Press, p. 317.

[210] Steichen, B., Conati, C., and Carenini, G. Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data. *ACM Transactions on Interactive Intelligent Systems 4*, 2 (jul 2014), 1–29.

[211] Steichen, B., Wu, M. M. A., Toker, D., Conati, C., and Carenini, G. Te,Te,Hi,Hi: Eye Gaze Sequence Analysis for Informing User-Adaptive Information Visualizations. In *User Modeling, Adaptation, and Personalization*. Springer International Publishing, 2014, pp. 183–194.

[212] Still, J., and Still, M. Influence of visual salience on webpage product searches. *ACM Transactions on Applied Perception (TAP) 16*, 1 (2019), 1–11.

[213] Strauss, E. Perception of emotional words. *Neuropsychologia 21*, 1 (1983), 99–103.

[214] Sullivan, G. M., and Artino Jr, A. R. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education 5*, 4 (2013), 541–542.

[215] Taber, K. S. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education 48*, 6 (2018), 1273–1296.

[216] Tangmanee, C. Fixation and recall of YouTube ad banners: An eye-tracking study. *International Journal of Electronic Commerce Studies 7*, 1 (2016), 49–76.

[217] Thilakaratne, M., Falkner, K., and Atapattu, T. A systematic review on literature-based discovery: General overview, methodology, & statistical analysis. *ACM Computing Surveys (CSUR) 52*, 6 (2019), 1–34.

[218] Toker, D., and Conati, C. Eye Tracking to Understand User Differences in Visualization Processing with Highlighting Interventions. In *User Modeling, Adaptation, and Personalization*. Springer International Publishing, 2014, pp. 219–230.

[219] Toker, D., Conati, C., and Carenini, G. User-adaptive Support for Processing Magazine Style Narrative Visualizations. In *23rd International Conference on Intelligent User Interfaces* (New York, NY, USA, mar 2018), ACM, pp. 199–204.

[220] Toker, D., Conati, C., and Carenini, G. Gaze analysis of user characteristics in magazine style narrative visualizations. *User Modeling and User-Adapted Interaction 29*, 5 (nov 2019), 977–1011.

[221] Toker, D., Conati, C., Carenini, G., and Haraty, M. Towards adaptive information visualization: on the influence of user characteristics. In *International conference on user modeling, adaptation, and personalization* (2012), Springer, pp. 274–285.

[222] Toker, D., Conati, C., Steichen, B., and Carenini, G. Individual user characteristics and information visualization: Connecting the dots through eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, apr 2013), ACM, pp. 295–304.

[223] Trielli, D., and Diakopoulos, N. Search as news curator: The role of google in shaping attention to news information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–15.

[224] Tsurukawa, J., Al-Sada, M., and Nakajima, T. Filtering visual information for reducing visual cognitive load. *UbiComp and ISWC 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the Proceedings of the 2015 ACM International Symposium on Wearable Computers* (2015), 33–36.

[225] Turpin, L., Kelly, D., and Arguello, J. To Blend or Not to Blend? In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (New York, NY, USA, jul 2016), ACM, pp. 1021–1024.

[226] Vallat, R. Pingouin: statistics in python. *Journal of Open Source Software 3*, 31 (2018), 1026.

[227] Velez, M. C., Silver, D., and Tremaine, M. Understanding visualization through spatial ability differences. *Proceedings of the IEEE Visualization Conference* (2005), 65.

[228] Villa, R., and Halvey, M. Is relevance hard work? Evaluating the effort of making relevant assessments. *Proc. 36th ACM SIGIR* (2013), 765–768.

[229] Villarreal-Narvaez, S., Vanderdonckt, J., Vatavu, R.-D., and Wobbrock, J. O. A systematic review of gesture elicitation studies: What can we learn from 216 studies? In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (2020), pp. 855–872.

[230] Visuri, A., Hosio, S., and Ferreira, D. Exploring mobile ad formats to increase brand recollection and enhance user experience. In *Proc. 16th MUM* (2017), pp. 311–319.

[231] Wan, H., Du, Z., and Yan, Q. The speed control effect of highway tunnel sidewall markings based on color and temporal frequency. *Journal of Advanced Transportation 50*, 7 (2016), 1352–1365.

[232] Weidt, F. N., and Silva, R. Systematic literature review in computer science-a practical guide. *Relatórios Técnicos Do DCC/UFJF 1* (2016).

[233] Wijntjes, M. W., and Rosenholtz, R. Context mitigates crowding: Peripheral object recognition in real-world images. *Cognition 180*, July (2018), 158–164.

[234] Wilson, M. *Interfaces for information retrieval.* Facet, 2011, p. 139–170.

[235] Wu, W.-C., Kelly, D., Edwards, A., and Arguello, J. Grannies, tanning beds, tattoos and nascar: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th information interaction in context symposium* (2012), pp. 254–257.

