Defining and Measuring Cost, Effort, and Load, within Information Seeking and Retrieval

PhD Thesis

Molly McGregor

Computer and Information Sciences
University of Strathclyde, Glasgow

January 22, 2026

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.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed: Molly McGregor

Date: 14.09.2025

Abstract

During the Information Seeking and Retrieval (ISR) process, users engage in interactions with search systems, submitting queries, retrieving documents, and examining results. These interactions place varying demands on users' internal and external resources. In ISR research, cost, effort, and load (CEL) are frequently invoked to explain and evaluate user behaviour and search experiences. However, the field lacks universally accepted definitions and standardised measurement methods, creating conceptual ambiguity and challenges in interpreting and comparing findings.

This doctoral research addresses these challenges through four novel contributions. First, a working definition framework provides operational definitions for CEL and a conceptual structure that researchers can adopt and refine, fostering a unified understanding within ISR. Second, a multi-stage relevance judgement model, grounded in existing theory, structures the document judgement task used in the empirical work and allows for the evaluation of user effort and load. Third, empirical studies characterise user effort and load between task stages, offering insights into how different judgement decisions influence these constructs. Finally, multiple measures of effort and load are evaluated to assess validity, triangulate findings, and examine their temporal dynamics.

Together, these contributions provide conceptual clarity and methodological guidance for ISR research, advancing understanding of CEL and offering empirically grounded approaches to modelling user behaviour during relevance judgement tasks.

Contents

List of Figures

List of Figures

List of Tables

List of Tables

Acknowledgements

I would like to thank my supervisors, Martin Halvey and Ian Ruthven, for their guidance, support, and encouragement throughout the course of this PhD. Their insights and feedback have been invaluable in shaping both the direction and quality of this work.

I am also grateful to my family and friends for their patience, understanding, and encouragement during this process. Many thanks to my parents, Mark and Tina, for their endless support, and for being on call as highly skilled childcare assistants whenever I've needed them.

Above all, I want to thank my husband, Leo, for his unwavering support, patience, and belief in me, which made the completion of this PhD possible.

Finally, I would like to thank my daughters, Hilvi and Elva, who were born during this PhD. Balancing this thesis with your daily adventures has ensured there's never a dull moment, and your joy and curiosity have made it all worthwhile.

I am also grateful to the Marie Skłodowska-Curie scholarship for providing the financial support that enabled my research within the DOSSIER project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 860721.

Preface

This thesis includes research that has been published in peer-reviewed journals. I served as the first author on these publications, contributing significantly to the conceptual development, data analysis, and writing of the manuscripts.

Part I

Introduction and Background

Chapter 1

Introduction

1.1 Motivation

The *Information Seeking and Retrieval* (ISR) process involves a wide range of user-system interactions. The user formulates and submits queries, examines results pages, and engages with a document to make a relevance decision [1]. All of these interactions, whether individually or collectively, will impose varying levels of demand on the user's internal (i.e., memory, attention) and external resources (i.e., time, money). In an ideal scenario, the user will have unlimited resources to attend to these demands and to source the perfect information to satisfy their need. However, in reality, these resources are finite in nature, and the user must therefore allocate their resources based on these limitations. For over 50 years, constructs related to user demand such as **effort** and **cost**, have been incorporated within *Information Retrieval* (IR) evaluation frameworks, and in more recent years the measurement of load (**cognitive load** and **workload**) has become increasingly prevalent within the field. A significant amount of ISR research has focused on cost, effort, and load (**CEL**) in relation to the search task [2–4],the system [5–7], and the individual characteristics of the user [8, 9]. Despite the wide research interest, there are two key challenges currently facing CEL examination within ISR. The first challenge relates to the lack of formal and universal definitions ascribed to these constructs, and the second challenge, relates to the scarcity of standardised or gold standard methods used to measure them.

This thesis seeks to address these challenges relating to **(a) the definition** and **(b) the measurement** of CEL within ISR. More specifically, the aim of this research is two-fold. Firstly, to establish working definitions and a conceptual framework for defining CEL, and secondly, to examine and evaluate existing measures of effort and load within an ISR context. To address part **(a) definition**, theoretical research was conducted in the form of a systematic review. From this, working definitions of CEL, and related constructs were developed, in addition to a conceptual framework. To address the primary methodological issues identified in the theoretical framework, the empirical component of this thesis (part b: measurement) was conducted through a series of user experiments. These experiments were designed around the ISR sub-task of document relevance judgment and were informed by the theoretical development of the **multi-stage relevance judgment model**, which served as the foundation for the user study task (see Chapter 4 for an outline of the model used). A variety of commonly used measures of effort and cognitive load were then incorporated to allow systematic evaluation.

1.2 Context

This thesis examines the definition and measurement of CEL within ISR. This thesis is structured into two primary components: (1) theoretical contributions and (2) empirical contributions. The theoretical work is used to inform the empirical work.

The theoretical contributions encompass two key parts: first, a comprehensive systematic review (Chapter 3) that offers an innovative framework for defining CEL in the context of ISR; and second, drawing on existing literature, the development of a multi-stage relevance judgement model (Chapter 4).

The empirical contributions outlined in this thesis are formed of four studies (Chapters 7, 8, and 9) (three different studies and one replication study). As aforementioned, the theoretical work was used to inform the empirical work. Firstly, the analysis conducted for the systematic review facilitated the identification of the most effective methods for measuring effort and load, in terms of their ability to accurately reflect the

properties of these constructs. When designing the experiments for the empirical work, the findings from the systematic review were utilised to inform decisions regarding the selection of appropriate measures and the optimal timing for their application. At this stage, it was determined that the empirical work would concentrate solely on the measurement of effort and load. Following the theoretical work, cost was considered less directly related to user behaviour, as it primarily reflects the consumption of external resources (e.g., time, money), whereas effort and load capture the internal processes that more accurately represent how users interact with and experience the task. Secondly, the multi-stage relevance judgement model developed in the theoretical work was utilised across all of the studies as a means to examine user effort and load at different document judgement stages. Although the primary focus of the empirical work is on the evaluation of effort and load measures at different stages of document judgement, certain user characteristics that may influence these measurements were also investigated. These included: **perceptual speed**, **working memory**, **motivation**, **topic knowledge**. All studies were conducted remotely using crowd-sourced participants. The first study served as a pilot, and subsequent studies were developed through an iterative process. In each iteration, the study design was refined to optimise the evaluation of effort and load measures. Additionally, adjustments were made to better understand how the various judgement stages impacted these constructs. A final study was carried out as a replication of Study 3, which was regarded as the strongest in terms of methodological rigour. This replication was undertaken to confirm and validate its findings.

1.3 High-Level Research Questions

The overarching aim of this thesis is to investigate how CEL are defined and measured within ISR. To achieve this aim, a set of high-level research questions and corresponding sub-questions has been formulated. HL-RQ1 establishes the theoretical foundation by reviewing how CEL have been defined and measured within ISR. Building on this, HL-RQ2 develops a multi-stage relevance judgement model integrating effort and load, which then provides the conceptual basis for HL-RQ3, where effort and load are em-

pirically measured across the stages of document judgement.

This thesis focuses on **(a) defining cost, effort, and load** and **(b) measuring effort and load** during a multi-stage relevance judgement task. More specifically, this research endeavours to address the following **High-Level (HL)** research questions and their corresponding sub-questions.

HL-RQ1 (Theoretical): How has cost, effort, and load (CEL) been defined and measured within Information Seeking and Retrieval (ISR)?

- (a) How have CEL, and their related constructs been defined within ISR?
- (b) Which methods have been used/proposed to measure CEL, within ISR?
- (c) What are the relationships between the different definitions of CEL and the methods used to measure them?

HL-RQ2 (Theoretical Model): How can effort, and load be integrated into a multi-stage model of relevance judgement?

- (a) What are the key stages involved in relevance judgement, based on existing theory?
- (b) How do effort, and load theoretically influence each stage of relevance judgement?
- (c) How can these theoretical relationships be structured to form a coherent multi-stage model?

HL-RQ3 (Empirical): How can effort and load be measured within the ISR sub-task of multi-stage document judgement, as defined by the theoretical model?

- (a) How do different stages of document judgement vary in terms of user effort and load?
- (b) To what extent is effort and load influenced by the type of judgement rating (i.e.,"no", "partially", "yes") between document judgement stages?

- (c) What are the operational relationships between effort and load?
- (d) How do measures of effort and load change over time during the task?
- (e) To what extent do user characteristics influence measures of effort and load?

1.4 Contribution

The primary contributions of this thesis stem from the theoretical work undertaken and the empirical findings derived from the four relevance judgement user studies. More specifically, the main contributions of this thesis are as follows:

1. **Working Definition Framework**: An initial contribution of this work is the development of working definitions and a conceptual framework for defining CEL as a means to overcome some of the conceptual challenges highlighted in the systematic review (Chapter 3). The purpose of this framework is to offer the ISR community a set of operational definitions that other researchers can adopt and further develop in their own studies. This collaborative approach seeks to foster a more unified understanding and analysis of CEL constructs within the context of ISR.

2. **Multi-Stage Relevance Judgement Model:** The second major contribution of this thesis is the development of a multi-stage relevance judgement model (Chapter 4). Drawing on established theory from the ISR relevance literature, this model underpins the document judgement task used across all studies in this work. Each stage of the model is theoretically designed to vary in task complexity, allowing for a natural manipulation of intrinsic cognitive load and enabling a nuanced examination of effort and load measures within a realistic task framework.

3. **Characterising User Effort and Load between Task Stages:** The set of user studies conducted for this thesis empirically tested theoretical predictions regarding the varying complexity of different document judgement stages. This

work provides novel insights into how specific judgement decisions influence user effort and cognitive load, highlighting stage-dependent variations in cognitive demands. By identifying these patterns, the research contributes to the development of more accurate and nuanced user models that reflect real-world behaviour during relevance assessment. These models can inform the design of adaptive IR systems, optimise task workflows, and support the creation of interfaces that better accommodate users' cognitive limitations, thereby bridging the gap between theoretical understanding and practical application in information-seeking contexts.

4. **Evaluating Effort and Load Measures**: Building on the insights gained from the theoretical work, the studies in this thesis employed multiple measures of effort and cognitive load within each study to rigorously evaluate their validity. This multi-measure approach enabled the triangulation of effort and load assessments, providing a more robust understanding of these constructs. By examining the correlational relationships between different measures, the research offers valuable guidance on which assessment methods are most reliable and informative. Furthermore, by strategically varying the timing of effort and load measurements across studies, the work provides novel insights into the temporal dynamics of these constructs vary during the document judgement task, shedding light on how cognitive demands evolve throughout the task.

1.5 Thesis Summary

The work presented in this thesis is organised as follows:

Chapter 1- Introduction: This section outlines the motivation, identifies the challenges, and highlights the key contributions of the thesis.

Chapter 2- Background: This section outlines the motivation for this research and reviews literature on CEL in ISR. It covers the origins and development of CEL in ISR, examines how these constructs have been measured and linked to user search behaviour, and presents definitions from both ISR and related disciplines. The section

also summarises methods for assessing CEL and identifies gaps in the literature that this thesis aims to address.

Chapter 3- Systematic Review: This section presents the first component of theoretical work undertaken for this thesis: a systematic review that offers a comprehensive analysis of CEL research within ISR over the past 50 years. It also introduces a novel framework of definitions, which was developed based on the analysis of the literature.

Chapter 4- Relevance Judgement Model: This section presents the second component of theoretical work undertaken for this thesis: the development of a multi-stage relevance judgement model. This model is included in all of the studies presented in this thesis.

Chapter 5- General Methodology: This section provides an overview of the general methodology used in all studies presented in this thesis. This includes, the motivation and theoretical justifications for the experimental design used. It also outlines the technical details related to the documents/topics used; the experimental interface; the experimental procedure; participant recruitment. The methods used to measure effort and load are also detailed.

Chapter 6- User Studies: This section outlines the order of studies and the way in which they differ from one another. Relevant work related to each research sub-question and motivations for each study is also reviewed.

Chapter 7- Study 1 (Pilot Testing the Model and Measures): Presents the first study which focused on examining the relationships between effort and load measures during multi-stage document judgement. The influence of user characteristics (perceptual speed and working memory) on user effort and load was also examined.

Chapter 8- Study 2 (Between-Subjects Evaluation of Model and Measures): Presents the second study which addressed several of the limitations from the initial study, while further investigating how document judgement stages differ in relation to effort and load.

Chapter 9- Study 3 (Within-Subjects Evaluation of Model and Measures): Presents the third and final study which explores how effort and load differ

between document judgement stages - in addition to examining the relationships between the different measures. The influence of the user characteristics, topic knowledge and motivation on effort and load is also investigated. This chapter also presents the replication results of the third study.

Chapter 10- General Discussion: This section summarises and discusses the findings from both the theoretical and empirical work, considering them in relation to the research questions outlined in Section 1.3. It presents the implications of the findings for the ISR research landscape, highlights the limitations of the thesis, and offers recommendations for future research.

Chapter 11- Conclusion: This chapter summarises how the the main research questions were addressed and concludes the thesis.

1.6 Publications

A portion of the work presented in this doctoral thesis has been previously published at the the following peer-reviewed conferences or journals. These are listed chronologically by publication date.

1. M. Mcgregor, L. Azzopardi, and M. Halvey, "Untangling Cost , Effort , and Load in Information Seeking and Retrieval," pp. 151–161, 2021. [10]

2. M. McGregor, L. Azzopardi, and M. Halvey, "A Systematic Review of Cost, Effort, and Load Research in Information Search and Retrieval, 1972–2020," ACM Transactions on Information Systems, vol. 42, no. 1, pp. 1–39, 1 2023. [11]

3. M. McGregor, "Defining and Measuring Cost, Effort, and Load in Information Retrieval," SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, no. 2014, p. 3490, 2023. [12]

Chapter 2

Background

2.1 Overview of Cost, Effort, and Load Research in Information Seeking and Retrieval

Many interactions take place during the Information Seeking and Retrieval (ISR) process, users type and submit queries, examine search engine result pages, and engage with the retrieved information to assess whether it satisfies their information need. All of these processes will tax both the users internal resources, i.e., those related to cognitive processing, such as their working memory and attention, and the users external resources, i.e., the time they spend on the search task. These resources are finite and such limitations will inform the decision making of the user during the search process. For more than five decades, ISR researchers have acknowledged how the search process and interactions can place specific demands on user resources. As a result, measures of effort and cost have been integrated into evaluation frameworks alongside other critical metrics, such as precision and recall.

Cost was first introduced in the context of Information Retrieval evaluation in the early 1970s [13] and during this time was mostly treated as a "fiscal cost" - predominantly examined for the purpose of comparing online bibliographic IR compared to manual IR within organisations. However, the advent of online search engines in the early 1990s ushered in a highly competitive era for search, wherein network latency and download speeds became integral in influencing user experience [14]. With this brought

a resurgence of interest in cost-related evaluation within ISR, however, researcher attention had turned to the "temporal costs"(i.e., "time cost"), [1,15,16] as a method for assessing system efficiency and effectiveness, rather than the fiscal-costs of earlier years. Users have long been assumed to attach significant importance to the time invested in their search. Miller [17] demonstrated this in 1968 with the 2-second rule, advocating that an interactive search system should respond within this time-frame to maintain user focus and attention on the ongoing task. More recent studies reinforced this notion, revealing that delays in network and download speeds adversely impact user's perceived usefulness and interactions with web-pages [18,19]. Research into the effects of time delays on the Google search engine demonstrated that even a modest 400ms delay in returning search results resulted in a 0.59% decrease in searches conducted over a 6-week period [20]. Other studies have echoed similar findings, emphasising that as the cost of querying rises, users tend to issue fewer queries [1,21].

The exploration of **effort** within ISR has undergone a similar trajectory as cost, receiving attention during the 1970s-1980s, followed by a decline until experiencing a revival as a notable construct in the late 1990s. Early examinations of effort was often intertwined with cost, primarily discussed and measured in the context of an organisation, and the labour involved (time/salary) in utilising different systems. Similar to cost, the resurgence of effort in the late 1990s may have been prompted by the rise of search engines, as it became more evident that effort played a crucial role in influencing user search decisions and behaviours. Nevertheless, effort examination within ISR still closely relates to the conceptualisation and operationalisation of cost, in the sense that in high-cost search scenarios, the user practices effort preservation tactics such as reducing the number of queries, viewing less pages, and spending less time examining a document. Thus, it is unsurprising that effort has long been conceptualised within ISR and subsequently measured in relation to both the users cognitive processes such as reading a document, formulating a search query, examining a results page, and their physical actions such as typing, scrolling, and mouse clicks. It has been proposed within ISR, that users base their search interaction decisions around the notion of *satisficing* - when an individual stops their search once the acquired information

is deemed sufficient to meet their information needs [22]. Satisficing relates closely to Zipf's "principle of least effort" Law [23], where the individual evaluates the value in continuing their search and potentially finding better information against the expected utility of that information [22]. This understanding of effort and its influence on the decision making process, has been a central focus of effort related research within ISR. As the demands placed on the users available resources increase during the search process, users tend to decrease their number of actions over time and may abandon the search task entirely if these demands exceed their resource capacity. For instance, Azzopardi [1] demonstrated that users performing their search on a "high-cost" interface, issued significantly less queries and examined more documents per query compared to those using the "low-cost" interface. Maxwell [21] observed a similar trend, indicating that as the relative cost of querying increased, users issued fewer queries and examined more documents per query rather than issuing additional queries. Furthermore, as the complexity of the search task itself increases, users subjectively report heightened levels of both effort [4] and workload [2]. Similar to task complexity, effort measures have been shown to correlate with all search task characteristics associated with difficulty, such as an increased number of steps to achieve the task goal; intellectual task product; and amorphous task goal. Increased task complexity and difficulty evoke the same effect on user behaviour as the "high-cost" search scenarios previously mentioned, with users executing fewer actions, clicks and bookmarks per query, as noted by Capra et al. [24]. Likewise, when the primary search goal of users is to meet their information needs, Yilmaz [25] observed that users tend not to invest time in determining the relevance of a document if they cannot quickly find the information they seek or of they perceive the relevant information as challenging to comprehend or process. In such cases, users often abandon the document completely and move on to the next. Even when a document is relevant to the query or the users information need, the documents utility to the user diminishes if they must exert more effort to locate and comprehend the relevant sections.

Gwizdka [26] used the construct of **cognitive load** to explain why certain aspects of the search process may influence these specific user behaviours. In their study, they

noted that average cognitive load varies across different task stages, highlighting that cognitive load is highest during stages such as query formulation and tagging of relevant documents compared to other stages such as examining search results and viewing individual documents. These findings imply that in a typical search scenario, the user may adapt their behaviour according to the level of demand imposed by a specific stage, and in alignment with the aforementioned studies, users are likely to invest their effort in task stages with lower demands, such as reviewing results or individual documents, while actively avoiding stages which impose higher demands, such as issuing queries or assessing the relevance of documents.

Although cognitive load is largely defined in relation to the demand-resource paradigm, the construct has received much less attention within ISR research compared to cost and effort. In fact, studies examining cognitive load appeared within ISR research almost two decades after cost and effort. Similarly, the construct of **workload** was not acknowledged within ISR until the late 1990s, however its measurement has since gained momentum within the last decade.