[236] Wu, W.-C., Kelly, D., and Sud, A. Using information scent and need for cognition to understand online search behavior. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (2014), pp. 557–566.

[237] Yeal Lee, S., and Cho, Y.-s. Do Web Users Care About Banner Ads Anymore? The Effects of Frequency and Clutter in Web Advertising. *Journal of Promotion Management 16* (2010), 288–302.

[238] Yeh, S.-L., He, S., and Cavanagh, P. Semantic priming from crowded words. *Psychological science 23*, 6 (2012), 608–616.

[239] Zen, M., and Vanderdonckt, J. Assessing user interface aesthetics based on the inter-subjectivity of judgment. In *Proceedings of the 30th International BCS Human Computer Interaction Conference 30* (2016), pp. 1–12.

[240] Zhai, C. Interactive information retrieval: Models, algorithms, and evaluation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 2444–2447.

[241] Zhang, F., Mao, J., Liu, Y., Ma, W., Zhang, M., and Ma, S. Cascade or recency: Constructing better evaluation metrics for session search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 389–398.

[242] Zhang, M., Feng, J., Ma, K., Lim, J., Zhao, Q., and Kreiman, G. Finding any Waldo with zero-shot invariant and efficient visual search. *Nature 9*, 1 (2018).

[243] Zhang, S., Grenhart, W. C., McLaughlin, A. C., and Allaire, J. C. Predicting computer proficiency in older adults. *Computers in Human Behavior 67* (feb 2017), 106–112.

[244] Zheng, B., Udeh-Momoh, C., Watermeyer, T., de Jager Loots, C. A., Ford, J. K., Robb, C. E., Giannakopoulou, P., Ahmadi-Abhari, S., Baker, S., Novak, G. P., et al. Practice effect of repeated cognitive tests among older adults: Associations with brain amyloid pathology and other influencing factors. *Frontiers in aging neuroscience 14* (2022).

[245] Zhou, K., Redi, M., Haines, A., and Lalmas, M. Predicting pre-click quality for native advertisements. *25th International World Wide Web Conference, WWW 2016* (2016), 299–310.

[246] Zhu, X., and Neupert, S. D. Dynamic awareness of age-related losses predict concurrent and subsequent changes in daily inductive reasoning performance. *British Journal of Developmental Psychology 39*, 2 (2021), 282–298.

[247] Ziefle, M., Brauner, P., and Speicher, F. Effects of data presentation and perceptual speed on speed and accuracy in table reading for inventory control. *Occupational Ergonomics 12*, 3 (2015), 119–129.

[248] Zimprich, D., and Kurtz, T. Individual differences and predictors of forgetting in old age: The role of processing speed and working memory. *Aging, Neuropsychology, and Cognition 20*, 2 (2013), 195–219.

Part VI

Appendices

Appendix A

Systematic Review

A.A Included and Excluded Studies in the Systematic Review

The complete list of both included and excluded studies in the Systematic Review, and categorised with the reason for exclusion, has been provided at the following link:

- https://doi.org/10.15129/9e4ddb4e-a417-4152-8609-212209b0176a

A.B Systematic Review Raw Results

The excel spreadsheet that was created in the Systematic Review, which categorised every included paper against a number of dimensions—including the type of results found and how the study was designed—can be accessed at the following link:

- https://doi.org/10.15129/226309e3-869e-495b-88a1-1c29fbd4a89c

Appendix B

Perceptual Speed Stimuli

B.A Finding A's Stimuli

The list of stimuli used for the Finding A's Perceptual Speed test has been colour-coded for the purpose of this Appendix to make it clear which words contain the letter 'a' (red font) compared to words which do not (black font). This can be viewed at the following link:

- https://doi.org/10.15129/7f921e83-4043-4c65-9844-2ee0b137860a

B.B Number Comparison Stimuli

The list of Number Comparison Stimuli, including the index of change, and what the numbers changed from and to, are provided in an excel document which can be accessed with the following link:

- https://doi.org/10.15129/1fb105cb-c025-47fa-9122-72745aa1a7ef

Appendix C

Experimental Instructions

C.A Introduction Page for Experiment

News Search System Study

This HIT involves performing four search tasks of approx. 5-10 minutes each using four Newspaper search engines. For each search task, you will be given a topic, such as:

Find and bookmark articles that discuss instances of piracy, where DIFFERENT BOATS/SHIPS have been illegally taken control of or boarded.

Your goal, for each search task, is to find and save several relevant and different documents.

NOTE: Your HIT might be declined if your performance is too low i.e. if you save too many non-relevant documents compared to relevant documents.

This means you may not be rewarded for your HIT or that your total reward is reduced, if you do not perform the tasks accurately enough.

A short practice task will be provided so you can decide whether to participate.

After each search task, you will be asked to recall aspects of what you found. In the above, example, the different boats/ships involved. This will be followed by a series of short survey questions asking you about the difficulty of the topic, your search experience and the performance of each search engine.

The HIT will start with a short demographics survey (1 min), and then you will have the chance to undertake a short practice task (5 minutes max), where you can decide to continue or not.

IMPORTANT NOTES

- The HIT will take approximately **60 minutes**.
- You need to be **at least 18 years old**.
- This study must be completed in a **single sitting**.
- Do not exit the process unless you intend to decline the HIT.
- The study will appear in a pop-up box - do not resize the window.
- You can only participate in this study once.
- In order to conduct this experiment you need to use:
  - **Safari, Firefox or Chrome only, ensuring:**
  - that **JavaScript is enabled**;
  - that **pop-ups are enabled** for this site;
  - that **ad blockers are disabled** for this site;
  - you are using a **laptop or desktop computer**;
  - you device has a resolution of **at least 1024x768**; and
  - you are using a **mouse.**

If any of these conditions are not met, then your HIT may not be accepted.

- You must complete the search tasks successfully to be paid.
- All the collected data will be treated as confidential and it will not be possible to identify you as a person with the data after the experiment as ended.

C.B Instructions for Experiment

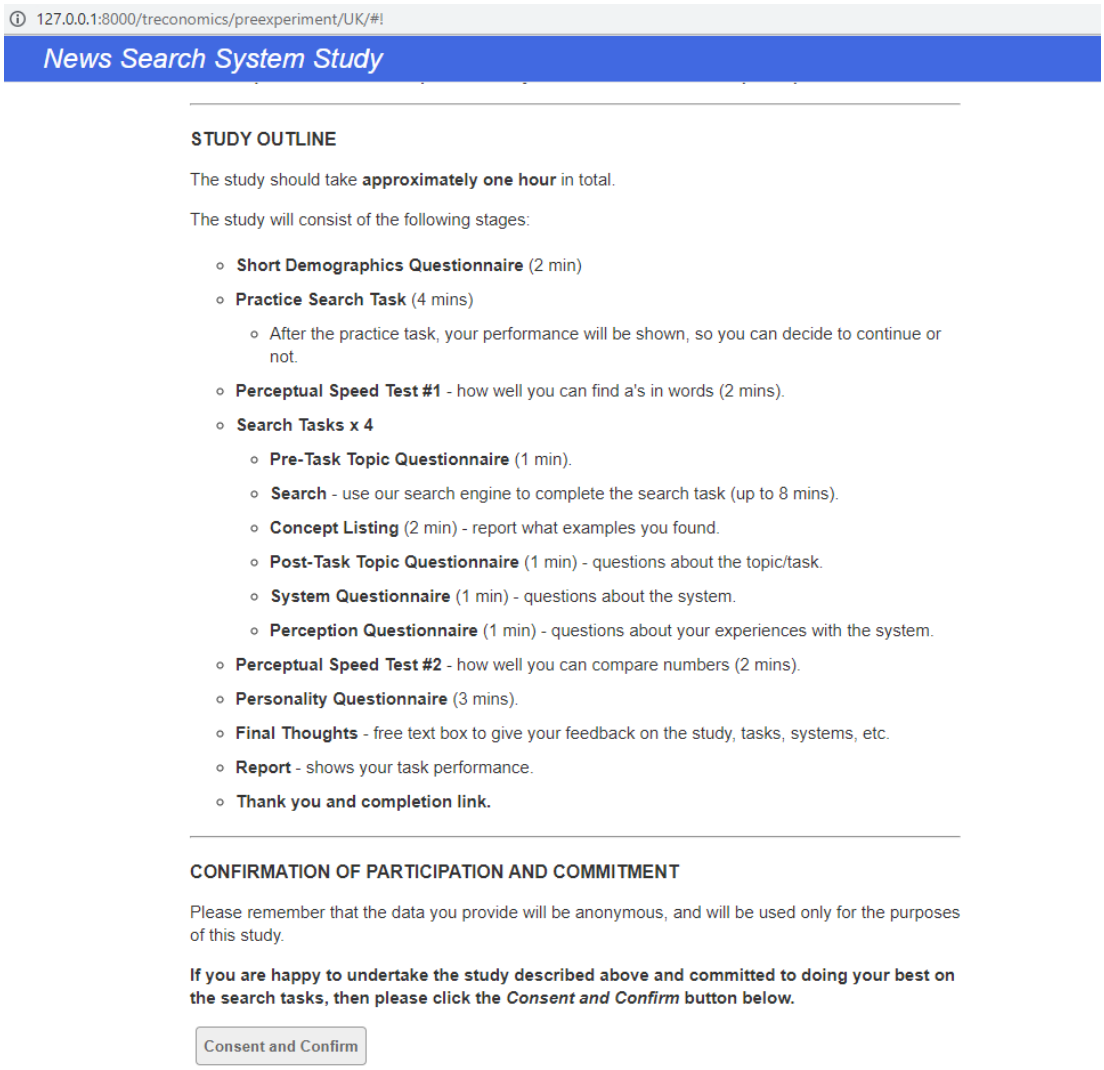


Figure C.1: The Study Outline provided to every user.