Given the significant impact that cost, effort, and load can have on user search decisions and behaviours, it is unsurprising that a considerable body of research in this domain has focused on developing search systems and interfaces designed to minimize the demands placed on users during the search process. Additionally, much of this research aims to understand which task elements and user characteristics may influence, or be influenced by, these demands. For example, in relation to the system and interface, studies have indicated that users exert less effort using systems which employ post-retrieval document clustering techniques compared to a less structured system [5]. In the context of search using a structured interface, users report significantly lower workload compared to those using a standard (traditional) interface [7]. Similar findings extend to user effort, where users invest less effort when search results are presented visually (i.e., 400 results per page) as opposed to a traditional vertical list (i.e., 10 results per page) [27]. Other studies have acknowledged however, that the influence of demands placed on users during these different search contexts may not be uniform across all users. For instance, users with higher cognitive abilities tend to

exert more effort as interface and task conditions become more complex [8]. Likewise, users with high working memory invest more effort during the search process than their low working memory counterparts [28], and users with high perceptual speed ability experience lower task demand compared to those with low perceptual speed ability [29]. Although these examples constitute a limited number of the studies investigating CEL within ISR, the discussion so far underscores the crucial role played by these constructs in the ISR process. Such findings, emphasise the significance of integrating these constructs into system evaluation frameworks. However, these constructs cannot be fully understood within the context of ISR until the field attains consensus regarding their definition, and in turn, their accurate measurement.

2.2 Overview of Existing Definitions: Cost, Effort, Load, and Related Constructs

The aim of this section is to present an overview of CEL constructs and their underlying theory. More specifically, this section refers to accepted CEL definitions from fields such as Psychology and Human Factors as a means to ground discussions surrounding CEL within ISR.

2.2.1 Conceptualisation and Operationalisation of Constructs

Before commencing this discussion, it is important to firstly present the notions of *conceptualisation* and *operationalisation*. Research methods from Psychology propose that in order to decipher an appropriate method of measurement, it it necessary that the construct is defined in relation to both its nominal and operational meaning [30]. Nominal definitions, otherwise referred to as conceptualisation, elucidate the *meaning* of the construct, while operational definitions, also known as operationalisation, precisely detail *how* the construct and its elements will be *measured* [31]. Given that conceptualisation precedes operationalisation, utilising ambiguous and vague definitions in the outset can lead to significant problems when it comes to measurement. For instance, when a term like "effort" is defined in multiple ways, the resulting operational

properties will also vary. As the operational properties determine which elements of the construct are measured, the measurement variables which emerge from the varying definitions will likely also differ, and subsequently will be used to measure what researchers believe to be the "same" construct. However, the lack of unity on conceptual definitions from the beginning, is likely to lead to research characterised by conceptual overlap, measures that lack clarity, and a diminished ability to establish causality. Despite, over 50 years of extensive discussion and exploration of CEL across various disciplines such as Psychology, Ergonomics, and Human Factors, a universal consensus on how to define these constructs has yet to be established. The following section will discuss each construct and their most commonly accepted definitions from disciplines outside of ISR.

2.2.2 Cost

The construct of cost has received various interpretations depending on the field of research. For instance, definitions from Cognitive Psychology and Neuroscience tend to associate cost with cognitive or mental costs [32,33]. In this context, individuals are viewed as valuing the effort they exert, treating the expenditure of effort as costly [34]. Human behaviour often involves a constant trade-off between effort and reward, but the limited capacity of cognitive resources constrains these trade-offs [35]. Consequently, the allocation of cognitive resources to a specific task at a given moment must be strategically chosen based on a cost-benefit analysis - such as "should I spend or conserve my resources to achieve this goal?" [35]. In this scenario, cost is operationalised in relation to energy expenditure - where in order to minimise negative outcomes such as fatigue, individuals are suggested to expend energy on tasks where the energetic costs are low but the benefits are comparably high. This cost-benefit analysis is not limited to metabolic cost. For instance, the cost-benefit analysis proposed in the domain of Behavioural Economics quantifies costs by assigning monetary value to low vs. high effort tasks [33]. Temporal costs are also frequently discussed in this trade-off process - in situations where time is considered valuable, for example in an organisational context, effort expenditure should be directed toward faster processes to achieve the task

goal [35]. Despite the different interpretations of what constitutes a "cost", a shared consensus among the various disciplines may be reached in the notion that the allocation of resources during the expenditure of effort leads to some degree of cost, and therefore cost is operationalised in this sense, as the *resources spent*, whether that be time, energy, or money.

2.2.3 Effort

Effort, also described as mental, cognitive, or physical effort, is a construct most humans are intuitively familiar with [36]. Similar to the construct of cost, interpretations of effort vary among disciplines. However, while universal definitions are yet to emerge, there are similarities among definitions. Psychology views effort as a purposeful and volitional phenomenon, involving the active participation of the individual rather than a non-autonomous, passive process [37]. Cognitive Psychology shares a similar standpoint, considering effort as both a physical and mental action or labour that varies in intensity and over time accumulates with the key purpose of accomplishing a task or goal [38]. Evolutionary Psychology offers a more transactional, extrinsic interpretation of effort, where the individual bases their effort exertion or the amount of "work" they are willing to put in, based on the incentives available to them - likened to a worker determining the amount of effort they are willing to expend based on incentives such as a bonus or salary.

2.2.4 Cognitive Load and Workload

Cognitive load and workload are often used interchangeably and while they emerged independently from different disciplines, they are arguably theoretically intertwined by the same core principles [39]. *Cognitive Load Theory* (CLT) [40] from Educational Psychology and *Multiple Resource Theory* (MRT) [41] from Ergonomics and Human Factors, underpin cognitive load and workload, respectively. Both theories overlap in relation to their understanding of human's limited resource capacity and competing task demands [40, 41]. The term *demand* is frequently discussed alongside definitions of load and refers to the elements of the task which determine how much physical or

mental exertion will be required to accomplish the task [38]. Cognitive Load Theory (CLT) has predominantly found application within the field of educational instruction and learning since it surfaced in the 1980s [42]. Within CLT, cognitive load is characterised as the overall mental burden placed on an individuals working memory at a specific point in time [43]. This characterisation reflects the roots of CLT and its derivation from the working memory model that underscores the constrained capacity of working memory vs. the abundant capacity of long-term memory in the human brain [44]. The degree of cognitive load that a person experiences is determined by the concurrent interaction of elements within their working memory. Given the finite capacity of working memory, there exists a limit to the volume of information it can manage simultaneously. Consequently, information overload occurs when an excess of information is presented at once, leading to reduced information processing by the individual [40]. Arguably, the most distinctive aspect of CLT is its differentiation among three distinct types of cognitive load: **intrinsic** (linked to the inherent characteristics of the task, such as difficulty); **extraneous** (imposed by the context in which the task is carried out); and **germane** (arising from the construction of schemas) [45]. CLT posits that these three cognitive load types are cumulative in nature [46].

Within Ergonomics and Human Factors research, mental workload can be considered as one of the most widely explored constructs. Nevertheless, researchers in the field have not yet achieved a unanimous agreement on its definition [47]. While certain aspects of CLT have been applied to characterise workload, definitions within Psychology generally interpret workload in relation to Multiple Resource Theory (MRT), specifically concerning processes such as task switching and attention allocation [48]. MRT suggests that the human brain possesses a fixed amount of mental resources. These resources can be likened to a collective pool of energy that can be utilised for various concurrent mental operations, spanning different tasks, modalities, and processing [41]. This theory proposes that a decline in performance arises when these resource pools become depleted, which can happen when two or more tasks necessitate a single resource [39]. Generally, mental workload is widely understood as the distribution of available resources to fulfill the requirements of a task (i.e., task demands) and the

cognitive experience which arises as a result of meeting those task demands [39, 48]. Considering the various interpretations of cognitive load and workload, it is clear that both constructs are conceptually linked by the principles of rising task demand and the consumption of available cognitive resources.

2.2.5 How are Cost, Effort, and Load Related?

It is apparent from the discussion in the previous sections that CEL are to a degree, conceptually entwined. Cognitive load occurs as a result of the task demand on our mental resources at a particular moment, and is subjectively experienced by the individual. Effort, on the other hand, is a deliberate reaction to the load and reflects the overall cognitive resources devoted to addressing the task demands over time, aiming to achieve a specific end goal [45]. From this perspective, it appears that effort is exerted, whereas cognitive load is experienced [45]. This dynamic process of effort exertion and cognitive load experienced in relation to the task demand are proposed to incur some kind of cost [38], whether that be affective/physical costs [49] (such as fatigue), temporal costs [35] (like the time spent), or economic costs [33] (such as monetary expense).

2.2.6 Summary

The previous sections have underscored that disciplines beyond ISR have encountered similar challenges in offering precise definitions of CEL. From the discussion provided, there exists a clear demarcation between these constructs in relation to their distinctive characteristics but also a degree of interrelation among them. The theoretical work carried out in this thesis, leverages these external representations of CEL to anchor the analysis of how these constructs are approached within the domain of ISR. Not only are these interpretations employed as a reference against which definitions can be assessed but they are also used as a foundation to scrutinise how the current measures employed within ISR possess the ability to accurately represent and mirror these complex constructs.

2.3 Measurement of Cost, Effort, and Load

The aim of this section is to present a general overview of the type of methods used to measure CEL constructs both within ISR, and also external disciplines.

2.3.1 Objective Methods

Objective methods of measurement are those which provide an impartial, and quantifiable outcome. These can be considered both *direct* and *indirect measures* [50]. Direct objective measures include eye-tracking, dual-task methods, or brain-activity measures such as functional near-infrared spectroscopy (fNIRS) [50].

Eye-tracking methods have been commonly used to assess cognitive load, and involve the measurement of trace fixation patterns, eye blinking, and pupil dilation [50]. Advantages of using eye-tracking approaches relate to their ability to collect data in real-time and blink activity, pupil size, and fixation duration have all been found to correlate significantly with varying levels of cognitive demand. The underlying theory suggests that each eye-tracking variable represents distinct yet complementary information about cognitive activity, where each variable is governed by different regions of the nervous system, contributing unique insights into cognitive activity [50]. However, a common criticism of eye-tracking measures is that some variables may capture phenomena unrelated to cognitive load or effort. For instance, pupil dilation has been associated with conditions such as depression and tiredness. It can also be affected by situational factors, such as the brightness of the room or stimuli [50].

Dual-task methodologies involve the undertaking of two tasks (primary and secondary task) concurrently, where performance on the secondary task (usually a simple activity which requires sustained attention) is proposed to reflect cognitive load [51,52]. The dual-task paradigm is grounded in the theory that the brain has a limited central processing system. When an individual is required to respond to or engage in a secondary task, it increases the load on the same working memory system that is concurrently being utilised by the primary task. Dual-task performance variables are normally taken as reaction time, accuracy, and error rate [46], where the outcome of

these variables are used as a direct measure of the load demanded during the dual-task scenario [53]. However, the results from the dual-task paradigm may be influenced by issues related to the task design itself. For example, if the task demand is too low, individuals may have enough cognitive resources to successfully complete both the primary and secondary tasks simultaneously [50]. Practice effects can also occur with repeated use and therefore performance differences may be attributed to learning rather than varying levels of cognitive load [50]. Taking these limitations into account, the dual-task methodology can be regarded as a sensitive and reliable technique, provided that task conditions are carefully considered.

Overall, direct objective measures are regarded as the most effective means of assessing cognitive load [50]. however, it is recognised that these methods raise questions of construct validity related to the potential problem of multi-causality. Whereby if differences in measurement data are detected, there could be several factors causing these differences. Therefore, it is crucial that experimental designs exercise great caution to ensure that the measure is not only reliable but also valid.

In addition to direct objective measures, effort and cognitive load can also be assessed using *indirect* objective measures. Such measures include physiological approaches like electroencephalography (EEG) and interaction data. These measures are considered indirect because they are influenced by the information processing and retrieval processes occurring during task performance. EEG records electrical activity across the surface of the brain which arise from currents in the brain. The data derived from EEG can provide temporal information during the processing of stimuli and can be used to infer increased demand [50]. Problems with EEG relate to its inability to distinguish between different load types and is also a costly method of measurement.

Data extracted from interaction logs can also be considered a widely used measure of effort and load. One of the most common measurement variables derived from interaction logs is time-on-task. It is widely acknowledged that all cognitive processes take time and these processes can be affected by a variety of factors such as task complexity, prior knowledge, time it takes to find information, and so forth. If these factors are properly controlled within a study, it becomes possible to identify which

processing variables are influencing time on task. However, it is crucial to recognise that the theoretical relationship between time-on-task and cognitive load is indirect, and there is limited evidence clarifying how these two variables are interconnected [54]. Other interaction data variables used to measure effort and load and often used within ISR, are variables such as number of clicks, pages visited, saved pages etc. While an increased number of clicks and pages visited may suggest greater exertion of effort and could thus be considered an indirect measure of cognitive load, these differences may also be attributed to a range of other factors. Therefore, experimental design must be approached with caution to ensure proper control over additional variables that may influence the results and are unrelated to cognitive load. Overall, interaction data is often considered a favourable choice for measuring effort and cognitive load, due to its ease of administration, objectivity, and the production of relatively reliable results [54].

2.3.2 Subjective Methods

Subjective approaches to measurement require human judgement of some kind, and are based on the notion that individuals can introspect on their cognitive processes and articulate this in some manner [50, 52]. Subjective measures can be considered one of the most popular methods of effort and load measurement [43]. Such measures most often involve self-report questionnaires, usually in the format of a Likert scale, which are composed of one or more semantic differential scales by which the individual can articulate the experienced level of mental burden. These scales are grounded in the assumption that individuals can accurately assess the level of demand they are experiencing during a specific scenario [43]. It is proposed that some constructs lend themselves better to human introspection than others. For instance, it has been shown that individuals are able to provide numerical indication of their perceived mental demand [52]. Many subjective methods are multi-dimensional and encompass a range of similar variables such as mental effort, frustration, temporal demand etc. Other scales are unidimensional, encompassing a single-item question related to the individuals perceived demand, for example a difficulty or effort rating. Despite, their unidimensional nature, these scales have been shown to be highly correlated with multi-dimensional tools and therefore are

just as valid and reliable in their assessment of load [46]. While the simplistic nature of subjective methods is appealing, they can have considerable limitations. Firstly, subjective measures primarily depend on the validity and accuracy of user reports. Secondly, they are often administered at task completion, and can therefore only provide a one-point post-hoc assessment of effort and load [50]. This results in a global assessment of effort and load, rather than providing insight into the specific aspects of the task that influenced the reported effort or load. This limitation can be mitigated by the repeat application of the self-report measure throughout the task, however very few studies have implemented this both within ISR and other domains [43]. Other issues with subjective measures relates to their content validity, primarily in regard to which type of cognitive load they are measuring but also how the measure relates to the task itself. Subsequently, these measures provide little information about which processes have influenced the level of demand experienced.

2.4 Summary

The previous section reviewed methods for measuring CEL, both objectively and subjectively, drawing on research from ISR and related disciplines. Across fields, the measurement of effort and cognitive load remains relatively underdeveloped and imperfect. ISR research has adopted many of these existing methods, but their application requires careful consideration, particularly given the dynamic nature of cognitive load and the influence of contextual and individual factors.

Part II

Theoretical Contributions

Chapter 3

Systematic Review

3.1 Motivation

This chapter presents a part of the theoretical work carried out for this thesis and is based on the article: [11]. I served as first author and was responsible for all aspects of the work, including study design, data collection, analysis, and manuscript preparation. In this thesis, the term *theoretical work* refers both to the examination and synthesis of existing theories and definitions from the literature, and to the development of a conceptual framework through which new definitions and relationships among constructs are proposed. This chapter seeks to explore the first high-level research question outlined in Section 1.3: **HL-RQ1: How has cost, effort, and load (CEL) been defined and measured within Information Seeking and Retrieval (ISR)?**, and the corresponding sub-questions:

- (a) How have CEL and their related constructs been defined within ISR?
- (b) Which methods have been used/proposed to measure CEL within ISR?
- (c) What are the relationships between the different definitions of CEL and the methods used to measure them?

To address these questions, this thesis draws on the findings from two interrelated literature reviews, however this chapter focuses primarily focuses on the systematic review. The first, a perspectives paper [10], provided a conceptual analysis of 26 studies,

highlighting key theoretical perspectives and identifying gaps in the literature; this review will be discussed briefly in the next section. The second, a systematic review [11], analysed a larger body of 91 studies using explicit inclusion criteria and structured synthesis methods. The main section of this chapter focuses on the systematic review, which builds on the conceptual insights from the perspectives paper while offering a more comprehensive and methodologically rigorous assessment of the literature.

3.1.1 Perspectives Paper

To address **HL-RQ1: How has cost, effort, and load (CEL) been defined and measured within Information Seeking and Retrieval (ISR)?**, the aims of the perspectives paper [10] was two fold: (*i*) to review the meaning of CEL related concepts used within ISR, and (*ii*) to create a shared taxonomy of the concepts relating to CEL in ISR. To achieve these aims, a literature review as conducted, where 397 papers were reviewed across a 20-year period, and 26 papers that explicitly proposed measures or definitions of CEL were selected for analysis. Identified definitions were extracted and organised according to their respective CEL constructs. Following the approach of Martinic and colleagues [55], the individual elements of each definition were further separated, categorised, and quantified into the following five conceptual categories:

- Interaction-Oriented/Count-Based
- Time-Oriented
- Cumulative/Total Work
- Meta-Cognition/Conscious Awareness
- Capacity-Based/Bounded Resources

To explore how CEL and related constructs were measured within ISR (RQ1b), the literature was examined to identify both subjective and objective measurement approaches. Finally, the relationships between CEL constructs, their conceptual categories, and the measurement methods were investigated (RQ1c). This analysis revealed

conceptual overlap, validity issues, and lack of theoretical justification for use of measures.

The findings emphasised the importance of clear and consistent definitions of CEL and identified several methodological challenges associated with their measurement. Drawing on theoretical perspectives from adjacent disciplines such as Psychology, the review also proposed a set of working definitions intended to support the development of a shared conceptual framework for CEL-related terminology and constructs.

3.1.2 Connection between Perspectives Paper and Systematic Review

With the findings and challenges outlined in the first literature review (perspectives paper) at the forefront, the purpose of the second literature review [11] (a systematic literature review) was to build on the previous work by providing a richer and more comprehensive analysis of existing CEL research within ISR over the past 50 years (i.e., from the start of the literature).

To this end, the systematic literature review can be considered an extension on the perspectives paper. The systematic review examined a much wider data set of 91 papers - it is important to note that 26 of these papers were already identified and analysed in the perspectives paper. As the categories of analysis were broader in scope than the initial review, a much wider data set was analysed which in turn provided new insights which were not previously discussed or examined. These new insights further led to a revised version of the working definition framework and respective diagram previously included in the first review.

Key differences between the perspectives paper and systematic review were as follows:

- The systematic review followed a rigorous and systematic literature analysis. The inclusion/exclusion criteria for included articles was expanded in the systematic review to include a wider range of articles.
- The perspectives paper includes 26 papers in the analysis, whereas the systematic review includes 91.

- The number of sources (i.e., conferences, journals) included in the literature search was also expanded from nine sources in the perspectives paper to 26 sources in the systematic review.
- The perspectives paper only examined articles published within a 20-year period (2000-2020), whereas the systematic review searched articles from the beginning of the literature (50 year period).
- The categories of analysis were much broader in scope in the systematic review than the perspectives paper.
- As the categories of analysis were much broader in scope for the systematic review, the results section was significantly more extensive than the perspectives paper. New insights were discovered that were not previously discussed or examined.
- While the perspectives paper guided the initial working definition framework, findings from the systematic review led to revisions of the framework and its diagram to incorporate new insights - namely the identification and inclusion of "physical" internal resources and its respective attributes.

The remainder of this chapter will focus on the findings of the **systematic literature review** with some reference to the initial literature review. Therefore, the work presented in this chapter is primarily based on the previously published article entitled, *A Systematic Review of Cost, Effort, and Load Research in Information Search and Retrieval, 1972-2020* [11].