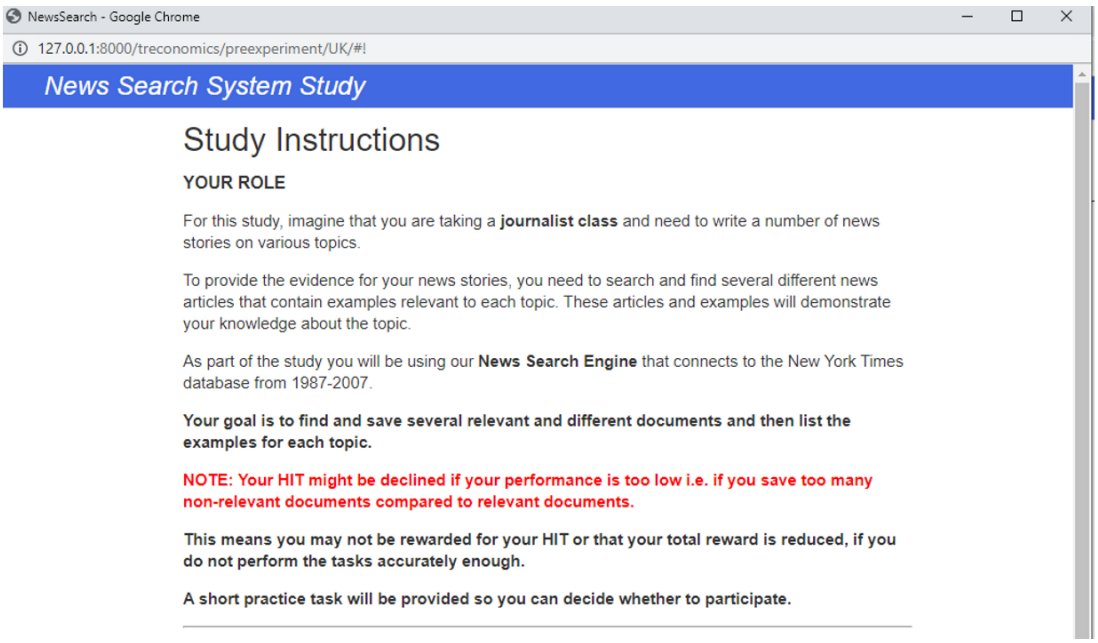
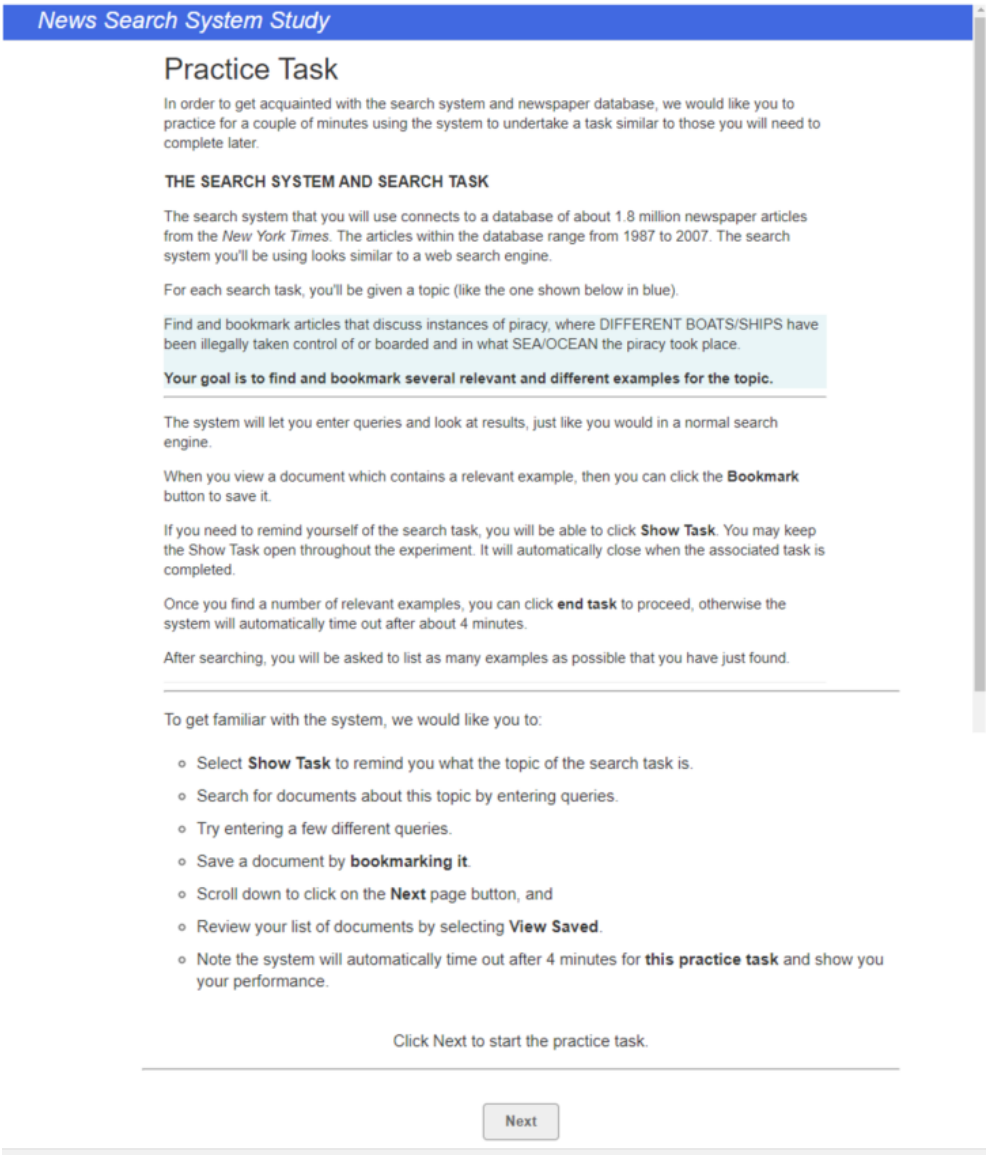


Figure C.2: The study instructions provided to every user.

C.C Practice Task Instructions



C.D Experiment Search Tasks

The topics and search scenarios used in the Experiment. Please note, the number in brackets refers to the topic number in the TREC AQUAINT test collection.

Piracy (№ 367). Find and bookmark articles that discuss instances of piracy, where DIFFERENT BOATS/ SHIPS have been illegally taken control of or boarded and in what SEA/OCEAN the piracy took place.

Wildlife Extinction (№ 347). Find and bookmark articles that discuss EXTINCTION PREVENTION MEASURES made by countries to protect DIFFERENT WILDLIFE SPECIES. Afterwards, you will be asked to recall these different wildlife species and their extinction prevention measures.

Airport Security (№ 341). Find and bookmark articles that discuss how DIFFERENT AIRPORTS use SECURITY MEASURES to better screen passengers and their carry-on luggage. Afterwards, you will be asked to recall these different airports and their security measures.

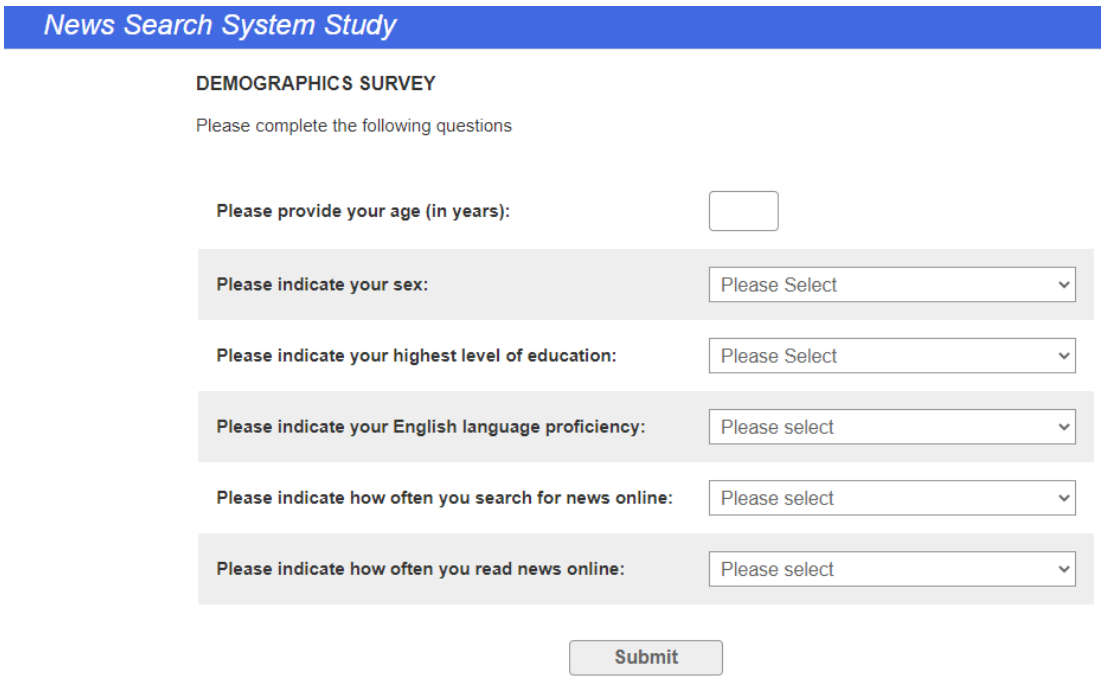
Tropical Storms (№ 408). Find and bookmark articles that discuss COUNTRIES that have had DIFFERENT TROPICAL STORMS (hurricanes and typhoons) which caused significant property damage and loss of life.. Afterwards, you will be asked to recall these countries and the different names of their tropical storms that caused fatal damage.