3.2 Methodology

The primary objective of the systematic review was two-fold: firstly, to collect, present, and synthesise existing ISR research that address the measurement of CEL, and secondly, to underscore the current challenges associated with defining CEL within ISR. These objectives help to fulfill the two main aims of the theoretical contribution of this thesis which are: (1) to establish more precise definitions of CEL and (2) to offer a

critical assessment of current measures and their prospective applications within ISR research.

A systematic review methodology was adopted to manage the large and diverse body of literature relevant to CEL. This approach enabled a rigorous and transparent synthesis of existing research, allowing for the identification of patterns, inconsistencies, and conceptual overlaps across studies [56]. Systematic reviews are particularly valuable for clarifying definitions, evaluating methodological trends, and developing integrative frameworks, goals which aligned directly with the theoretical objectives of this thesis. Moreover, by applying explicit and replicable procedures, the systematic review helped reduce bias and improve the reliability of the resulting conceptual framework [56].

The process of conducting the systematic review encompassed five crucial steps: (1) identifying research questions; (ii) establishing sources containing potentially relevant articles for selection; (iii) creating a search strategy and determining search terms/keywords; (iv) defining a set of inclusion/exclusion criteria; (v) establishing categories for coding and analysis. Each of these steps will be elaborated upon in subsequent sections for a more comprehensive understanding.

3.2.1 Stage 1: Research Questions

The research questions formulated for the systematic review were derived from the high-level research questions outlined in Section 1.3. Subsequently, sub-questions were incorporated to attain a comprehensive and in-depth understanding of how CEL are currently being defined, measured, and the interconnection between these within ISR. To this end, the systematic review sought to answer the following research questions:

[RQ1] How have CEL and their related constructs been defined within ISR? Specifically,

(a) How are ISR researchers using these constructs?

(b) Are there any similarities/differences?

(c) How are ISR researchers using CEL related constructs such as resource and demand?

(d) To what extent are the definitions used informed by existing theory?

[RQ2] Which methods have been used/proposed to measure CEL within ISR? Specifically,

(a) Which methods to measure CEL are used/proposed within ISR?

(b) What are the investigated dependent variables for CEL measurement within ISR?

(c) What are the investigated independent variables for CEL measurement within ISR?

[RQ3] What are the relationships between the different definitions of CEL, and the methods used to measure them? Specifically,

(a) Are there similarities/differences between the construct measured and the methods used to measure it?

(b) How do CEL conceptual categories align with methods and their unit of analysis?

3.2.2 Stage 2: Sources

The second stage involved identifying sources from which articles measuring CEL in the context of ISR could be selected. The initial set of sources was drawn from Kelly and Sugimoto's systematic review of Interactive Information Retrieval (IIR) evaluation studies [57], which provided a well-established foundation of 31 relevant publications. This list had been previously validated by four IIR experts, lending it credibility and relevance for further analysis. From this set, 17 sources were selected as appropriate for the specific focus of this systematic review. To ensure the source list was comprehensive and up-to-date, further consultations were conducted with two ISR experts - Martin Halvey and Leif Azzopardi (both from the University of Strathclyde, UK). Based on their recommendations, nine additional sources were identified and incorporated, resulting in a more robust and representative set of materials for the review. Overall, 26 sources were selected (9 journals; 16 conferences; and one workshop). In addition to

structured database searches (refer to table 3.1), supplementary manual search methods were employed to ensure comprehensive coverage of relevant studies. These included both chaining techniques, "backward" chaining (examining reference lists) and "forward" chaining (identifying later works that cited a given article) as recommended by the Cochrane Handbook [58, 59], as well as targeted Boolean search strategies within databases. While there is no universally standardised protocol for manual searching, combining these methods helped identify studies that may not have been captured through initial database queries alone. This process led to the identification of 26 additional articles from 16 different sources (refer to table 3.2).

Table 3.1: Source and publications examined in database search (T = title only search; A = abstract only search; T-A = title & abstract search; F-T = full text search)

Source Title	# Papers Retrieved	# Papers Included for Analysis
Journals		
Information Processing & Management (IP&M)	113 (T-A)	5
Journal of the Association for Information Science & Technology (JASIS&T)	5 (T-A)	1
International Journal on Digital Libraries	0 (T)	0
International Journal of Human-Computer Studies (IJCCI)	15 (T-A)	2
Information Retrieval Journal (IRJ)	0 (T)	0
Journal of Information Science	41 (A)	1
Journal of Documentation	100 (F-T)	2
ACM Transactions on Computer-Human Interaction (TOCHI)	5 (A)	1
ACM Transactions on Information Systems (TOIS)	35 (A)	2
Conferences/Workshops		
ACM/IEEE Joint Conference on Digital Libraries (JCDL)	71 (A)	3
European Conference on Digital Libraries (ECDL)	28 (F-T)	2
European Conference on Information Retrieval (ECIR)	17 (F-T)	1
ACM International Conference on Information and Knowledge Management (CIKM)	248 (A)	4
ACM International Conference on Intelligent User Interfaces (IUI)	14 (A)	0
Proceedings of the Association of Information Science & Technology (ASIS&T)	7 (T-A)	5
ACM Special Interest Group on Information Retrieval Conference (SIGIR)	169 (A)	12
ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR)	32 (A)	10
Information Interaction in Context Conference (IIiX)	11 (A)	5
ACM International Conference on Web Search & Data Mining (WSDM)	57 (A)	2
Conference on Human Factors in Computing Systems (CHI)	79 (A)	2
International Conference on the Theory of Information Retrieval (ICTIR)	23 (A)	3
ACM Conference on Recommender Systems Conference (RecSys)	5 (A)	0
The Australasian Document Computing Symposium (ADCS)	6 (A)	0
The Asia Information Retrieval Societies Conference (AIRS)	1 (F-T)	0
Conference on Human-Computer Information Retrieval (HCIR)	2 (A)	0
European Workshop on Human-Computer Interaction (EuroHCIR)	47 (F-T)	2

3.2.3 Stage 3: Search Strategy

The next stage of the review process involved the development of keyword search terms to identify papers in the database search. The search term ***(effort OR cost* OR "mental workload" OR "cognitive load" OR workload) AND (search* OR "information retrieval" OR"information seeking")*** was employed for all of the

Table 3.2: Source and publications examined in manual search (backwards & forward chaining)

Source Title	# Papers Included
Journals	
Journal of Management Information Systems Quarterly	1
Journal of the American Society for Information Science & Technology (JASIST)	7
Information Research	1
International Journal of Industrial Ergonomics	1
Sensors (MDPI, peer-reviewed open-access journal)	1
Journal of Advances in Human-Computer Interaction	1
Journal of Innovation in Health Informatics	1
Conferences	
Proceedings of the International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)	1
Proceedings of the ACM Conference on Human Information Interaction & Retrieval (CHIIR)	1
Proceedings of the Information Interaction in Context Symposium (IIiX)	1
Proceedings of the Association for Information Science & Technology (ASIS&T)	5
Proceedings of the International Conference on Multimedia, Interaction, Design & Innovation (MIDI)	1
Proceedings of the ACM International Conference on Digital Libraries	1
Proceedings of the European Conference on Information Retrieval (ECIR)	1
Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval	2
Proceedings of the Human-Computer Interaction & Information Retrieval (HCIR)	1

search databases beside the following three which did not facilitate search terms with truncation symbols: the Journal of Information Processing & Management (IP&M); the International Journal of Human-Computer Studies (IJCCI); and the European Workshop on Human-Computer Interaction (EuroHCIR). For those databases the following search term was used: ***(effort OR cost OR "mental workload" OR "cognitive load" OR workload) AND (search OR "information retrieval" OR "information seeking")***. For the EuroHCIR database, Boolean search was not possible and therefore the database had to be manually searched. It is also important to note that while it was preferred to refine the search to "title" and "abstract" search, this was not always possible for some databases. Therefore, in some cases "full-text" searches were used to avoid missing relevant articles. Overall, the database search yielded 1,083 articles.

3.2.4 Stage 4: Inclusion/Exclusion Criteria

The purpose of this stage was to define and assess inclusion and exclusion criteria that would be used to methodically accept or reject articles for review. Note that the development of these criteria took place *before* the literature search. The set of criteria were aligned with the intended article type: the measurement of CEL within ISR. Empirical studies (i.e., employing experimental methods) were required to measure at

least one CEL construct within an ISR context. Non-empirical articles, such as review articles, were expected to propose or describe at least one CEL construct in an ISR context. To ensure the retrieval of higher-quality articles, it was imperative that the selected articles were scholarly publications and retrieved from peer-reviewed venues (including journals, conference proceedings, and workshops). The articles could be full-length or short papers - providing they were written in English. A publication date time-frame was not established as the intention was to retrieve as many relevant articles as possible. Refer to Table 3.3 for an overview of the inclusion/exclusion criteria.

Table 3.3: List of inclusion criteria/exclusion criteria

Inclusion/Exclusion Criteria
For empirical studies, the article should measure at least one CEL construct in an ISR context.
For non-empirical studies, the article must propose/describe measures for at least one CEL construct in an ISR context.
CEL must be measured/described from a user-sided perspective
The article should be a full-length research article, short paper, or equivalent.
The article should be peer-reviewed and published in either a journal, workshop or conference proceedings.
Time frame: from the start of the literature
Articles must be written in English

Using these inclusion/exclusion criteria, the title, abstract, and results sections of the 1,083 articles retrieved in the database search were screened. Although it is customary in a systematic review to screen only the title and abstract, the results section was also included in this screening process due to the emphasis on methods. After the initial screening, 67 articles were identified as meeting the inclusion criteria. Each article was then read in its entirety to conduct a more comprehensive assessment of eligibility. At this stage, two articles were excluded, leaving 65 articles for analysis. Leveraging these 65 articles a source, an additional 26 articles were obtained through manual searches ("backward chaining", N=16; "forward chaining", N=10). Overall, 91 articles were included in the final analysis.

3.2.5 Stage 5: Categories for Analysis

The categories for coding and analysis were formulated in accordance with the key research questions, and were as follows:

RQ1: How have CEL and their related constructs been defined within ISR?

- Studies that include/do not include a definition of CEL.
- Similarities/differences between the conceptual categories associated with CEL.
- Different terminology applied to CEL.
- Studies that include/ do not include a definition of resources, demand, and the similarities/differences between these definitions.
- Studies that reference theory, i.e. Multiple Resource Theory or Cognitive Load Theory.

RQ2: Which methods have been used/proposed to measure CEL within ISR?

- Studies that use only objective methods or subjective methods.
- Studies that use both objective and subjective methods.
- Dependent variables used in objective measurement.
- Questions used in self-designed questionnaires.
- Independent variables used.

RQ3: What are the relationships between the different definitions of CEL, and the methods used to measure them?

- Similarities/differences between the conceptual categories associated with CEL and the methods used to measure these constructs.
- Studies which use/propose the same methods to measure different constructs.

3.3 Results and Discussion

3.3.1 Defining Cost, Effort, Load (CEL) and Related Constructs in Information Seeking and Retrieval (ISR)

This section presents the findings of the systematic review in relation to the first research question (**How have CEL, and their related constructs been defined**

within ISR?).

While 91 articles were reviewed, only 38 provided definitions of the construct they were examining. To identify similarities and differences between the definitions, key elements from each definition was extracted, categorised, and quantified based on previous work by Martinic and colleagues [55]. Through this analysis, each definition was associated with one of the five conceptual categories outlined in the previous literature review [10]. Refer to Table 3.4 for an overview of each construct, their associated conceptual categories, and the corresponding articles.

Conceptual Category	Construct & Number of Articles	Citation
Time-Orientated	Cost ($N = 7$) Effort ($N = 2$)	[1, 4, 13, 16, 21, 60, 61] [14, 62]
Interaction Orientated/Count-Based	Cost ($N = 4$) Effort ($N = 7$)	[16, 21, 60, 63] [3, 64–68]
Cumulative/Total Work	Effort ($N = 8$) Cognitive Load ($N = 2$) Workload (N = 4)	[5, 62, 69–74] [6, 75] [71, 76–78]
Meta-Cognition/ Conscious Awareness	Effort ($N = 3$) Workload ($N = 1$)	[4, 70, 79] [2]
Capacity Based/ Resource Bound	Effort ($N = 1$) Cognitive Load ($N = 3$)	[80] [26, 81, 82]

Table 3.4: CEL construct and conceptual categories

Cost

While sixteen articles examined cost, only eight provided definitions and from these two conceptual categories emerged:

- *Time Orientated:* Seven articles defined cost in relation to the time spent during the user-system interaction [1,4,13,16,21,60,61]. For instance, *"cost is often measured by the time of a series of actions, such as formulating queries, examining snippets, clicks on results etc"* [4]; and *"cost is often considered as the amount of time spent"* [16].
- *Interaction Orientated/Count Based:* Four articles defined cost as the number of actions/interactions between the user and the system [16, 21, 60, 63]. For in-

stance, *"consider the cost of information search results from query generation, documentation examination, search engine result pages and task description examinations"* [63].

Effort

Of the 54 articles which examined effort, only 20 provided definitions. These definitions were found to align with all five conceptual categories:

- *Cumulative/Total Work:* Eight studies [5, 62, 69–74] defined effort in relation to the cumulative or total amount of physical/mental work that the individual applies towards an outcome. For instance, *"the total work done to achieve a particular goal"* [83], and *"how much work must an assessor exert"* [73].

- *Interaction Orientated/Count Based:* Similar to definitions of cost, seven studies [3, 64–68] defined effort as the interaction or number of actions which occur between the system and the user. For instance, *"we assume that users perform actions to make progress on a search task; every action costs effort"* [66] and *"counts of actions that require a cognitive assessment (e.g., evaluate a SERP) or production (e.g., enter a query) from a searcher"* [68].

- *Meta-cognition/Conscious Awareness:* Three articles [4, 70, 79] defined effort as a volitional and intentional process in which the individual is consciously aware. For example, *"effort reflects a voluntary allocation of effort that can be reported by the individual"* [79].

- *Time-Orientated:* Again, sharing similarities with cost, two studies [14, 62] referred to effort in relation to the time spent during the user-system interaction. For example, *"a natural candidate for measuring user effort is time; the longer it takes the user to reach an answer, the more effort is expended on the user's part"* [14].

- *Capacity Based/Resource Bound:* Only one study [80] described effort in relation to the notion of limited capacity and mental resources. For example, *"study of*

mental demands and effort can involve an assessment of users' mental load, a control of mental demands imposed by a task, by a system, or characterization of users by their levels of mental capacity" [80].

Cognitive Load

Of the nine articles which examined cognitive load, six definitions were extracted. Definitions of cognitive load aligned with two conceptual categories:

- *Capacity Based/Resource Bound:* Similar to definitions of effort, three studies [26, 81, 82] defined cognitive load in relation to the limited mental capacity and resources. For instance, *"cognitive load is closely related to the notion of limited mental resources"* [26] and *"people's cognitive capacities are so limited that they can process only limited information chunks concurrently"* [82].
- *Cumulative/Total Work:* Two articles described cognitive load [6,75] in alignment with the "work" related elements also identified in the characterisation of effort. For instance, *"cognitive load is usually evaluated according to the quantity of information to be memorised and the amount of processes involved to perform the task"* [6].

Workload

While 22 articles examined workload, only five definitions were provided. These definitions aligned with two conceptual categories:

- *Cumulative/Total Work:* Similar to definitions of effort, four articles [71, 76–78] describe workload in relation to the cumulative or total amount of physical/mental work. For instance, *"can be intuitively defined as the amount of cognitive work necessary for a person to complete a task over time"* [77].
- *Meta-cognition/Conscious Awareness:* Also similar to definitions of effort, one article [2] defined workload as a volitional and intentional process in which the individual is consciously aware. For instance, *"mental workload represents the subjective experience of a decision maker"* [2].

Related Constructs

To gain a broader understanding of how CEL constructs are being defined within ISR, the systematic review also investigated related constructs such as **demand** and **resources**. Definitions of these constructs were also organised into conceptual categories, outlined below.

Demand: Definitions of demand were provided in eleven articles measuring effort [64, 68, 79, 80], cognitive load [26, 75, 82] and workload [2, 76–78]. The following two categories emerged from these definitions:

- **Task demand:** Seven articles mentioned the demands arising from the search task itself, encompassing various task characteristics such as: difficulty [77]; time pressure [77]; complexity [82]; concurrent task requirements [76].
- **System demand:** Five articles discussed the demands arising from the information system, encompassing aspects such as the interface and information displays [26]; sensory modality [77]; interruptions [77].

Resources: Definitions of resources were provided in fifteen articles measuring cost [4]; effort [64,68,70,71,79,80], cognitive load [6,26,29,75,82], and workload [2,16,71,77]. The following two categories emerged from these definitions:

- **Cognitive resources:** Eleven articles described the cognitive resources (i.e., internal resources) that the user has available to them in order to process contextual stimuli, such resources include: working memory [2,15,26,29,68]; perception [16,77]; attention [16,70,82] , and motor control [16].
- **Limited capacity of resources:** Eight articles referred to the limited capacity of resources, inferring that there is a finite amount of resource that can be distributed to a task/system demand at any given time - this was often conceptualised as a "constraint" or "limitation" [16,26,64,68].

Theory: In total, seven articles referred to existing theory to justify their research - two key theories emerged as follows:

- Cognitive Load Theory (effort: $N = 1$; cognitive load: $N = 5$)
- Multiple Resource Theory (cognitive load: $N = 1$; workload:$N = 1$)

3.3.2 Measuring Cost, Effort, and Load (CEL) in Information Seeking and Retrieval (ISR)

This section presents the findings of the systematic review in relation to the second research question (**Which methods have been used/proposed to measure CEL within ISR?**).

The systematic review revealed that ISR studies employ a variety of objective and subjective methods to examine CEL. Objective methods can be considered as those which are impartial and yield a quantifiable outcome, whereas subjective methods require a human judgement of some type (refer to section 2.3 for a more detailed overview of objective and subjective measures). Note that this could also include a subjective judgement about an objective search behaviour, for instance, "*how many queries did you issue during your search?*".

Objective methods:

Out of the 91 articles reviewed, 59 articles employed or suggested objective methods to measure CEL. Table 3.5 outlines each objective method and the associated construct measured. The table highlights that among the empirical studies (i.e., conducted an experiment), search interaction logs stood out as the most frequently used method, with 48 articles incorporating at least one search interaction measure. In nine of these studies, search interaction logs were complemented by another objective method, such as dual-task and eye-tracking. Other objective methods were employed less frequently, with only twelve articles utilising eye-tracking, eight using dual-task, and one article employing a range of physiological measures such as electroencephalogram (EEG), temperature, electrocardiogram (ECG), and electro-dermal activity (EDA). For articles which were non-empirical, but instead *proposed* methods, seven articles proposed the use of search interaction logs, and the two remaining articles proposed the use of two

physiological methods: functional near-infrared spectroscopy (fNIRS) and electroencephalogram (EEG).