Curbing Population Growth (№ 435). Find and bookmark articles that discuss DIFFERENT COUNTRIES that have been successful in reducing population growth and the MEASURES they have taken to do so. Afterwards, you will be asked to recall these different countries and their measures to reduce population growth.

Appendix D

Experiment Surveys

D.A Demographic Survey



The drop-down menu bar provided the following options for each question:

1. Please indicate your sex:

   (*Female; Male; Other; Prefer not to say.*)

2. Please indicate your highest level of education:

   (*High School; College / Diploma; Undergraduate / Bachelors; Masters; PhD; Prefer not to say.*)

3. Please indicate your English language proficiency:

   (*Native; Bilingual; Professional Working; Limited Working.*)

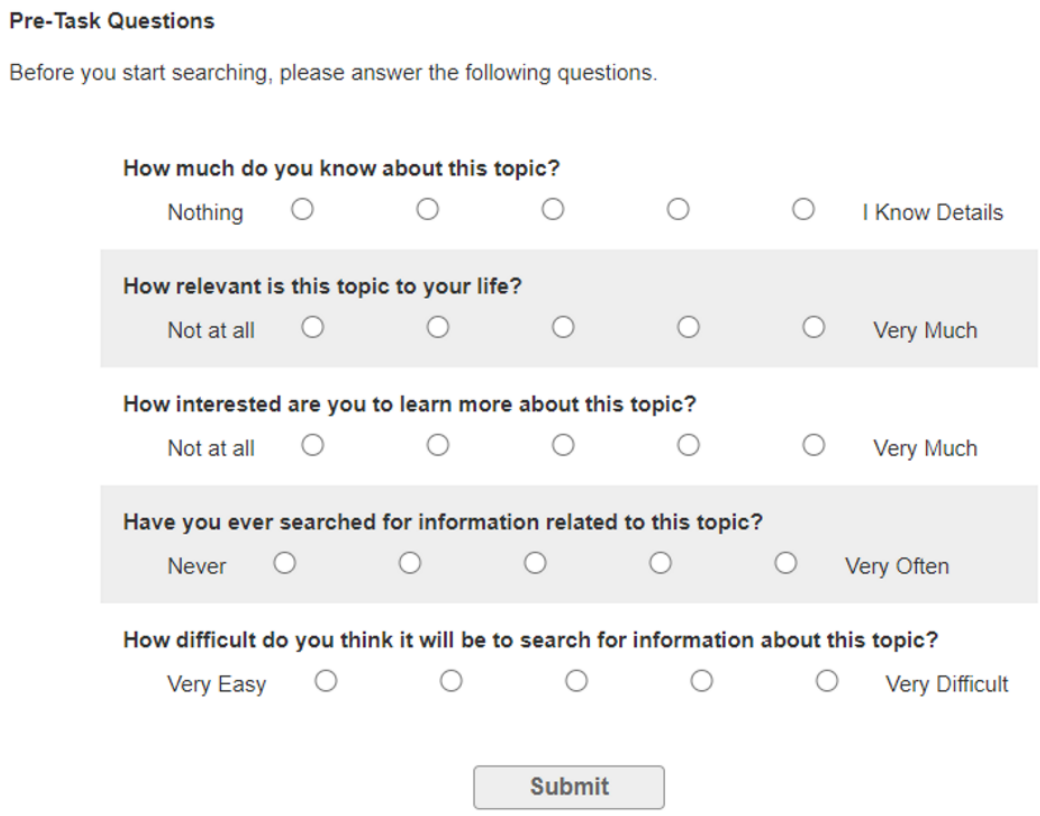
4. Please indicate how often you search for news online:

   (*Never; Sometimes; A few times a week; Many times a week; 1-2 times a day; Several times a day.*)

5. Please indicate how often you read news online:

   (*Never; Sometimes; A few times a week; Many times a week; 1-2 times a day; Several times a day.*)

D.B Pre-Task Survey



D.C Post-Task Surveys

Every post-task survey was completed immediately after every search task finished.