Objective Method Proposed/Used	Construct & Number of Articles	Citation
Search Interaction Logs	Cost ($N = 14$)	[1, 4, 13, 15, 16, 21, 60, 61, 63, 84–88]
	Effort ($N = 38$)	[5, 14, 24, 27, 69, 70, 72, 79, 83, 89–94] [3, 4, 8, 9, 25, 62, 64–68, 73, 80, 92, 95–102]
	Cognitive Load ($N = 2$)	[75, 82]
Eye Tracking	Cost ($N = 1$)	[63]
	Effort ($N = 9$)	[3, 27, 60, 64, 68, 80, 103–105]
	Cognitive Load ($N = 1$)	[82]
	Workload ($N = 3$)	[71, 78, 106]
Dual Task	Effort ($N = 3$)	[70, 79, 97]
	Cognitive Load ($N = 4$)	[6, 26, 75, 107]
	Workload ($N = 1$)	[108]
Other Physiological (EEG, Temperature, ECG, EDA, fNIR)	Cognitive Load ($N = 1$)	[81]
	Workload ($N = 2$)	[78, 109]

Table 3.5: Objective methods used/proposed to measure CEL in reviewed articles

To summarise, four objective methods used to measure CEL were identified in the ISR literature. Note that each of these measures are associated with specific dependent variables (i.e., the variable used to measure the construct). For instance, as we can see from Table 3.6 that CEL have been measured by 23 different dependent variables associated with search interaction logs, and these have been grouped into three measurement categories: *total interaction counts; rates of interaction; and time-based.*

Subjective Methods

As observed in Table 3.7, four different types of subjective methods were used to measure CEL in 39 articles. The NASA-Task Load Index was employed in 22 articles, and can therefore be considered as the most widely used subjective method of measuring of CEL within ISR. It was noted however, that the administration and analysis of the NASA-TLX differed between articles. For example, only eleven articles used the full

Objective Method Proposed/Used	Dependent Variable & Number of Articles	Cost	Effort	Load
Search Interaction Logs	**Total Interaction Counts:**			
	Total number of:			
	Documents viewed/browsed/read/opened (N = 28)	X	X	
	Queries issued (N = 18)	X	X	
	Clicks & scrolls (N = 10)	X	X	
	Query reformulations/iterations/refinement (N = 9)		X	
	Book*marks (N = 7)*	X	X	
	SE*RPs clicked/visited/viewed (N = 6)*	X	X	
	Unique *search terms issued (N = 4)*		X	
	R*elevant documents browsed/marked as relevant (N = 3)*		X	
	Queries without a bookmark (N = 2)		X	
	Interaction Rate:			
	Number of:			
	Clicks: per *query (N = 2); per snippet (N = 2); per document; without a bookmark (N = 2)*	X	X	
	Words *per query (N = 6)*	X		
	Time-Based:			
	Dwell time (N = 5)	X	X	
	Time taken to:			
	Complete task/session (N = 16)	X	X	X
	View/examine search results (N = 6)	X	X	
	Formulate first query (N = 4)	X	X	
	Read/assess/judge documents (N = 6)	X	X	
	Enter a query (N = 2)	X	X	
	Average time:			
	per *click (N = 2)*		X	
	per search action (N = 2)		X	
	between queries & clicks (N = 2)		X	
Eye Tracking	*Fixation duration (N = 10)*	X	X	X
	Number of eye fixations (incl. fixations on documents; SERPs; task descriptions) (N = 10)	X	X	X
	Pupil Size*/Diameter/Dilations (N = 4)*		X	X
	Perceptual span (N = 3)		X	
	Length of saccade (N = 2)		X	
Dual Task	Re*action time (N = 8)*		X	X
	Miss *frequency (N = 3)*		X	X
	Accuracy (N = 2)		X	
Other Physiological (EEG, Temperature, ECG, EDA, fNIR)	*EDA: electric resistance of the skin (N = 1)*			X
	ECG: electric activity generated by the heart (N = 1)			X
	PPG: blood volume changes (N = 1)			X
	Temperature: fluctuations in body temperature (N = 1)			X
	EEG: electrical activity in the brain (N = 1)			X
	fNIR: detects hemo-globin changes in the brain (N = 1)			X

Table 3.6: Dependent variables used/proposed to measure CEL in reviewed articles

Subjective Method Proposed/Used	Total Number of Articles	CEL Construct Measured & Number of Articles	Citation
NASA Task Load Index (NASA-TLX)	22	Cost ($N = 1$)	[87]
		Effort ($N = 4$)	[28, 73, 110, 111]
		Cognitive Load ($N = 2$)	[75, 112]
		Workload ($N = 17$)	[7, 9, 29, 87, 102, 113–116] [1, 2, 76, 77, 102, 108, 117, 118]
Self-Designed Questionnaire	14	Cost ($N = 2$)	[4, 85]
		Effort ($N = 11$)	[119–123] [65, 70, 79, 96, 124, 125]
		Workload ($N = 2$)	[123, 126]
Workload Profile (WP)	1	Workload ($N = 1$)	[77]
Mental Workload Test (MWT)	1	Workload ($N = 1$)	[71]

Table 3.7: Subjective methods used to measure CEL in reviewed articles

six-component version (physical demand; mental demand; temporal demand; performance; frustration; and effort). Additionally, the majority of studies (64%) used the NASA-TLX to assess workload across search systems, with the remaining studies (36%) using the tool to measure workload across search tasks. Self-designed questionnaires were utilised in fourteen articles, as depicted in Table 3.8, the questionnaires differ in relation to their format, scale, and unit of analysis. Finally, the Workload Profile (WP) and the Mental Workload test (MWT) were employed in two studies.

Objective and Subjective Methods

Ten articles employed a mixed methods approach, utilising a combination of both objective and subjective methods within a single study. As shown in Table 3.9, the majority of these studies used two different methods, mainly self-designed questionnaires and search interaction logs.

Independent Variables

A variety of different independent variables were examined in relation to their influence on CEL within ISR. Of the 91 articles included in the review, 121 independent variables were identified and categorised into five groups: task characteristics; search engine

Construct	Questions	Unit of Analysis
Effort		
	[119] Eight search behaviour questions related to use of search features; advanced search features; query terms entered; and frequency of web search engine use.	Total scores of user effort from 1(low) - 17 (high)
	[120] One question: "How much effort did it take to complete the task?"	Scale from 1 (very little effort) to 7 (a lot of effort)
	[121] Two questions: (1) Search Result Judgement effort: "How much effort did you spend on this web page?" (2) Session effort: "How much effort did this task take?"	(1) Scale from 1 (none) to 7 (a lot) (2) Scale from 1 (minimum) to 7 (a lot of)
	[121] Two questions: (1) Session effort: "How much effort did this task take?" (2) Post-click result judgement effort: "How much effort did you spend on this webpage?"	For both questions - Scale from 1 (minimum/none) to 7 (a lot of)
	[125] One question: "Rate your effort to answer this question well"	Scale from 1 (low) to 5 (high)
	[123] One question:"How much mental effort you used to complete the task"	Scale from 1 (low) to 5 (high)
	[124] Five questions related to search behaviour and difficulty: number of sessions; number of sources consulted; difficulty in selecting useful references in search for essay; number of read but not cited articles; number of channels used	Difficulty: Scale from 1 (very difficult) to 5 (very easy)
	[70, 79] One question: "Rate your effort invested in searching"	Scale from 0 (no effort) to 10 (a great deal of effort)
	[4] Perceived time estimation	Perceived time (s) examining each document
	[96] One question: "Did you put in a lot of effort to complete the task?" (after each task)	Scale (range not specified) - however, low = not much effort.
Cost		
	[85] Three questions relating to: Ease (type of source referred to); Time (self-reported time to complete the task); and Number of Sources (total number of sources consulted)	Ease: Low, Medium, High (depending on the type of source) Time: Low (<30mins); Medium (30-90mins); High (>90mins) Number of Sources: Low (<1); Medium (2-4); High (>4)
	[4] One question: "Estimate the duration spent on searching" (after each task)	Perceived dwell time (s)

Table 3.8: CEL constructs measured and the questions used in self-designed questionnaires

	Self- Designed Questionnaire	NASA-TLX	Mental Workload Test	Search Interaction Logs	Dual-Task	Eye-Tracking
Cost [4, 85]	X			X		
Effort [70, 79]	X			X	X	
Effort [73]		X		X		
Effort [4, 96]	X			X		
Workload [71]			X			X
Workload [108]		X			X	
Cognitive-Load [75]		X		X	X	

Table 3.9: Studies which use both subjective & objective measures of CEL

results page (SERP); the system; individual differences; and the document/webpage. The majority of studies (35%) manipulated some kind of task characteristic, such characteristics included the tasks: *structure; product; determinability; goal; complexity; and difficulty.* The second most widely examined independent variable (31%) was the search engine results page, or the query interface, which was manipulated in a range of different ways, such as adaptation to: the results list; SERP presentation; query features; and result quality. In relation to examining the effects of the system on user CEL, the majority of studies compared two independent systems (i.e., system A vs. system B). The examination of user characteristics on user CEL were less frequent within the reviewed articles. However, for those which did, working memory was the most frequently examined cognitive ability. Other cognitive abilities included associative memory; perceptual speed; verbal closure; inhibition; visualisation ability; and spatial ability. Other user characteristics included typing speed; domain knowledge; and level of user experience. Finally, the least examined independent variable was the document/web-page/landing page, adaptations included: the level of document relevance; and the visual complexity of the document.

3.3.3 Relationships between Constructs and Measures

This section presents the results from the systematic review which address the third key research question; **What are the relationships between the different definitions**

of CEL and the methods used to measure them?. Figure 3.1 illustrates the complex and interconnected relationship between CEL constructs and their respective measures. Table 3.10 further illustrates these relationships and the overlap between constructs and their measurable conceptual elements. These relationships are discussed in more detail below.

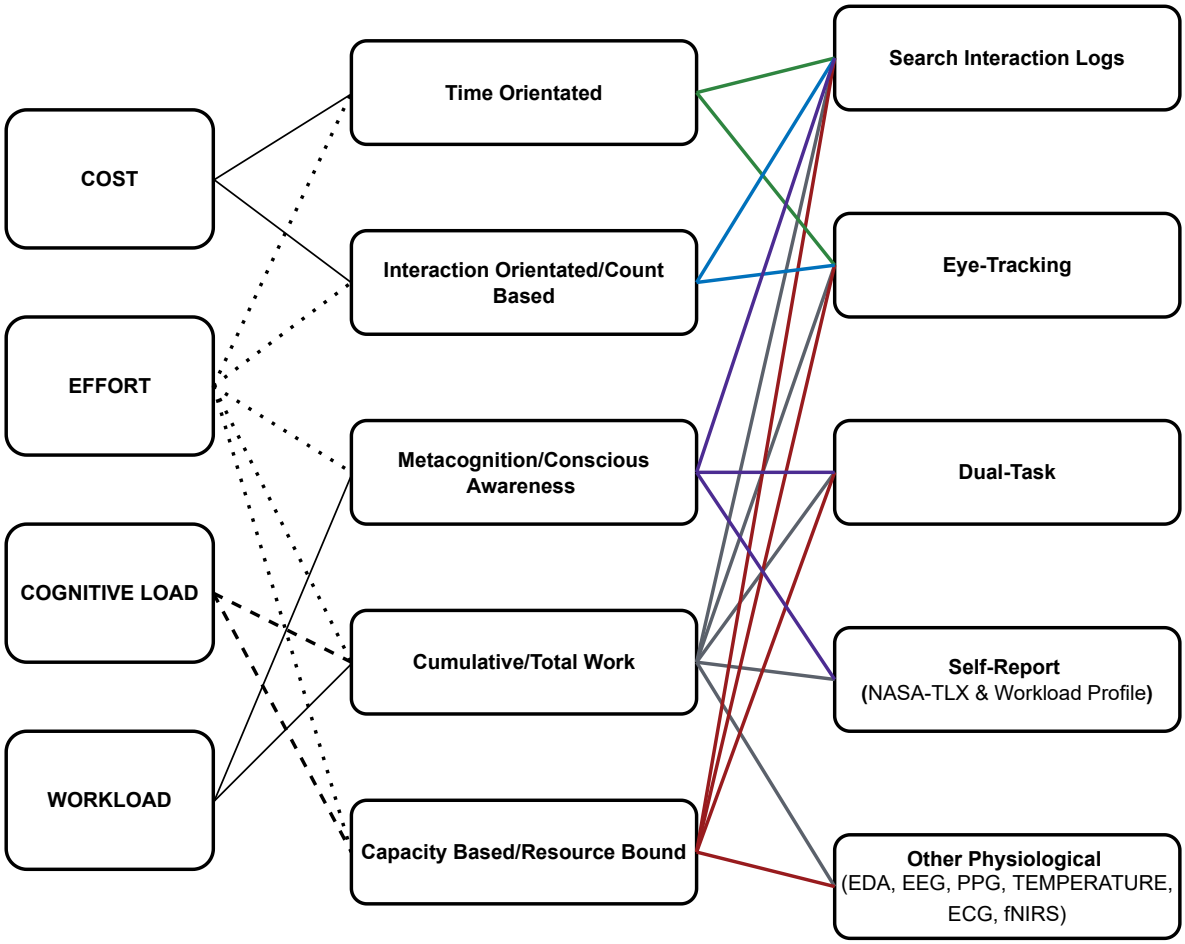

Figure 3.1: Relationship between constructs and measures

Construct	Conceptual Category	Measure	Dependent Variable(s)
Cost	Time-Orientated	Search Interaction Logs	*Time taken to*: issue a query; enter a word; examine a snippet/query/suggestions/SERP; make a relevance judgement; & dwell time. *Total number of*: clicks.
		Eye-Tracking	*Number of*: examined results; & length of examined result sequence.
	Interaction-Orientated/ Count Based	Search Interaction Logs	*Time taken to*: enter queries; examine SERP; & dwell time *Total number of*: clicks; words; queries; terms used; snippets hovered over; documents viewed; documents marked relevant.
		Eye-Tracking	*Number of*: examined results; fixations on documents/SERPs/task descriptions; & length of examined result sequence.
Effort	Interaction-Orientated/ Count Based	Search Interaction Logs	*Time taken to*: complete task (mins); enter a query; examine result snippets/clicked results. *Total number of:* mouse actions; interactions; queries issued/reformulated; result summaries/web-pages visited.
		Eye-Tracking	*Number of*: eye fixations/fixation regressions/saccadic movements. Fixation duration; length of saccades/reading sequences; pupil size.
	Cumulative/ Total Work	Search Interaction Logs	*Time taken to*: complete task (mins); view results; perform search (s/mins); make relevance judgement. *Total number/amount of*: commands; scrolling/navigation; documents read/opened; queries issued.
		Dual-Task	Reaction Time (ms); miss frequency.
	Meta-Cognition/ Conscious Awareness	Search Interaction Logs	*Time taken to*: complete search session (min/s); formulate first query; view search results; read documents; & dwell time. *Total number/amount of*: clicks; documents viewed/read.
		Dual-Task	Reaction time (ms/s); miss frequency.
		Self-Report	Perceived time length (s)
	Time-Orientated	Search Interaction Logs	*Time taken to*: task completion(s). Number of: queries issued per task.
	Capacity Based/ Resource Bound	Search Interaction Logs	*Total number/amount of*: query reformulations; visits to web-pages.
		Eye-Tracking	Fixation duration (s); Number of regression fixations.
Cognitive Load	Capacity Based/ Resource Bound	Dual-Task	Rate of missed events; reaction time (ms).
		Eye-Tracking	Fixation duration (ms); fixation count.
		Search Interaction Logs	Task completion time (ms).
		Physiological	EEG.
	Cumulative/ Total Work	Dual-Task	Average reaction time (ms).
Workload	Meta-Cognition/ Conscious Awareness	Self- Report	NASA-TLX.
	Cumulative/ Total Work	Eye-Tracking	Pupil diameter; eye fixation; fixation duration (ms); saccadic peak velocity (visual degree per second).
		Self-Report	NASA-TLX; Workload Profile.
		Physiological	EDA (average electric resistance of skin:kilohms); ECG (millivolts: mV); PPG (heart rate:average mV); Body temperature (degrees celcius); EEG (power & phase of analytical signal Hz).

Table 3.10: Relationships between constructs and measures

Cost

The assessment of cost typically mirrors its conceptualisation within ISR. Articles that define cost as *time-orientated* predominantly employ time-based metrics derived from search interaction logs. Even in the absence of clear definitions of cost, a similar pattern emerges, where time-based measures are commonly utilised. Analysing each study

and its respective measures of cost in isolation may suggest some internal validity, as it could be argued that time-based metrics clearly align with the time-orientated nature of cost. Nevertheless, despite the consistency of the "time" unit across studies, its operationalisation varies widely - for example, the time taken to: examine search results; issue a query; read a document. Consequently, making comparisons between these studies becomes challenging, as each implies a "fixed" value for cost. For instance, is the time taken to assess a document for relevance the same as the time taken to issue a query? Moreover, the articles reviewed would suggest that *time* is the only cost borne by users in the ISR process. It is plausible that a limitation of these articles lies in the narrow scope and limited generalisability of their research. Throughout the articles reviewed, the majority of empirical work into cost relied on laboratory studies involving university students as participants. In this context, cost is predominantly user-sided i.e., the time taken to complete actions and tasks. However, we also know from fields outside ISR, that costs can also be operationalised as monetary, financial costs or costs related to the expenditure of human resource. Translating this back to ISR, these additional types of cost for instance may be applicable in a professional search context where both the user and the organisation bear cost in terms of time, money, and human resources expended during the search process.

Effort

Following the synthesis of definitions, effort was represented by five conceptual categories. There could be various reasons for this, for instance it may reflect the authors varying perspectives of effort, the multi-faceted nature of the construct, or a consequence of the absence of a universally accepted definition. The diversity in effort definitions is reflected in how the construct has been operationalised within ISR - and in many cases, the operationalisation of effort does not align with how it has been conceptualised. For example, across all five conceptual categories, effort was operationalised using search interaction log metrics such as total interaction count and time-based variables. It can be argued however, that we cannot assume with certainty that these metrics are adequately capable of reflecting conceptual categories such as

Capacity-Based/Resource Bound which can only really be reflected by direct, cognitive measures. In articles which completely lack a definition, the array of measures used to operationalise effort become even more extensive. This prompts the question of *why* the conceptualisation and operationalisation of effort has faced so much diversity within the field. A closer examination of the provided definitions suggests that researchers often label individual components of effort rather than offering a holistic definition of the construct as a whole. For example, definitions such as "effort reflects a voluntary allocation of effort that can be reported by the individual" and "the total work done to achieve a particular goal" capture specific dimensions of effort but fall short of defining the entire construct. This issue, coupled with a general lack of clarity in effort conceptualisation across ISR and other domains, may contribute to the proliferation of variables operationalised as "effort". Consequently and particularly with the use of search interaction variables, the examination of effort within ISR has resulted in a "kitchen-sink" approach to measurement. However, rather than stemming from simple disagreement among researchers in relation to its meaning, it has arisen from a shared lack of common language to comprehensively define the construct.

Besides search interaction log variables, another widely used measure of effort was self-designed questionnaires. In general, there seemed to be a coherence in the use of subjective self-report measures across studies that defined effort within the context of the corresponding conceptual category, namely, *"Meta-Cognition/Conscious Awareness."*. However, issues became evident upon closer scrutiny of the self-designed questionnaires. As illustrated in Table 3.8, it becomes evident that there is a general lack of standardisation regarding the questions, scales, units of analysis, and format employed in these questionnaires. The variety of scales used for example, include a a seventeen-point scale (1-17), a seven-point scale (1-7), a five-point scale (1-5), and an eleven-point scale (0-10). The absence of standardised thresholds in these scales may imply different interpretations of effort levels; for instance, a rating of "5" might signify a high level of effort on one scale, while a rating of "17" could represent a high level of effort on another. This variability complicates comparisons between studies. Although these scales may assume face validity, the ability of individuals to accurately

assess their own cognitive capacity raises questions. Consequently, this can result in challenges when comparing individual ratings.

Cognitive Load

A limited number of articles offered a definition of cognitive load, and this observation may be tied to the sparse use of formal theory (i.e., Cognitive Load Theory) in which the principles of cognitive load is inherently rooted. The scarcity of explicit definitions restricts the ability to extract specific conceptual elements, which therefore constrains the inferences that can be drawn about the conceptual categories and their associated measures.