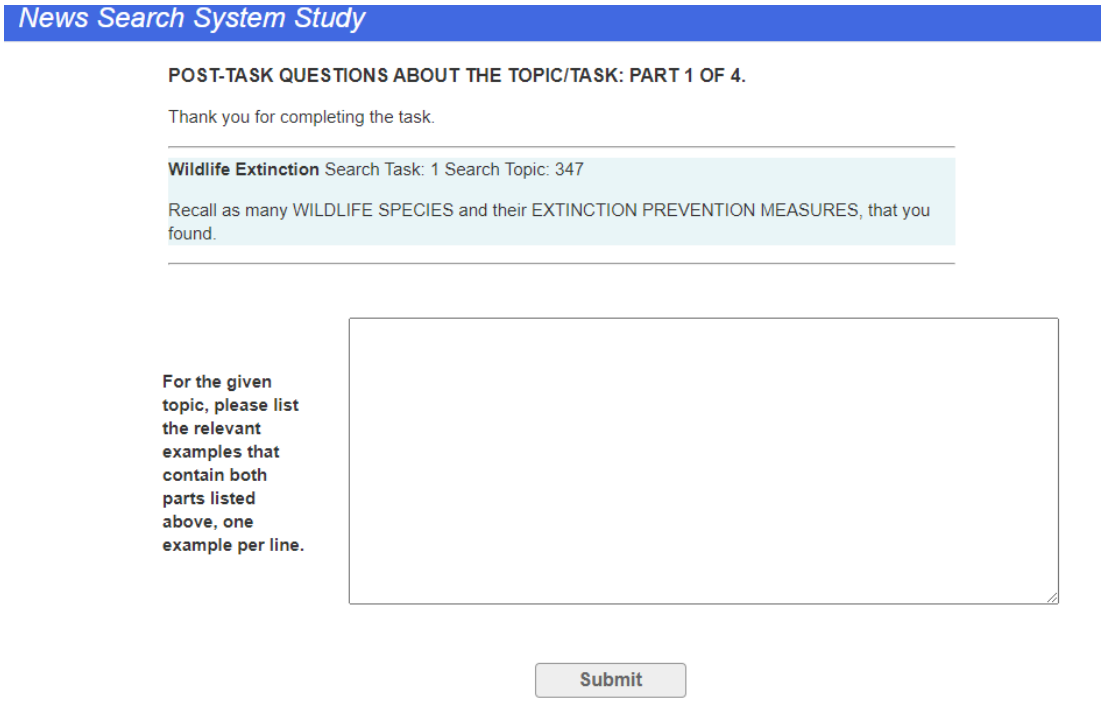


Figure D.1: Post-Task Concept Recall Survey

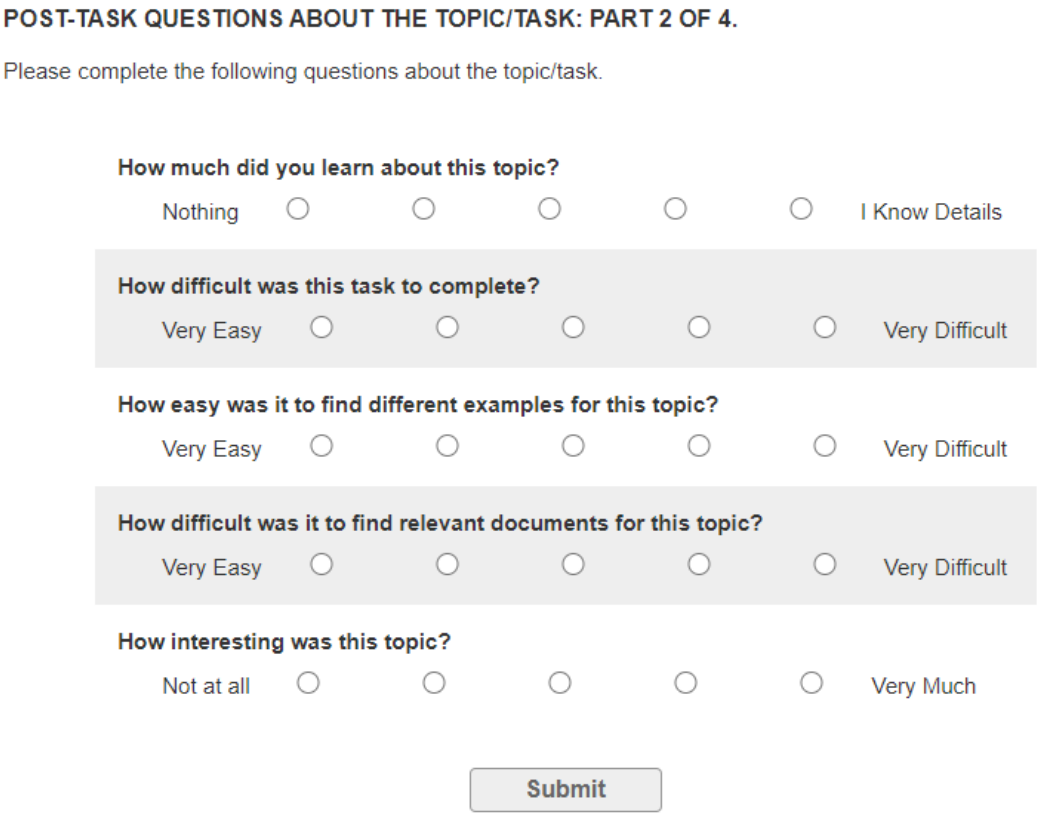


Figure D.2: Post-Task Topic Survey

POST-TASK QUESTIONS ABOUT YOUR PERCEPTIONS: PART 3 OF 4.

Please complete the following questions about your perception of completing the search task.

I felt frustrated while doing the task.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

I was confident in my decisions.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

I enjoyed completing this task.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

I was satisfied with my search performance.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

I checked each document carefully before saving.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

If present, I understood why the advertisements were being shown.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

I felt tired when completing this task.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

Submit

Figure D.3: Post-Task Perception Survey

POST-TASK QUESTIONS ABOUT THE SEARCH SYSTEM: PART 4 OF 4.

Please complete the following questions about the Search System.

The system was aesthetically appealing.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

The system was boring.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

The system was annoying.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

The system was easy to use.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

The system was confusing.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

The system was engaging.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

If present, the system showed related advertising.

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

Submit

Figure D.4: Post-Task System Survey

Appendix E

Experiment Datasets

E.A Original Experimental Logs

In the system that was designed to run the experiment, raw data was automatically generated for every measure of behaviour and performance in relation to each individual query a participant issued. This can be accessed at the following link:

- `https://doi.org/10.15129/852b32d0-b1a0-4a2f-9aa8-ba0dc0b1a6f4`

E.B Original Perceptual Speed Scores

The original scores for each participant in both PS tests can be viewed at the following link:

- `https://doi.org/10.15129/0227155f-242a-4a44-ae5a-baa0dc671172`

E.C RQ2 Correlations

To compute the correlations for answering Research Question 2: '***What is the relationship between Perceptual Speed and visual clutter in the form of advertisements during an IIR task?***', the original per-query data was transformed. Specifically, the sum of every query for each condition was created. Then, to create the *All-Ads* condition, the three conditions that contained ads were averaged. Every participant then had 1 row of data, which contained data for *No-Ads* (NA) and *All-Ads* (AA). This data can be accessed at the following link:

- `https://doi.org/10.15129/7849e287-1dc8-485e-bdab-5cb77420e87c`

E.D RQ2 ANOVAs

The ANOVAs compared Low-PS and High-PS, which had been categorised using Extreme Group Analysis in both PS tests. Every user had two rows of data: one represented all metrics from the *No-Ads* condition; and the other row represented the *All-Ads* condition.

- For Finding A's, there were 9 Low-PS participants and 9 High-PS participants. The data is available here:
  `https://doi.org/10.15129/9d28ea34-5487-433e-aaad-5f6a9ff78744`
- For Number Comparison, there were 12 Low-PS participants and 10 High-PS participants. The data is available here:
  `https://doi.org/10.15129/b3bb63d9-3bbb-46b7-8712-63440cd95e9b`

E.E RQ3 Correlations

To compute the correlations for answering Research Question 3: '***Does clutter that is congruent with the task improve or worsen the search experience for users with Low Perceptual Speed?***, the original per-query data was again transformed into the sum of each condition for every participant. This was then merged into 1 row of data per participant, where each participant had data for: *No-Ads* (NA), *Congruent-Ads* (CA), *Incongruent-Ads* (IA), and *Mixed-Ads* (MA). The data can be accessed at the following link:

- `https://doi.org/10.15129/ba09affb-e5b0-4038-88cf-fc6bd0d4fb89`

E.F RQ3 ANOVAs

The ANOVAs compared Low-PS and High-PS, which had been categorised using Extreme Group Analysis in both PS tests. Every user had four rows of data: the *No-Ads* condition; *Congruent-Ads* condition; *Incongruent-Ads* condition; and *Mixed-Ads* condition.

- For Finding A's, there were 9 Low-PS participants and 9 High-PS participants. The data is available here:
  `https://doi.org/10.15129/ea53890c-994f-44cb-b059-3b9c31515ea7`
- For Number Comparison, there were 12 Low-PS participants and 10 High-PS participants. The data is available here:
  `https://doi.org/10.15129/e260a744-7d22-4638-adc6-b680694d4848`

E.G Survey Data

The raw data for every participant's response to every survey administered can be accessed here:

- `https://doi.org/10.15129/c7e97aa3-bf69-441a-9585-206f495da81c`