Articles which characterised cognitive load as *Capacity Based/Resource Bound* utilised or proposed measures such as dual-task, eye-tracking and electroencephalogram (EEG). Given that these methods are regarded as direct and objective, specifically designed to capture the dynamic and instantaneous aspects of cognitive load, they can be assumed to reasonably capture the *Capacity Based/Resource Bound* conceptual properties of the construct [127]. Moreover, these methods were also widely used in articles where a definition of cognitive load was not provided. Unlike the measurement of effort, this finding suggests a certain level of consensus among researchers regarding their understanding of cognitive load and awareness of appropriate measurement methods. However, for these measures to precisely reflect the conceptual properties of cognitive load, it is crucial that they undergo analysis at the appropriate level of granularity. In other words, to understand the intricate trends and patterns of cognitive load, it is imperative that the data is analysed at a fine-grain, discrete level. However, such detailed examination was not frequently conducted across the reviewed articles. For instance, in many cases where the dual-task paradigm was used, reaction time and missed event measures were averaged across the entire search process, rather than individual stages or interactions. Similar trends were noted for eye-tracking measures, like fixation duration and number of eye fixations - which were also frequently averaged across a search session. Moreover, when data is analysed at a higher level of granularity such as across a whole session, only a static, post-hoc snapshot of cognitive load can be achieved - potentially

obscuring the dynamic interplay between demand and user resource consumption. It is noteworthy, however, that in most instances where the article included a definition of cognitive load, the data was analysed a level of finer granularity, such as the task segment level. This further underscores the assertion that conceptualisation should precede operationalisation for effective measurement.

Akin to the absence of explicit definitions and reference to existing theory, the majority of articles exploring cognitive load did not explicitly reference or differentiate between the various load types (intrinsic, extraneous, and germane) outlined by Cognitive Load Theory (CLT), in their experimental design. This poses several problems. Without clearly defining the type of load being measured, it is difficult to identify which aspect of a task consumes users' cognitive resources. This becomes especially problematic when both task-related (intrinsic) and interface-related (extraneous) factors are manipulated within the same experiment. While some studies report "overall load" (assuming load is cumulative), failing to distinguish between load types risks conflated or misleading conclusions. The lack of reference to established theory and recognised load types further undermines the empirical basis of such manipulations.

Workload

Similar to cognitive load, the reviewed articles provided very few definitions of workload. Those which provided definitions which aligned with the *Cumulative/Total Work* conceptual category employed a range diverse methods, such as eye-tracking, self-report, and physiological measures. For those articles which did not provide a definition, there was a notable reliance on the NASA-TLX as the chosen measurement method. This trend is not exclusive to ISR; since its inception in the 1980s, the NASA-TLX has been the most widely utilised tool for workload assessment. Some argue that the NASA-TLX is so closely tied to workload that it has almost become synonymous with the construct [128] - perhaps explaining why most of the reviewed studies using the NASA-TLX did not offer a specific definition of workload. The prevalent use of the NASA-TLX has been endorsed by its founders, Hart and Staveland, as the most valid and sensitive indicator of workload [129]. While the credibility of the tool itself is less doubtful, var-

ious concerns have been raised regarding its application within ISR. For instance, the NASA-TLX is intended for workload comparison among tasks, yet the predominant use in the reviewed articles involved comparing workload among systems. This inappropriate application of the tool raises concerns about the validity of the assertions made in these studies. Additionally, the demands of the task are likely to vary from moment to moment during a search task [130], but most articles applied the NASA-TLX *after* completion of the search task. As this scenario relies on a post-hoc evaluation, the user must reflect on and amalgamate their task-related memories to form an overall score of their perceived demand - sometimes this can be reasonably precise, at other times less so. This integration process may be additionally complicated and disproportionately affected by the recollection of peak episodes or deviations in workload [130]. Consequently, the overall workload scores obtained retrospectively may not precisely mirror the users' experienced workload throughout the task. Lastly, it is worth questioning whether the NASA-TLX is really suitable for ISR-specific tasks? Originally intended for the Aviation industry, the NASA-TLX continues to be primarily employed as a workload measurement tool in high pressure operational settings like Air Traffic Control, Military, and Healthcare, where tasks involve flying, driving, surgical operations and so forth [131]. Given that ISR tasks typically involve lower demands than these examples, there is uncertainty regarding whether the tool is sufficiently sensitive to detect lower levels of workload. Additionally, the relevance of NASA-TLX items such as "physical demand" to ISR tasks remains uncertain.

3.4 Implications

The systematic review emphasises a collection of high-quality research articles that mostly all illustrate and substantiate the use of certain measures - exemplifying the potential strengths and limitations of each measure and serving as valuable insight for guiding future research. In the previous section, the relationships between constructs and measures were examined, shedding light on both the commonalities and variations among ISR researchers and potential concerns regarding the current examination of CEL.

The following section will discuss the implications of these issues in relation to the first high level research question outlined in Section 1.3 (**HL-RQ1: How has cost, effort, and load (CEL) been defined and measured within information seeking and retrieval (ISR)?**). Additionally, this section will also help set the foundation for the empirical work carried out for this thesis.

The outcomes of the systematic review seem to support an existing assertion that research within the ISR field is predominantly influenced by innovation and technology, rather than focusing on theory development or the enhancement of conceptual understanding [31]. This orientation toward practical aspects of science has prioritised results over theoretical explanation, resulting in numerous studies lacking solid theoretical foundations [31]. The review's findings indicate that a significant number of articles under scrutiny did not offer a clear definition of the CEL construct. Instead, intuitive understandings were relied upon, bypassing established definitions or existing theory. Consequently, in these cases, the measurement of CEL often lacked a theoretical basis, as researchers leaned on the face validity of the instrument - wherein the measure lacks formal validation but is deemed appropriate based on intuitive credibility and its popularity within the research community [31]. Unsurprisingly, the past 50 years of ISR have shown limited advancement or progress in characterising and measuring CEL. Universal definitions and standardised methods for CEL measurement are still elusive, and there is no singular method recommended as "gold standard" based on this review. To identify such "gold standard" methods for CEL measurement within ISR, it is imperative to first address effective measurement by fulfilling the crucial step of conceptualisation. Therefore, to advance understanding and measurement of CEL, it is essential for the ISR field to collectively focus on grounding CEL research in established theory.

Definitions

As outlined in Section 1.3, a key aim of the theoretical work carried out in this thesis was to establish working definitions and a conceptual framework for defining CEL as means to overcome some of the conceptual challenges highlighted in the systematic re-

view. In the first literature review [10], a provisional framework to define CEL and its associated constructs was developed. The intention of this framework was to provide the ISR community with a set of working definitions which other researchers could adopt and expand upon within their own research. This collaborative approach aims to cultivate a more cohesive understanding and examination of CEL constructs within ISR. Notably, these working definitions and the accompanying diagram have undergone updates based on fresh insights derived from the systematic review, particularly concerning the physical resources accessible to the user. The demand on users' physical resources during the search process gained increased prominence in light of this review. Unlike the previous framework that mainly focused on the user's internal resources related to cognition, such as working memory and perception, the current review highlighted search interactions like clicks, scrolling, and typing as frequent indicators of user effort. It was recognised that these actions not only draw upon cognitive resources but also engage the user's physical resources, including strength, motor action, and metabolic energy. Understanding the consumption of physical resources in the search process and its implications for user effort is crucial, and it underscores the importance of considering accessibility factors in future research.

Below we describe the working definitions framework extracted from the systematic review and show how these constructs are related in Figures 3.2 and 3.3.

Resources: In accordance with CLT and MRT, individuals possess multiple resources at their disposal. In the context of ISR, these resources can be broadly categorised as follows: (*i*) **Internal Resources:** These are associated with the user and can be either cognitive (e.g., working memory, attention) or physical (e.g., metabolic energy, strength), and; (*ii*) **External Resources** These are resources available to the user from the external environment and may include factors such as time, money, labour, and other external support.

Resource Capacity: Every resource operates within limited capacity, such as the number of items that can be stored in working memory or the time available for task completion. Importantly, these capacities are not fixed and may fluctu-

ate over time. For instance, users can enhance their working memory capacity through practice or training. Conversely, factors like stress or fatigue might diminish this capacity. Additionally, external circumstances can impact resource capacities; a deadline shift to an earlier date decreases available time, while an extension increases the time at hand for a task. These variations underscore the dynamic and adaptable nature of resource capacities.

Demand: Demands arise based on the characteristics of the task, system, and the context itself. Demand will regulate how much of the internal resources (cognitive/physical) need to be exerted or expended, and also direct how much of the external resources will need to be paid or spent to perform the task using the system in the given context. Demand is dynamic in nature and varies over the duration of the task.

Load: Aligned with psychological theory, cognitive load and workload in this thesis work will be classified as part of the overarching term "load". Considering a specific resource and the demands posed by the task, system, and context, the construct of load can be generalised from CLT and refers to the quantity of resources (internal or external) being consumed at any given moment.

Overload: Collectively, the construct of overload occurs when the demands imposed by the task, system, and context exceed the capacity of the available resource(s). For instance, if the working memory or attention required surpasses an individual's capacity, they are likely to encounter overload.

Effort: Within the context of ISR, effort is a user-centric construct that reflects the total amount of *internal* resources that are *exerted* or *expended*, over a given duration, to fulfill the demands of the task, system, and context. In Figure 3.3, the lower plot illustrates the relationship between effort and load, with effort representing the total load experienced over time (i.e. the area under the curve).

Cost: Cost is distinguished from effort by focusing on the resources they pertain to. Cost is examined in relation to *external* resources (such as money, time, human

resources etc.) that the user *spends* or *pays* to fulfill the demands of the task, system and context.

Within the context of ISR, the aforementioned definitions converge as follows: During an interactive search task, demands emerge from both the inherent characteristics of the task itself, such as task difficulty, and the features of the system, such as the layout of the search engine result page. The user possesses internal resources (cognitive and physical) to address these demands, such as retaining information in their working memory, or they may utilise external resources, such as seeking assistance from a colleague. If the demands become excessive, these resources will reach their limit, leading to user overload. In such instances, the user may witness a decline in performance or cease the search task altogether. To efficiently allocate resources throughout the task duration and achieve the task goal, the user must consciously engage in physical activities (such as typing a query) and cognitive processes (such as analysing a results page). The level of effort exerted is contingent on the amount of load experienced. As the user completes the search task, cost can be viewed as the external resources consumed or spent, such as the time invested in the task.

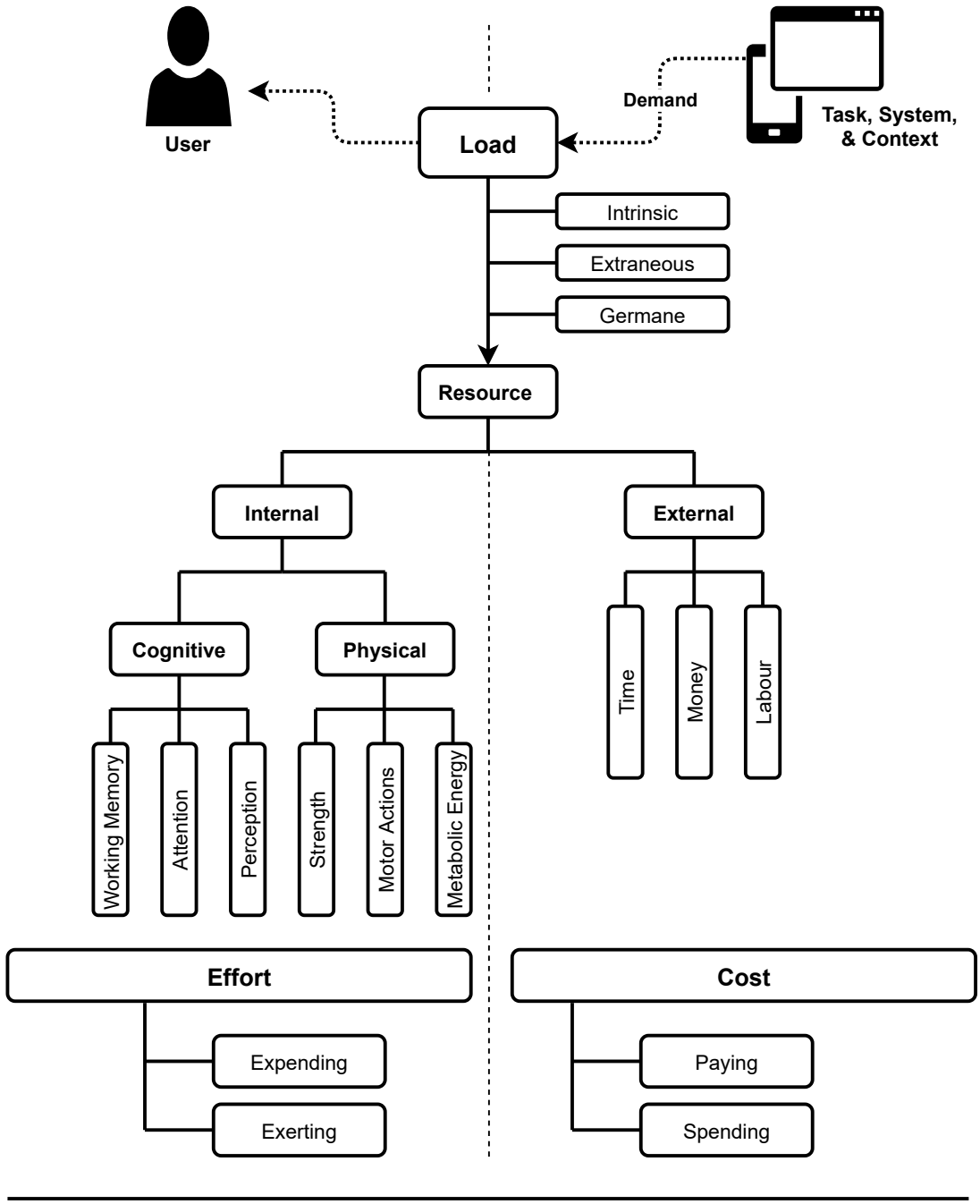

Figure 3.2: Relationships between CEL constructs

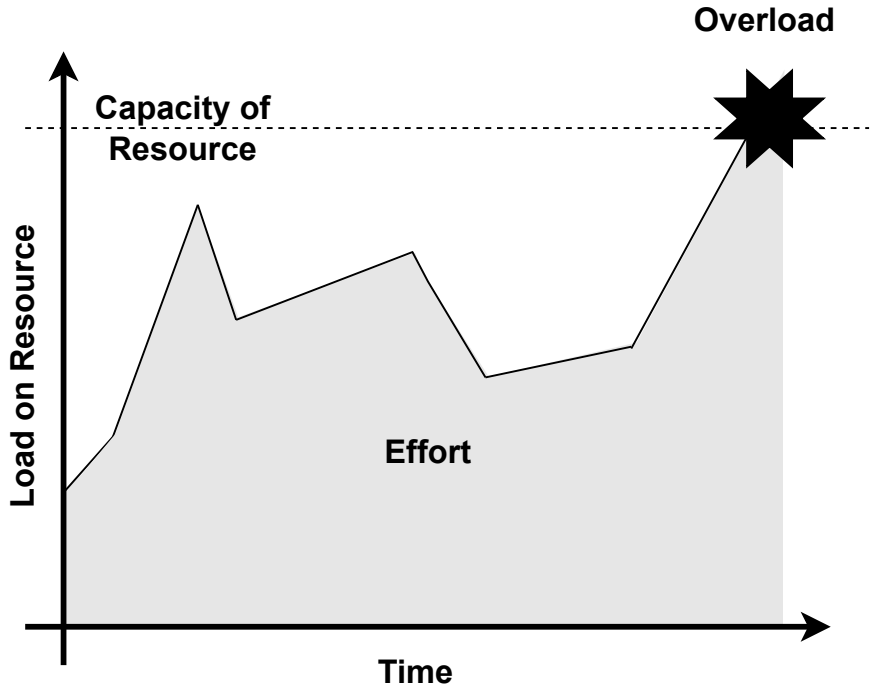

Figure 3.3: Relationship between load over time
A graphical depiction of the relationship of the load experienced by a user over time for a given internal resource. When the load demanded by the task, system and context exceeds the capacity of the user's resource, then they hit overload. The effort experienced by the user is the total load over time (i.e. the area under the curve).

Measures

The systematic review highlights a lack of innovation in the advancement of measures for CEL within ISR. For instance, the utilisation of search interaction logs as a means to measure cost and effort remains as prevalent today as it was four decades ago. Although search interaction log metrics provide a non-intrusive and relatively straightforward way to gauge cost and effort, their application has faced criticism. Notably, the use of time as an indicator of cognitive processing has been deemed "too simplistic", with arguments asserting that it fails to adequately account for the influence of other confounding variables, such as the nature of the task, the topic, and individual characteristics of the user [132]. The reliance on search interaction metrics presupposes that users will exhibit predetermined or stereotypical behaviours. However, given that each

user's interaction with a system and its information is distinct in terms of the physical, cognitive, and emotional aspects, can these fixed values truly capture the individual experiences of users? It has been posited that the meaningful value of search interaction signals can only be attained when the context, purpose, and nature of *what* is being examined is taken into consideration [31]. Despite this, ambiguity in relation to which construct is being measured will persist if the *same* metrics are being used to measure *different* constructs. For example, "number of clicks" have been used as measure of cost, effort, and in studies beyond the scope of this review, as a robust indicator of topical interest [133, 134]. Similarly, commonly used metrics like dwell time and time-on-task, traditionally associated with cost and effort, have also been employed as measures of user interest [133], user satisfaction [135], and relevance [136]. Given their prevalence and longstanding use, it is probable that search interaction logs will continue to be a prominent method for measuring cost and effort within ISR. Therefore, it is crucial that these measures are employed as accurately as possible. As discussed, relying solely on measures derived from search logs may not adequately capture and elucidate complex phenomena such as cost and effort. However, methodological triangulation presents a valuable technique to enhance the credibility and validity of research findings [30]. This approach generally advocates for the incorporation of multiple data collection methods within a research study to mitigate potential biases inherent in relying on a single method [30].

More broadly, effort and workload were the most frequently examined constructs across the reviewed studies. However, these constructs were seldom the central focal points of the study. Instead, it seemed that many studies incorporated measures of effort and workload, often in the form of self-report, as a "quick and dirty" approach to gauge users perceptions regarding task or interface demands. While self-report measures can offer insight into overall task demand, the subjective and usually post-task nature of these measures provides limited information about the particular elements or aspects of the task, system, or interface contributing to heightened effort or workload. In contrast, appropriately measured cognitive load has the potential to indicate specific instances when users encounter increased demand during the task. As heightened levels

of cognitive load can have adverse effects on both individuals and task performance, identifying the areas where this increased load occurs can greatly benefit the advancement of interface design and the development of tools to support users during tasks of higher complexity. Despite the valuable insights that measuring cognitive load can offer, this construct has been the least examined CEL construct within ISR, and has noticeably experienced a sharp decline since 2014 when workload examination began and has since dominated. While reasons for this are unclear, it may be that the complex characteristics of cognitive load, coupled with the absence of a "quick and dirty" measure, has diminished the appeal of examining this construct. Nonetheless, delving into constructs like cognitive load within ISR could offer additional insights. This is especially important in comprehending the specific elements of the search process that impose higher demand and assessing how this may influence system usability and user task performance.

3.5 Summary

Overall, the objective of the systematic review was to compile and critically examine how CEL, are currently defined and measured in articles across the field of ISR. A key outcome of this work has been the identification and analysis of commonalities among ISR researchers concerning the definition and subsequent measurement of each CEL construct. Despite the absence of explicitly stated definitions, the review unearthed several conceptual similarities among researchers, suggesting a potential foundation for the development of a unified approach to CEL definition within ISR. The systematic review highlighted several methodological challenges with the measurement of CEL. This revealed a need for the field to develop experimental designs which seek to minimise multi-causality while also utilising multiple methods to enhance the validity and robustness of findings. The next section of this theoretical work relates to the identification and development of an ISR sub-task which seeks to address some of these challenges, allowing for the examination of effort and load measures in a tightly controlled experimental context.

Chapter 4

Relevance Judgement Model

The systematic review provided a comprehensive overview of Information Seeking and Retrieval (ISR) research on cost, effort, and load (CEL), identifying several areas that remain under-explored despite their importance across ISR tasks and domains. Much of the existing literature has examined CEL at the level of the overall search session, with less attention given to how these constructs emerge within specific component tasks. One such task is **relevance judgement**, in which users assess the value of individual documents. Document judgement is integral to every search session, underpinning the process by which users filter, select, and make sense of retrieved information, yet it has rarely been examined in detail from the perspective of effort and load. Focusing on CEL at this level provides the opportunity to gain more precise insight into how these constructs influence user behaviour during evaluation, as well as to refine the methods used to measure them. This chapter introduces the second theoretical contribution of this thesis: a **multi-stage relevance judgement model**. Building on existing models of relevance assessment in ISR, this model extends prior work by explicitly incorporating the roles of effort and cognitive load in the document judgement process, an aspect not systematically addressed to date. The model forms the foundation for the empirical studies presented in this thesis, with each judgement stage analysed in relation to user effort and load using a range of complementary measurement techniques.

It is important to note, that following the development of working definitions in the previous chapter (Chapter 3), cost was conceptualised as the external resources

that a user pays or spends, while effort and load were defined as the consumption and expenditure of the user's internal resources. As the next stage of this thesis is concerned with user behaviour and cognitive processes during the document judgement task, the focus therefore shifts specifically to effort and load.

Based on this framing, this chapter introduces the research questions that guided the development of the multi-stage relevance judgement model, focusing on how effort and cognitive load can be theoretically integrated into document evaluation and meaningfully represented across different stages of judgement.

The research questions are as follows: **HL-RQ2 (Theoretical Model): How can effort and load be integrated into a multi-stage model of relevance judgement?**

- (a) What are the key stages involved in relevance judgement, based on existing theory?
- (b) How do effort, and load theoretically influence each stage of relevance judgement?
- (c) How can these theoretical relationships be structured to form a coherent multi-stage model?

4.1 Relevance Judgement Literature

A common ISR task which requires varying levels of intrinsic cognitive processing is the task of judging documents for relevance. The concept of relevance judgement has been of continual and significant interest within the field of ISR for many years [137–140]. Relevance judgements form an integral aspect of the search process, the development of test collections, the evaluation of information retrieval systems, and professional search tasks across most professional domains [25, 110]. In recent years, the focus on relevance as a system or algorithmic problem has noticeably shifted within the research community towards a more user-oriented and subjective perspective of relevance [138–143]. So with this emerged a quantity of research examining various factors that may influence user behaviour during the judgement process.

4.1.1 Relevance Judgement User Studies

As noted earlier, relevance judgements are a core component of the search process, underpin the development of test collections, and play a central role in the evaluation of information retrieval systems [25,110]. Consequently, many ISR studies have employed relevance judgement tasks as an empirical method of investigation. Given the breadth of literature on relevance judgement, this thesis necessarily focuses on a subset of studies, specifically those examining relevance judgement in relation to user behaviour (e.g., time on task, mouse movement) or cognitive processes (e.g., attention, effort).

Some studies have examined the processes and behaviours specific to individual assessors as they make relevance judgements. For example, Al-harbi [144] examined the relevance judging behaviour of secondary assessors using a think-aloud user study. Users performed four search tasks and were asked to judge 36 documents, one at a time, as either relevant or non-relevant in relation to four search topics. As they performed their task they were asked to verbalise their thoughts. The findings revealed that assessor relevance judgements are associated with a range of certainty levels, ranging from low to high. The think-aloud data highlighted factors which contributed to incorrect relevance judgements such as; difficulty in applying the search topic; difficulty in processing the document; and assessor lack of knowledge/concentration.

Smucker [145] examined the degree to which mouse movements can indicate assessor attention during a relevance judgement task. Similar to the Al-harbi [144] study, this article asked users to judge the relevance of document and document summaries, one by one, against four search topics. As users performed the task, their mouse movements were recorded. The results showed that the average participant judges 76% of the document summaries without making any indication of what their attention is focused on, and 41% of the full documents were also judged with no indication of user attention.

Villa and Halvey [110] were interested in the effort users expend when making relevance judgements, and whether the size of the document or relevance grade of the document have any influence on user effort or accuracy. Similar to the other study designs, users were presented with a search topic followed by a document and asked to judge the relevance of the document to the topic. Effort was measured through the

NASA-TLX; the mean time to make relevance judgements; and mean number of topic view clicks. Results revealed that document length does influence the amount of effort invested when judging a document, but not the accuracy. Whereas document relevance level affects both the accuracy and effort.

Some studies have examined assessor disagreement while making relevance judgements, and the factors that may influence this. For example, Yilmaz et al. [25] examined the disparity between real-user relevance judgements and relevance judgements given by IR relevance assessors, arguing that the level of effort invested can explain the differences between the two. To test their hypothesis, a user study was carried out using three data sets, where highly trained assessors and crowd-sourced judges were asked to judge query-document pairs on a binary relevance scale. The authors were mainly interested in any cases where the utility of a document to a real-user differs from the relevance of the document. Results revealed that relevance judges tend to spend more time on documents that require the user to invest more effort when finding and gathering relevant information. In these cases, users perceive the document to be of low utility due to a reluctance to invest the effort, even when the document is actually relevant. Therefore, effort is considered as the key factor which may influence the disagreements between the relevance and utility of a document. The authors further validate these findings by identifying features of a "high effort" document such as: readability level of the document; document length; and location of query terms in the document.

Similar to the interest in real-user relevance assessment compared to artificial relevance assessment, Chu [146] examined relevance assessment within a realistic context where the assessors were also searchers. More specifically, the article examined factors which may influence relevance judgement of retrieved documents in the context of legal search - as part of the TREC 2007 Legal track interactive task. Participants were asked to search for documents and submit their top 100 results, ranked on a three-point relevance scale. Data related to relevance judgements was collected via a questionnaire and users were also asked to select and rank three relevance factors from 6 categories (from a list of 80 relevance factors). The findings identified factors such as; having specific and adequate information in the request; ease of relevance scale use; topicality

and kind of response required as the most important factors to participants as they make relevance judgements.

Damessie [147] was also interested in the key factors that may influence the judging scenario. More specifically, the authors were interested in how search topic difficulty; document degree of relevance to search topic; and document ordering influenced the time taken to make a relevance judgement decision. To determine these relationships, a user study was conducted where each participant made relevance assessments (using four levels of relevance grade i.e., highly relevant, relevant, marginally relevant, or non relevant) for two search topics (one easy; one hard). Participants were presented with a sequence of topic-document pairs, one at a time. The top of the screen displayed the Title, Description, and Narrative of the TREC search topic followed by a single document underneath. Dwell time was measured from when the current documents page loaded until the time the assessor clicked the submit button to record their relevance judgement. Subjective topic difficulty was also considered and measured in the exit questionnaire item - "how easy was it to identify relevant documents for the search topic?". The findings revealed that judges were found to read documents more quickly for easy topics than the hard ones, and that assessors process non-relevant documents more quickly than marginally relevant, relevant, or highly relevant documents - i.e., assessors spend less time on non-relevant documents.

Professional Search

The sub-task of judging or assessing documents is undertaken across a range of search domains and is often a critical component in various professional contexts, such as legal and clinical information retrieval. In these domains, the process of document assessment is inherently complex and can have significant consequences if carried out inadequately.

Within the legal search domain, the process of retrieving and assessing legal documents can be a cumbersome and time-consuming task, even for professionals in the field. A typical legal search task entails querying an extensive database of previous cases to locate documents pertinent to a narrowly defined situation [148]. In such

tasks, users may seek to: (1) understand how the legal issue aligns with established legal concepts; (2) identify potentially applicable legal actions; and (3) examine the outcomes of comparable cases. In summary, the user must navigate a substantial corpus of legal documents and evaluate them based on their relevance [148]. Inaccurate relevance assessments can have significant consequences, including undermining knowledge acquisition and even access to justice [149]. Not only do legal databases possess a vast number of documents, the documents themselves often possess distinct characteristics that can potentially influence cognitive effort and judgement, such as considerable length, necessitating metadata and summaries to reduce the cognitive load. The domain is also highly heterogeneous, encompassing diverse document types (e.g., legislation, court decisions, parliamentary documents, case-law notes), therefore requiring a variety of skills for accurate interpretation. Additionally, legal language is both precise and ambiguous, further contributing to cognitive load by requiring sustained attention and interpretative effort during relevance assessment. While no studies have directly examined user effort and load within the legal search domain, there has been examination of related constructs such as task difficulty and judgement speed (often used as a proxy for cognitive load). For example, Wang [150] investigated user relevance judgement in legal e-discovery, where electronically stored documents (e.g., emails, balance sheets) are assessed for relevance in legal cases. The study examined accuracy, agreement, speed, and perceived difficulty in binary relevance judgements. Users rated judgement difficulty and provided justifications. Results indicated lower accuracy for difficult-to-judge documents, while easier judgements were made more quickly.

Clinical information retrieval systems play a vital role in enabling clinicians and researchers to efficiently access relevant medical information and literature. In recent years, the volume of information available in electronic health records (EHRs) has become vast and can inform many aspects of a patients care including, conditions, medical treatments and so forth [151]. As in legal search, clinical information retrieval systems encounter similar challenges such as vocabulary differences, conceptual ambiguity, and term overlap when retrieving information [152]. From a user-perspective, the professionals interacting with these systems are most likely limited by time, memory, and

access to information [153]. As Electronic Health Records (EHRs) are used to support critical clinical decision-making and a range of secondary purposes, such as identifying candidates for clinical trials and contributing to systematic reviews, it is essential that these systems function effectively to prevent potentially harmful outcomes. Approximately 30 percent of electronic health record (EHR) systems have been shown to underperform in clinical settings, largely due to insufficient attention to user-system interaction and the resulting disruption to clinical workflows [154]. User interface evaluations of these clinical IR systems identified the key user concerns as (1) the large number of steps to perform a task (i.e., number of clicks and menus); and (2) the mental workload due to the time taken to perform all the necessary tasks. In addition to multiple data entry steps, the user is also required to make several document judgement decisions throughout to assist the diagnosis of the patient. The demand imposed by the process is further compounded by a vast quantity of information, which medical professionals are required to filter. The two-fold effect of a poorly designed system combined with information overload can result in mental fatigue, adverse patient outcomes, and also the well-being of the medical professional [155]. Koopman and Zuccon [152], specifically examined the demand involved in assessing the relevance of clinical health records. The researchers asked four medical professionals to rate the relevance of clinical documents as "highly relevant," "somewhat relevant," or "not relevant," with optional justifications, and then asked them to report task difficulty. Findings showed that while document length had no impact on cognitive load, assessing "somewhat relevant" documents was most cognitively demanding. Furthermore, query characteristics such as temporality, subjectiveness of meaning, and the inclusion of multiple dependent elements in the query were found to increase cognitive demand. The study concluded that it is possible that high levels of cognitive demand may negatively impact the ability of the user to make accurate relevance judgements.

This discussion highlights the central role of relevance judgement in professional search tasks. In such contexts, the process of determining relevance can be cognitively demanding and cumbersome, with potentially high-stakes consequences if performed inadequately. However, there remains a notable lack of empirical research that explicitly

investigates the associated cognitive effort and load in these settings.

4.1.2 Summary

The studies discussed in this section reflect the different ways in which relevance judgement has been examined within ISR. While constructs such as effort have been acknowledged as important in this process, there have been very few studies which examine effort or cognitive load as the key variable of interest. Furthermore, although the user-oriented, cognitive perspective of relevance judgement suggests that users evaluate information based on multiple levels of relevance and a diverse range of criteria to satisfy their information needs [139, 141, 156], there has been limited systematic investigation into this multidimensional conception of relevance and its impact on user experience and behaviour. Moreover, limited attention has been given to understanding these stages in relation to CEL.

4.2 Multi-Stage Relevance Judgement

Interest in the concept of multi-stage relevance judgement primarily emerged when the focus on relevance as a system or algorithmic problem within the ISR community began to move towards a more user-oriented and subjective interpretation of relevance [138, 141,143]. From the literature discussed in the previous section, relevance judgement has been a core concept in ISR research. While early IR models conceptualised relevance as a binary, system-orientated outcome (i.e., relevant or not relevant), decades of research in this area have revealed that relevance is a far more nuanced, context dependent, and dynamic process. Within this broader understanding, multi-stage relevance judgement is based on the idea that relevance judgements are dynamic, multidimensional, and subjective [141, 157]. This means that as a user performs and progresses through a search task, their cognitive processes will change. For instance, a document that may seem relevant at the beginning of their task may not be considered relevant at the end of their task. Relevance judgements are also subjective and therefore will likely differ among users. It seems that relevance judgements are influenced by time, context,

and situation. From this perspective, a more complex framework which goes beyond a dichotomous and static view of relevance, is necessary. The recognition of multi-stage relevance judgement evolved in tandem with advances in interactive IR and user-centered evaluation. In the 1960s and 70s, system-focused IR research typically used binary, assessor based judgements - often derived from test collections such as Cranfield, to evaluate retrieval effectiveness [158]. This approach treated relevance as an absolute, overlooking user context, situation, and cognition. By the 1980s and 90s, scholars such as Saracevic [141, 159] began to distinguish between different types and levels of relevance, emphasising that judgements were subjective, context-dependent and often temporal. This period marked a shift toward cognitive and situational perspectives, laying the groundwork for models of iterative or multi-dimensional relevance evaluation.

The multi-dimensional perspective of relevance led to a body of research which examined the attributes and manifestations of relevance, and more specifically the criteria used by searchers to make relevance judgement decisions. This research is grounded in the premise that users make relevance judgements based on multiple criteria. A prominent model of multidimensional relevance is Saracevic's stratified framework [141, 159], which situates relevance within context and conceptualises its manifestations as distinct attributes or dimensions that extend beyond mere topicality. In Saracevic's [141] *Stratified Model of Relevance*, five different types of relevance are proposed, which are as follows:

1. **System or algorithmic relevance** - relates to how well the query and document align (as retrieved or not retrieved by the system) through a procedure or algorithm.

2. **Topical or subject relevance** - relates to how well the query topic and the topic detailed by the document align - here topicality is associated with *aboutness*.

3. **Cognitive relevance or pertinence** - relates to the level of alignment between the individuals information need and the document. This type of relevance encompasses criteria such as *informativeness*, *novelty*, *information quality*.

4. **Situational relevance or utility** - refers to the relationships between the con-

text, task, or information problem, and the document. This type of relevance includes criteria such as *usefulness*, *appropriateness of information*, and *lessened uncertainty*.

5. **Motivational or affective relevance** - relates to the how well the goals and motivations of the user align with the document. This type of relevance encompasses criteria such as *satisfaction*, *success*, and *achievement*.

This theoretical framework conceptualises relevance as a multi-layered, interactive phenomenon involving both the system and the user. The model identifies multiple strata (or levels) at which relevance can be understood and assessed, with a strong emphasis on the cognitive and situational aspects of users. Borlund [138] suggests that the models cognitive and situational levels signify the main types of *subjective relevance* - a type of relevance that is user-centered rather than system centered. Subjective relevance is also proposed to be situation-dependent, and relates to the aboutness, utility, or usefulness of a document relative to the individuals' goals, interests, work tasks, or problem-solving needs. Cosijn and Ingwersen [160], further developed the work of Saracevic [141, 159], establishing a revised framework of attributes and manifestations of relevance. The manifestations of relevance included: topical, cognitive/pertinence, situational/utility, and socio-cognitive. These manifestations align with Saracevic's [141, 159], "affective relevance", which are proposed to represent expressions of cognitive change. User studies conducted by Barry and colleagues [137, 161, 162] identified relevance criteria and categories which further aligned with "affective relevance' - indicating that through user interaction cognitive processes change over time. Cosijn and Ingwersen [160] argue that it is the temporal dimension (i.e., the progression of time) that affects users' relevance decisions, with cognitive changes emerging through ongoing interaction playing a central role in shaping this influence. If relevance judgements are influenced by the interaction of temporal, affective, and cognitive factors, then relevance may be understood as situational. Within this conceptualisation, relevance is defined as the user's perception of a document's significance in relation to their information need. It is characterised as multi-dimensional (differing across users), dynamic (changing over time), and complex yet systematic in

nature [138]. Taylor [156] proposes that this holistic approach to relevance judgement involves multiple stages in an extensive cognitive process. Research such as that from Kuhlthau [163] examined the information search process from the perspective of the user, which led to the development of a model of information seeking which acknowledged cognitive and affective components and described the process as a "series of encounters with information... transforming information into meaning" (p.361) [163].

The acknowledgement that relevance is multi-faceted and the information search process consists of multiple stages is reflected in a vast body of ISR research. Numerous studies have established a variety of classifications and groupings of relevance. For instance, Barry and colleagues [137, 161] examined the subjective evaluation of relevance, identifying a set of relevance criteria. Barry [161] conducted a study examining document selection during search which identified 23 categories of relevance with several criteria within these categories. These criteria applied to a variety of factors including information content, subjective interpretations of the document relating to topic knowledge; contextual factors; and quality of the document. This study revealed insights into both the relevance judgement process and the cognitive processes which may occur. Consistency in these criteria were later supported in a study conducted by Park [162] and a follow-up study by Barry and Schamber [137]. Xu and Chen [142] examined a subset of relevance criteria identified in a previous study with the aim of establishing core criteria used by users when making relevance judgements. Five criteria were identified as follows: *topicality*, *novelty*, *understandability*, *reliability*, and *scope*. Similarly, Wang and Soergel [164] propose *topicality*, *orientation*, *quality*, and *novelty*, as the main criteria users associate with relevance. Later, Greisdorf [165] suggested *topicality*, *pertinence*, and *utility* as criteria which together infer the relevance of a document.

It is important to note, that these criteria are not considered static, rather relevance behaviour and the criteria used to assess relevance will change depending on the stage of the information search process and the user's situation [156]. Tang and Solomon [166] found changes in *clarity*, *importance*, *newness*, *recency*, and *topicality* as users progressed from evaluation of document surrogates to the evaluation of full-text

documents. Vakkari [167] found that searchers consider more documents as relevant in the early stages of the information search process and fewer documents as relevant in the later stages. Wang [164] examined different search stages and found that criteria such as *topicality*, *novelty*, and *recency* were the most commonly selected criteria- with more diverse relevance criteria applied in the later stages. Hirsh [168] found that the *topicality* criteria, was less important to searchers in the later search stages.

The literature suggests that relevance can manifest itself in different ways and users consider a range of criteria beyond topicality when making relevance judgements. Foundational studies, such as those by Saracevic [141, 159], conceptualised relevance manifestations at a philosophical and abstract level. More recently, research has focused on the plurality of relevance by examining the diverse judgement criteria employed by users. The process of judging relevance, as well as the criteria applied, are influenced by cognitive changes that unfold during the assessment. Such cognitive changes are likely to be closely associated with fluctuations in user effort and cognitive load, given that shifts in attention, memory, and decision-making demands can place varying levels of strain on the user's cognitive resources. However, what remains unclear is how these different criteria and stages of the relevance judgement process specifically affect effort and cognitive load.

Building on the preceding literature and the identified relevance criteria, the next section outlines a preliminary framework for exploring how different stages and factors in the relevance judgement process may influence user effort and cognitive load.

4.2.1 Multi-Stage Relevance Judgement Model

The multi-stage relevance judgement model presented in this thesis underwent several iterations throughout the course of the empirical work, as will be discussed in detail in Chapters 7, 8, and 9. Nevertheless, the underlying theoretical foundation and the four proposed stages remained largely consistent. This section provides an overview of the model as it was applied in the empirical studies, along with the theoretical rationale underpinning its development.

Purpose of Model

The systematic literature review presented in Chapter 3 revealed that constructs such as effort and cognitive load are seldom examined as primary variables of interest. Instead, they are often measured in a superficial or ad hoc manner - typically using a single-item measure within study contexts where numerous variables remain uncontrolled. An additional challenge in measuring effort and cognitive load lies in the frequent failure of ISR study designs to differentiate between distinct types (i.e., intrinsic, extraneous, germane) of cognitive load. Consequently, it remains unclear which specific type of cognitive load is placing greater demands on the user's cognitive resources during the task. These challenges underscore the need for an experimental design and task that enables the isolated measurement of a single type of cognitive load within a controlled context, in which the influence of other load types is minimised through careful design. As previously discussed, relevance judgement is a central and integral component of ISR; however, the influence of user effort and cognitive load during these tasks remains under explored. The multi-stage relevance judgement model presented in this section was developed to address the need for isolating and examining a specific type of cognitive load - intrinsic load - within the context of a critical and broadly generalisable ISR task. By structuring the task around a series of relevance judgement questions, the model enables a systematic investigation of intrinsic cognitive load, with each question designed to elicit varying levels of cognitive complexity. Moreover, to address common methodological limitations in the measurement of effort and load, the model integrates multiple assessment points, capturing both in-task and post-task data. This approach facilitates a more nuanced and reliable evaluation of user effort and cognitive load across different dimensions of the relevance judgement process.

The Model

The **Multi-Stage Relevance Judgement Model** developed in this thesis is grounded in, and informed by, established research within ISR, rather than being proposed as an entirely new theoretical account of relevance. Its conceptual foundations lie primarily in Saracevic's Stratified Model [159] of relevance, which characterises relevance as

a multi-level construct encompassing topical, cognitive, situational, and motivational dimensions. In addition, the model draws on prior empirical studies [142, 164, 166] outlined in Section 4.2 that operationalise relevance through discrete judgement criteria in document evaluation tasks, as well as research in cognitive load theory that highlights differences in processing demands across evaluative activities [42]. The present model represents a synthesis of these strands of prior work, adapted to support controlled experimental investigation of relevance judgement as a cognitively demanding sub-task within ISR.

While numerous multi-stage judgement models exist within ISR, many converge on similar conceptualisations of relevance criteria. This thesis focuses specifically on topical, cognitive, and situational relevance as defined in Saracevic's Stratified Model [159]. System relevance was excluded as it is largely determined algorithmically, and motivational relevance was omitted due to its reliance on affective and long-term goal structures that extend beyond the scope of controlled laboratory studies. While system relevance is largely determined algorithmically and motivational relevance extends into the domain of individual affect and long-term goals. Topical, cognitive, and situational relevance were the focus directly reflect how users interact with, interpret, and apply information in context. The strata of topical, cognitive, and situational relevance were chosen as they capture core user-driven relevance processes while remaining amenable to operationalisation within document judgement tasks. Within these strata, the criteria of *aboutness*, *relevance*, *novelty*, and *usefulness* were selected as they are well-established in ISR literature, frequently employed in relevance judgement studies, and conceptually distinct in the type and depth of cognitive processing they require. Together, these components allow the model to represent increasing levels of evaluative complexity while remaining suitable for controlled experimental investigation. These levels encompass both surface-level content matching and deeper evaluations involving knowledge integration and task alignment, processes that are inherently linked to cognitive load and user effort.

Throughout this thesis, the term **relevance** is used in two related but distinct ways. First, relevance is treated as a multi-faceted construct, consistent with ISR theory out-

lined in Section 4.2, encompassing multiple dimensions through which users evaluate information. Second, relevance is used in a narrower, operational sense to denote a specific stage within the Relevance Judgement Model. In this latter case, relevance refers to a criterion-based evaluative judgement that extends beyond surface-level topical matching, but precedes comparative (novelty) and task-oriented (usefulness) assessments.

To further operationalise the cognitive demands associated with relevance judgements, this study draws on a set of commonly used relevance criteria - ***aboutness***, ***relevance***, ***novelty***, and ***usefulness*** - and situates them within Saracevic's stratified model. *Aboutness* (often referred to as topicality) aligns with topical relevance, reflecting users' initial assessments of whether a document is about the queried subject. *Relevance* and *novelty* are mapped onto cognitive relevance, as they require users to critically evaluate the relevance of the content to the information need and its contribution to their existing knowledge as gathered from previous documents. *Usefulness* corresponds with situational relevance, capturing the extent to which the information supports a specific task or goal. This mapping allows for a structured understanding of how different criteria invoke varying levels of cognitive processing and, consequently, contribute to user effort and cognitive load across the relevance judgement process.

For all studies conducted for this thesis, participants were required to complete a document judgement task. For the purposes of this study, relevance judgements are defined as the process by which a user evaluates a document based on a set of relevance criteria, employing a categorical rating scale. Relevance criteria refer to the factors that inform and influence the user's assessment of a document's relevance, in this case, these are *aboutness*, *relevance*, *novelty*, *usefulness*. Although the model is described in terms of stages to reflect increasing cognitive processing, this should not be taken to suggest a fixed or linear sequence. Relevance judgements are inherently dynamic, with users often moving fluidly between different evaluative processes.

Although the judgement questions varied slightly in wording due to participant feedback, as the studies progressed, each question was based around the four stages outlined in the model: *aboutness*; *relevance*; *novelty*; and *usefulness*. Figure 4.1 depicts how each stage of document judgement maps to the judgement questions used.

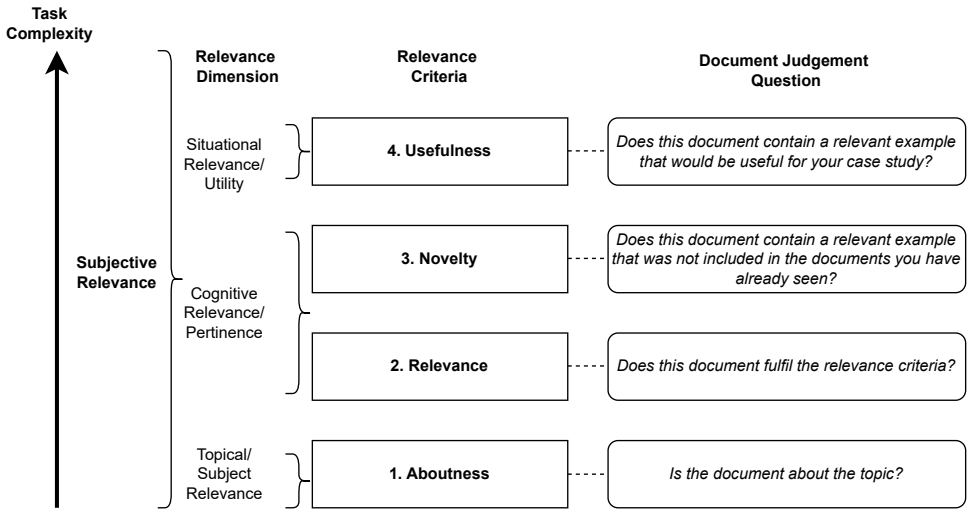

Figure 4.1: Conceptual representation of the multi-stage relevance judgement model used in the empirical studies of this thesis.
Note: The staged structure reflects increasing levels of cognitive processing for experimental purposes and does not imply fixed or linear sequence of relevance assessment in naturalistic contexts

It was important that the task allowed the manipulation of one load type - intrinsic load (refer to Section 5.1 for more detail). Each stage outlined in the multi-stage relevance judgement model is proposed to vary in relation to the level of intrinsic processing required. The higher the stratum (or stage), the greater the potential cognitive complexity and subjective effort required. This model provides a structured way to understand how user effort and load may vary in the document judgement process, grounded in user-centered relevance dynamics.

It is important to note that within ISR, *complexity* has faced challenges similar to CEL in terms of measurement and operationalisation (see Wildemuth et al., [169] for a comprehensive review). In the analysis of complexity conducted by Wildemuth et al., *search task complexity* is primarily examined and *cognitive complexity* is positioned as a property of the broader work task, tied to processes that either motivate the search or shape its outcomes (e.g., decision making). In contrast, the present thesis focuses on *document judgement*, a sub-task within the larger search process, rather than a full

search task, which is defined as "goal-directed activities carried out using search systems" (p.1134, [169]). While extensive research has investigated complexity at the level of search tasks, little attention has been given to defining or operationalising complexity within sub-tasks such as document judgement. The task of judging documents can certainly be linked to elements of cognitive complexity, particularly decision making; however, a full treatment of complexity in this type of task lies beyond the scope of this thesis. Accordingly, complexity is here considered in terms of intrinsic load and the interrelatedness of elements within the document judgement task.

Stage 1 involves an initial *aboutness* judgement, in which the user assesses whether the document pertains to the topic of interest. The term "*aboutness*" is commonly used interchangeably with topical relevance, and this judgement typically relies on surface-level content features such as titles, keywords, or headings. As such, it is hypothesised to involve the lowest level of element interactivity, given that the decision is based on a single, relatively straightforward criterion.

Stage 2 corresponds to a *relevance* judgement in the narrower, operational sense of the term. Although all stages of the model represent facets of relevance assessment, this stage captures a specific form of cognitive evaluation in which users assess the degree to which a document satisfies pre-defined relevance criteria beyond simple topical alignment. Including relevance as a distinct stage allows the model to differentiate between initial topical screening (*aboutness*) and deeper evaluative reasoning that involves integrating multiple document features (*novelty*) and subjective interpretations of the information need *usefulness*. While the reuse of the term relevance to describe both the overarching construct and a specific judgement stage may appear redundant, it reflects established practice in ISR research, where relevance is frequently operationalised as a discrete evaluative judgement within broader models of information seeking. Therefore, unlike Stage 1, this process demands the integration of multiple informational elements, thereby increasing the cognitive complexity and element interactivity involved in the judgement.

Stage 3 introduces a *novelty* judgement, which requires the user to determine whether the document contains new and relevant information compared to previously

viewed documents [142]. This stage is cognitively more demanding, as it involves both an assessment of content relevance and a comparison across memory traces of earlier documents. The element interactivity at this stage is considered high, given the need for retrieval, comparison, and integration of prior information. Stage 4 reflects a *usefulness* judgement, which is regarded as the most cognitively demanding of the four stages. Here, the user must evaluate the document's utility in relation to the overarching task or goal, rather than its topical alignment or contribution to knowledge alone. This requires task-based reasoning and contextual evaluation, and thus involves the highest level of element interactivity, integrating content understanding, prior knowledge, and goal alignment.

Although the model is described in terms of stages to reflect increasing levels of cognitive processing, this should not be interpreted as implying a strict or natural sequence of relevance assessment. In practice, users may revisit or collapse these evaluative processes as their understanding of the information space evolves. The staged representation is therefore a methodological abstraction designed to support experimental control and the systematic examination of cognitive load during document judgement, rather than a claim that relevance assessment unfolds linearly in naturalistic search contexts. As a result, while the model provides a structured and theoretically grounded means of examining cognitive load during document judgement, its staged representation necessarily simplifies the fluid and iterative nature of relevance assessment in naturalistic search, which should be considered a limitation of the approach.

4.2.2 Summary

This chapter addressed the second high-level research question of this thesis (**HL-RQ2: How can effort and load be integrated into a multi-stage model of relevance judgement?**) by developing and presenting a novel model of multi-stage relevance judgement. In answering the high-level research questions, the following sub-questions were also addressed:

- (a) What are the key stages involved in relevance judgement, based on existing theory?

- (b) How do effort, and load theoretically influence each stage of relevance judgement?
- (c) How can these theoretical relationships be structured to form a coherent multi-stage model?

In response to sub-question (a), the chapter identified the key stages of document evaluation by drawing on existing theory, particularly Saracevic's Stratified Model of Relevance. Four stages were operationalised, *aboutness*, *relevance*, *novelty*, and *usefulness*, each reflecting increasing levels of cognitive processing and aligning with topical, cognitive, and situational relevance.

In relation to sub-question (b), the chapter examined how effort and cognitive load can be theoretically mapped onto these stages. Building on the working definitions developed in Chapter 3, effort and load were conceptualised in relation to the demand on a user's internal cognitive resources, with intrinsic load hypothesised to vary systematically across the four stages according to the level of element interactivity and processing demands. This perspective positions effort and load as central factors shaping the dynamics of the relevance judgement process.

Finally, in response to sub-question (c), the chapter demonstrated how these relationships can be structured into a coherent multi-stage model. By integrating user-centred relevance criteria with cognitive load theory, the model provides a framework for isolating and examining intrinsic load within a controlled ISR task. The staged structure enables a systematic analysis of how user effort and load vary during document judgement, and it forms the theoretical foundation for the empirical studies presented in Chapters 7–9.

Part III

Empirical Contributions

Chapter 5

General Methodology

This chapter provides an overview of the *general methodology* used in the empirical work of this thesis. The chapters (7,8,9) which follow, report on four empirical contributions, relating to the third high-level research question outlined in Section 1.3: **HL-RQ3 (Empirical): How can effort and load be measured within the ISR sub-task of multi-stage document judgement, as defined by the theoretical model?**

- (a) How do different stages of document judgement vary in terms of user effort and load?
- (b) To what extent is effort and load influenced by the type of judgement rating (i.e., "no", "partially", "yes") between document judgement stages?)
- (c) What are the operational relationships between effort and load?
- (d) How do measures of effort and load change over time during the task?
- (e) To what extent do user characteristics influence measures of effort and load?

5.1 Motivation

While the systematic literature review (Chapter 3) revealed that effort and load, particularly workload, are widely researched within ISR, a key observation was that they are seldom examined as the primary construct of interest (refer to sub-section 3.4). Rather, effort and load measurements tend to be included in these studies in the form

of a "quick and dirty" approach where one measurement is taken, often with little justification provided for its use. A further challenge in effort and load measurement, relates to the lack of distinction made between the three cognitive load types (i.e., intrinsic, extraneous, germane) within the reviewed ISR studies experimental design. Subsequently, it is unclear in these studies (i.e., those reviewed in the systematic review) which type of cognitive load is imposing higher demand on the users cognitive resources during the task. The empirical work conducted for this thesis focused on examining and measuring the **intrinsic** load only. Previous research [42, 170] interprets germane load as a dimension of intrinsic load, and therefore does not examine germane load as a distinct load type. The following empirical work is based on this assumption, and therefore the experimental design used only controls for extraneous load (i.e., the load imposed by the design of the task presentation and context).

Also note, that the construct of **cost** is not included in the empirical work for this thesis. As aforementioned in Section 3.4, cost during the ISR process is considered to arise from the consumption of the users *external* resources, such as time and money. As measures of cost can be considered more tangible and objective to gather, empirical examination of these measures seemed less fruitful. Rather, the empirical work in this thesis will focus on the consumption of the users *internal* or *cognitive resources*, which relate to the effort and load exerted and experienced by the user. Measures of effort and load are more difficult to interpret and were therefore deemed to be in need of further investigation.

In light of the methodological issues highlighted in the systematic review (Chapter 3) and to answer part three of the high-level research questions outlined in Section 1.3, it was important that the empirical work contained the following attributes:

1. Manipulates only one cognitive load type.

2. Reflects a real-life search task that can be generalised across different search domains.

3. Is narrow and controlled in design in order to control for confounding variables which may affect the demand imposed on the user.

4. Allows the inclusion of both objective and subjective measures of effort and load (i.e., methodological triangulation), that can be administered both during and after the task.

To address the requirements outlined above, a task based on the **multi-stage relevance judgement model** (Chapter 4) was deemed appropriate. As discussed in Section 4.2.1, the different stages and decisions involved in judging a document for relevance entail varying levels of intrinsic cognitive processing, thereby providing a means to examine how intrinsic load shifts across the user's evaluation process. Although some studies have investigated effort and load in the context of document relevance judgement, they have primarily focused on the influence of document attributes, such as length [110], relevance level [110], and readability [88], on user effort and load. However, the theoretical work presented in Chapter 4 suggests that no studies to date have examined effort and load within the framework of a multi-stage document judgement process.

To design a document judgement task which fulfilled the requirements, a model of multi-stage document judgement (see Figure 4.1) was developed, involving four document judgement stages with corresponding questions in which the users answers in relation to a document. This multi-stage document judgement task was considered to possess the following advantages:

- Reflects a real-life search task that forms an integral part of the search process and can be applied to a variety of search contexts, such as professional search, the development of test collections, and general web-search tasks.
- The narrow and controlled design allows one type of load to be measured, while the other load types are controlled.
- Can be conducted via different modalities - both an in-person lab environment or via an online platform.
- Multiple measures of effort and load can be integrated, and the use of different judgement stages allow for during-task and post-task measures of effort and load which allows for a deeper analysis of relationships between constructs.

- The simplicity of the experimental design and functions makes the task relatively user friendly, and requires no prior topic knowledge to take part.

5.2 Documents and Topics

For all of the user studies conducted, documents from the TREC Washington Post collection were used. The collection consists of 728,626 articles from the period of 2012-2020. Four topics were chosen from the TREC 2018 Common Core Track [171] which contains pre-assessed relevance judgements for documents. The four chosen topics were as follows: Wildlife Extinction; Airport Security; Transportation Tunnel Disasters; and Tropical Storms. From each topic collection, ten documents were selected, five relevant and five non-relevant as judged by TREC. These four topics were selected for their comparable topical interest, difficulty, and document format. This consistency was important in order to control for extraneous variation across user studies and to ensure that observed differences in behaviour or perception could be more confidently attributed to experimental conditions rather than topic-related factors.

5.3 Experimental Interface

Each user study was designed on the web based survey generating software, Qualtrics. Upon commencing the experiment, the study would launch in a new window of the web-browser being used.

5.4 Experimental Procedure and Flow

Each user study was designed to take around 30 minutes, this included the completion of the document judgement task and the pre-task and post-task questionnaires. All experiments followed a similar structure, where participants were asked to complete two questionnaires before beginning the document judgement task, and one post-hoc questionnaire following completion of the task. These questionnaires allowed us to collect demographic information; level of topic knowledge (Study 3) and motivation

(Study 3) of participants; and post-task workload assessment. Figure 5.1 provides an overview of the general experimental procedure used for our studies.

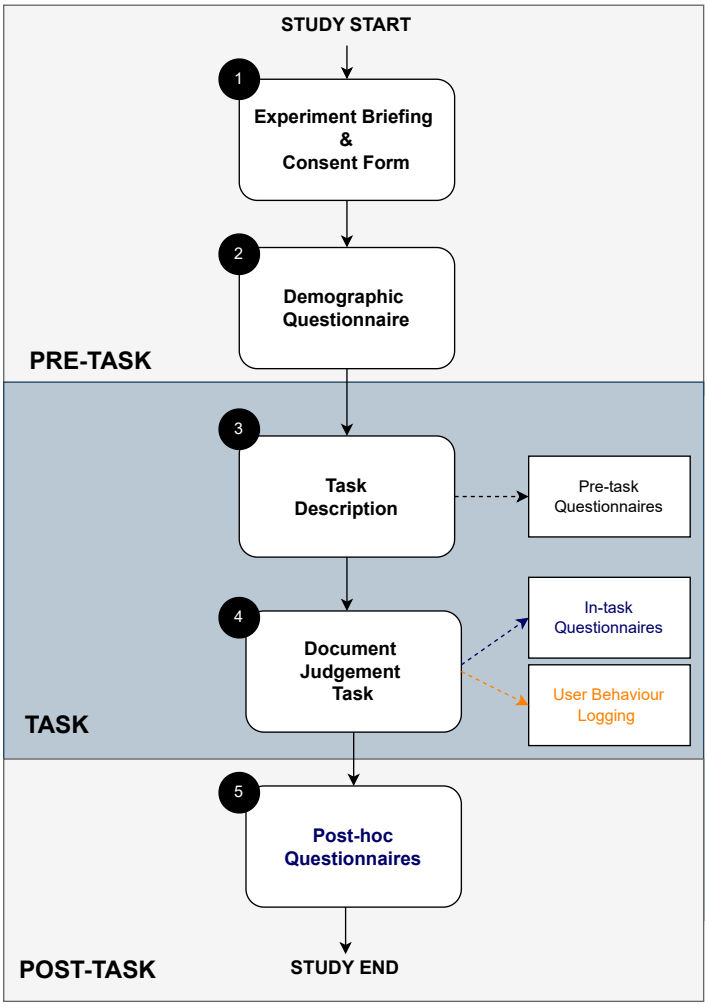

Figure 5.1: General experimental procedure
In-task questionnaires and post-hoc questionnaires (in purple), and user behaviour logging (in orange) were used to measure participant effort and load at various points in the study.

5.4.1 Document Judgement Task

The experiments conducted in this thesis employed either within-subjects or between-subjects designs, meaning that participants were required to answer either all four document judgement questions per document or only one type of judgement question per document. Despite this variation, several aspects of the document judgement task

remained consistent across all experiments, as outlined below:

- To make the task more concrete, participants were asked to imagine they were delegates involved in a fictional assessment centre for a vacant job role for a large organisation. The ability to synthesise and process information comprised a key aspect of this role and this skill would be assessed during a document judgement task. The task design reflects Borlund's [138] approach of improving the realism of search tasks through their integration into simulated work task scenarios.
- Participants were firstly provided with a task description describing the criteria associated with a relevant document. For example, for the topic *Wildlife Extinction* participants were explicitly informed that: "a relevant Wildlife Extinction document will specify: the country, the species involved, and the steps taken to save the species".
- Participants judged ten documents each (five relevant and five non-relevant).
- The order of documents were presented at random for each participant to minimise order effects.
- Participants could not return to a previously viewed document.
- All questions used across each of studies reflect the four stages (*aboutness, relevance, novelty* and *usefulness*) outlined in the model of multi-stage relevance judgement (See Chapter 4 and Figure 4.1).

5.5 Crowd-Sourced Participant Considerations

For all of the user studies, crowd-sourcing was used as a means of recruiting participants. Since its emergence in 2006, crowdsourcing has been used as a low cost and effective means of conducting large-scale studies [172]. The term *crowdsourcing* refers to the outsourcing of tasks to a large group of people with some kind of monetary incentive [173]. Within ISR, crowdsourcing has been successfully implemented across a variety of search tasks [173–175]. Several papers have examined the use of crowdsourcing

for relevance evaluation. Traditionally, relevance evaluation has been recognised as a cumbersome and expensive task and was often carried out by a small sample of individuals due to the time and resources involved [173]. However, several researchers examined the benefits of crowdsourcing as a means to gather large numbers of human-generated relevance judgement labels to establish IR evaluation collections [174]. An important consideration when using crowdsourced relevance judgement is ensuring the quality of the judgements. Alonso and Mizzaro [176] compared crowdsourced relevance judgements to TREC assessor judgements, finding moderate correlations between the two. Maddalena and colleagues [177] also propose that relevance judgements collected through crowdsourcing are reliable and when used for IR evaluation, are replicable. Through a review of the ISR literature, the following advantages of using crowdsourcing for relevance judgement were derived:

- Crowdsourcing can significantly reduce the time and cost to annotate a large test collection [172, 173]
- Crowdsourcing offers a relatively quick turnaround compared to laboratory-based studies, where a large volume of judgements can be generated in little time [173].
- Crowdsourcing offers the opportunity to reach a greater number of participants compared to a laboratory-based study.

Given the advantages highlighted in the ISR literature, the use of crowdsourced participants for the empirical work conducted in this thesis is well-justified.

5.5.1 Crowd-Sourcing Platform Details

The crowd-sourcing platform Prolific was used to recruit participants for all studies. Prolific was chosen for several reasons. Firstly, Prolific enables a pre-screening function, where specific demographics and cognitive profiles can be selected to align with the study need. Secondly, unlike other platforms, Prolific enforces relatively strict participant quality measures, making it less susceptible to fraudulent or inattentive responses. Finally, on Prolific it is easy to exclude or re-invite previous participants, which is useful for multi-phases studies like those undertaken in this thesis.

5.5.2 Participant and Technical Requirements

Participants were paid between a £7.20-£9.50 hourly rate, determined by the UK national minimum/national living wage at the time of the experiments. To fulfill eligibility for participation in the studies, we required that participants were:

- from the US or UK;
- had not participated in any of our previous studies.

Participants were recruited via Prolific as those living in the UK and US to ensure they were more likely native English speakers and likely familiar with the style and context of the TREC articles, which are based on U.S. news content from The Washington Post. These screening criteria were chosen to minimise any potential confounding effects on the effort and load measures. Participants were made aware of the length of the experiment prior to participating, so had the opportunity to reject the task if they felt it was too long.

In addition to the participant eligibility criteria, participants attempting the study were required to use either a desktop, laptop, or tablet device. The use of mobile devices were prohibited as the screen size was not considered large enough to avoid scrolling.

5.6 User Study Data

Qualtrics was chosen as the survey platform due to its robust functionality for designing complex, interactive tasks and its ability to handle randomisation, branching logic, and timed responses - all essential for the structure of the user studies. Additionally, its secure data handling and compatibility with Prolific made it a reliable and efficient tool for online experimental research.

In this section, the evaluation methods used to measure effort and load are described by dividing them into two distinct measurement categories: objective and subjective. In this thesis, subjective measures are considered as those which require some degree

of human judgement, and objective measures are those that provide an impartial measurement with a quantifiable outcome.

5.6.1 Objective Measures

Objective measurements were recorded solely from the interaction log data collected via Qualtrics. The following key objective measures were implemented:

- **Clicks:** The number of clicks issued by participants. This measure considered all of the clicks issued by the participant on the document judgement page.
- **Time:** The time taken to make a relevance judgement. This measure detailed the time a participant spent on the document judgement page. This was captured as the moment the judgement question appeared on the screen to the moment that they clicked the arrow to move to the next page.

Number of clicks have been operationalised in a number of ISR studies as a measure of user effort [24, 27, 60], where a greater number of clicks infers higher effort.

Time is also a frequently used measure within ISR, often used as a proxy for user effort and load [92, 93, 95, 178]. In the user studies conducted for this thesis, for each document judgement made by the participant, the system recorded the total dwell time on the page and also produced a timestamp. It could then be determined from this data how long the participant spent making the judgement, and also when the event occurred. As document ordering was randomised for each participant, the time stamp was valuable when examining document ordering effects.

It should be noted that objective measures collected from the **document judgement question page** only, were used in the analysis.

5.6.2 Subjective Measures

Subjective measures of effort and load were collected by the Qualtrics system via responses derived from both during-task and post-task questionnaires. The subjective measures employed across the studies were as follows:

During-Task

The following section details the subjective measures that were administered *during* the document judgement task. This allowed the examination of user cognitive load and effort at a more discrete level, and gain a fuller understanding of how effort and load may fluctuate as the task progresses.

- *Perceived Difficulty Scale:* Designed by Kalyuga and colleagues [179], this single-item scale is used as a measure of both cognitive load and effort. The scale requires individuals to rate their perceived difficulty on a 7-point Likert scale, ranging from 0= "very easy" to 7="very difficult". For this thesis research, participants rated their perceived difficulty on a visual analogue scale (VAS) rather than the traditional 7-point Likert scale. A VAS scale presents numbers on a continuum, from 0 (very easy) to 100 (very difficult) and participants rank their perceived difficulty response using a sliding bar. This type of scale was used for several reasons. Firstly, it has been identified as having both high test-retest reliability and minimal measurement error [180]. Secondly, using a 0-100 point scale provided consistency with the other subjective measures used, such as the NASA-TLX which also employs this type of scale. Therefore, providing standardisation across measurement scales to help reduce user uncertainty or error.

- *Subjective Rating Scale:* Developed by Paas and colleagues [181], this single-item measurement of effort has been shown to possess high validity and reliability, and is sensitive to very small differences in task complexity and design [182]. Individuals are asked to rate the amount of effort they invested during a task using a 7-point Likert scale, ranging from 0="very very low mental effort" to 7="very very high mental effort". For continuity with the measurements of difficulty and workload (i.e., NASA-TLX), a visual analogue scale was also employed for this measure, with participants rating their response between 0="very low mental effort" to 100="very high mental effort".

Post-Task

This section describes the subjective measure administered *post-task* - after all document judgements have been completed. The purpose of administering measures after the task completion is to firstly assess participants **overall** perception of the task, and secondly, to examine how this relates to the during-task measures.

- *NASA Task Load Index (NASA-TLX)*: Developed by Hart and Staveland [129], this subjective measure is used to assess participants perceived workload. The measure consists of six component scales (physical demand, mental demand, temporal demand, performance, frustration, and effort) which the individual rates on a 0-100 scale. The scores from each scale are then averaged to calculate the final overall score ranging from 0-100, referred to as the overall task load index.

5.6.3 Demographics

Several questionnaires were administered to participants prior to commencing the judgement task. These included a demographic questionnaire, and also questions pertaining to user individual characteristics.

General demographic information were collected about the participants using a demographic questionnaire, including: age; gender; highest level of education (from either Bachelors; Masters; Doctorate; None of the Above; Prefer not to say). Additionally, questions were included that related to the technical aspects of the task such as: the device used; the approximate screen size of device; and the type of pointing device used.

5.6.4 Individual Differences

While there are a large number of individual factors which may influence an individuals effort and load, the user studies conducted for this thesis focused on the following:

Working Memory: Previous research has proposed that individuals with high working memory capacity recall more information than those with low working memory capacity, and can also formulate better interpretations of the information presented [45].

Additionally, as task complexity increases, those with lower working memory capacity examine fewer result documents, whereas those with high working memory capacity continued performing more actions [68]. To test participants working memory capacity the Digit Span Memory Test (backwards and forwards) was employed. Within clinical psychology, this task, derived from the Weschler Adult Intelligence Scales [183], is considered one of the most prevalent approaches to working memory capacity assessment. The task asks individuals to view a series of digits and then repeat the series in the correct backward or forward order. The difficulty of the task is increased by increasing the number of digits presented in a series. In this thesis work, level of difficulty increased from 3-digits to 8-digits for both the forward and backward span tests. The test discontinued when the participant incorrectly recalled two series of digits consecutively. One point was allocated for each correct response. Participants were shown their score for this test at the end of the experiment.

Perceptual Speed: Previous research has suggested that individuals with high perceptual speed experience lower levels of workload during search tasks. [102]. Additionally, individuals with low perceptual speed are shown to have more difficulty scanning for relevant information (based on precision and recall) [102]. To test participants perceptual speed, the "Finding A's" test from *Ekstrom's kit of Factor Referenced Cognitive Tests* [184] was chosen as an apppropriate test. This test asks individuals to scan columns of words and select those which contain a letter "A". For this thesis work, participants were given six blocks of words, with each block containing 120 words that were shown for 20 seconds before changing to the next block. One point was allocated for each correct response, and participants were shown their score for this test at the end of the experiment.

Motivation: *How motivated they were to learn more about the topic.* Participants could select from: "Looks like it might be an interesting topic"; "Depends on the document"; "Looks like quite a boring topic". This measure was included to assess users' initial motivational orientation, particularly capturing aspects of intrinsic motivation. Prior ISR research has shown that intrinsic motivation, defined as the internal drive to engage in an activity out of personal interest or satisfaction, can influence user in-

volvement in a task [185]. It has also been associated with users' enjoyment and their willingness to acquire new knowledge, which may in turn affect the degree of mental effort exerted and the cognitive demands experienced during information-seeking activities. Measuring motivation, therefore, helps to account for a potentially confounding variable when interpreting results related to user effort and cognitive load.

Topic Knowledge: *How much they knew about the topic already.* Participants could select from: "Expert"; "Same as most people"; and "Almost nothing". Topic knowledge was measured as a potential confounding variable in the interpretation of effort and cognitive load. Prior research in ISR has shown that topic knowledge can influence relevance judgement outcomes. For instance, Ruthven et al. [143] found that assessors who self-identified as having a high level of topic knowledge were more likely to predict documents as relevant, but were less accurate in anticipating their final relevance decisions. Other studies have similarly suggested that topic knowledge can shape a searcher's strategy and influence the types of information they consider necessary [186]. These findings indicate that individual differences, such as topic knowledge, may interact with task performance and decision-making processes, and should therefore be accounted for in the analysis of effort and cognitive load.

Summary

Individual differences were examined pre-task and were included in the data analysis to examine the extent to which individual factors may influence user effort and load during the document judgement process. *Note* that each study varied in relation to which individual differences were examined. More details are provided in the chapters specific to each user study.

Chapter 6

User Studies

6.1 Overview of User Studies

The following three chapters (**Chapters 7, 8 and 9)** of this thesis present the three independent user studies (and one replication study) which were carried out for the empirical work of this thesis. When measuring constructs such as effort and load, it is important that the experimental design is tightly controlled and any confounding variables are mitigated. The user studies conducted for this thesis can be considered as a several step process of developing the most appropriate experimental method for examining effort and load in the context of multi-stage relevance judgement. For example, both within-subjects and between-subjects designs are tested, as is the extent to which individual differences such as working memory, perceptual speed, topic knowledge, and motivation which may influence results. Additionally, examining effort and load measures under different experimental conditions and levels of granularity provides broader insight into how these constructs vary depending on their measurement context. For instance, assessing effort and load at the level of individual document judgement stages (Study 2 - Chapter 8; Study 3 - Chapter 9) may produce different results compared to evaluating them holistically across all stages combined (Study 1 - Chapter 7). As each study is described in more detail, an explanation will be provided in relation to how the experimental design was revised from the previous and the decision making process behind this. This thesis proposes that the third and last user study, provides an

effective way of measuring effort and load within the context of multi-stage document judgement, and this proposal is further reinforced by the follow-up replication study which supports the original results.

6.2 Related Work and Motivation

All the user studies conducted for this thesis and presented in the following chapters seek to answer the third high-level research question outlined in Section 1.3, **HL-RQ3 (Empirical): How can effort and load be measured within the ISR sub-task of multi-stage document judgement, as defined by the theoretical model?**. Since all the studies seek to address the same main research question, the underlying body of literature is largely consistent across them. However, for clarity and in line with the slight variations in the sub-questions addressed across the studies, this section organises the literature in relation to each sub-question and its corresponding study (studies).

6.2.1 Relationships between Effort and Load Measures

All three studies (plus the replication study) conducted for this thesis examined the relationships between effort and load measures, and how they vary over the duration of the task. Specifically, they seek to answer the following sub-questions:

- (c) What are the operational relationships between effort and load?
- (d) How do measures of effort and load change over time during the task?

Previous ISR studies have employed multiple methods to measure effort and load within a single study [3, 60, 64, 68, 70, 73, 78, 80, 82, 110]; however, none have systematically examined the relationships between these measures. Failure to examine these relationships can undermine result validity, as conclusions may be based on isolated or contradictory findings. Understanding correlations between measures can clarify whether they assess the same construct or different aspects of it, or reveal the need for refinement in their definitions. Without this analysis, conflicting results may arise,

therefore reducing the relatability of the findings. Moreover, unexamined correlations could lead to redundant measures, wasting time and resources, especially when measures are highly correlated. To address these issues, each study employs at least three different measures of effort and load, both objective and subjective, and examines these relationships.

Our understanding of the relationships between effort and load at different stages of the relevance judgement process remains limited. The few studies which have examined effort and load in the context of document judgement have taken measurements following task completion only. Measuring constructs such as effort and load throughout a task, rather than just at the end, is important for several reasons. For instance, effort and load are not static, rather they tend to fluctuate as the task progresses. Measuring effort and load throughout allows researchers to capture these changes in real-time, providing a more accurate representation of how cognitive demand varies at different stages of the task. This can reveal how specific parts of a task are more taxing than others. Measuring effort and load at different intervals can also offer temporal insights, for example, researchers can associate specific moments in the task with a rise in cognitive load or effort. For example, a difficult decision point or a complex subtask may cause a temporary increase in cognitive load. Understanding when these peaks occur can help identify critical moments that may need re-design or improvements. Similarly, if effort or load is consistently high at certain points, task design can be adjusted to alleviate unnecessary strain, improving user performance and experience. Post-task measurements may also be vulnerable to recall bias or a lack of contextual relevance, where participants may not remember how demanding certain parts of the task felt. Measuring throughout the task helps mitigate this issue, capturing more precise and context-specific data. Overall, measuring effort and load both during the task and post-task, should offer a more nuanced, accurate understanding of cognitive load and effort, subsequently leading to improved task design, better performance outcomes, and deeper insights into the cognitive processes involved.

6.2.2 Effort and Load between Document Judgement Stages

Studies 2 and **3** (plus **replication study**) examine the extent to which effort and load differ between document judgement stages, seeking to answer the following research question:

- **(a) How do different stages of document judgement vary in terms of user effort and load?**

Few studies have examined effort and load within the context of document judgement tasks in ISR, and none have investigated these constructs within a multi-stage framework of relevance judgement. Existing work [73, 110] has typically focused on a single stage of judgement, most often whether a document is relevant or not. By treating relevance judgement as a single-stage process, these studies make it difficult to disentangle which specific aspects of the judgement task contribute to user effort and load.

As outlined in Chapter 4, document relevance judgement is inherently a multi-stage process, with different stages likely imposing varying levels of demand on the user. Cognitive load, in particular, is dynamic and expected to fluctuate not only across the duration of a task but also between stages of document judgement. For instance, the initial act of scanning a document may involve relatively low demand, whereas deeper evaluation of its *relevance* or *usefulness* may require substantially greater effort and cognitive resources.

Measuring effort and load at multiple stages therefore provides a more nuanced understanding of how demand varies across the judgement process. This approach also has practical value: identifying the stages that impose the greatest demand can inform the optimisation of search interfaces and task flows. For example, if cognitive load peaks during detailed relevance evaluation, features such as improved document previews or more effective result presentation could help alleviate this demand.

In summary, assessing effort and load across different stages of document judgement allows for a more comprehensive account of user experience, highlights critical points for task optimisation, and sheds light on how users navigate complex evaluative processes.

6.2.3 Effort, Load and Judgement Rating

Studies 2 and **3** (plus **the replication study**) also seek to examine how user effort and load are influenced by the type of judgement rating between document judgement stages, characterised by the following research questions:

- **(b) To what extent is effort and load influenced by the type of judgement rating ("no", "yes") between document judgement stages? (Study 2)**
- **(b) To what extent is effort and load influenced by the type of judgement rating ("no", "partially", "yes") between document judgement stages? (Study 3)**

In most relevance judgement tasks, the user is presented with a search topic or query and then a document. The user is usually then asked to make one type of relevance judgement, either a binary relevance assessment [73] (i.e., "relevant"/"not relevant") or a graded relevance assessment [110] (i.e., "not relevant"/"relevant"/"highly relevant"). Previous research examining document relevance judgements within ISR has suggested that the relevance rating of a document can impose varying levels of cognitive effort and load on the user. For example, Villa and Halvey [110] measured the amount of effort a user expends while making document relevance assessments, their results revealed that "relevant" judgements required more effort than "highly relevant" judgements. [104] Gwizdka also examined effort within the context of document relevance judgement, where users were asked to explicitly judge the perceived relevance of a document by marking "yes" or "no", while an eye-tracking device measured their level of cognitive effort. "Partially relevant" judgements were found to require higher levels of cognitive effort than "non-relevant" documents - differences in reading patterns and fixations also varied in relation to the perceived effort of the relevance assessment. Given the evidence presented, it is likely that the type of rating given at the *relevance* judgement stage of the multi-stage model will affect the users level of effort and load. As previously mentioned in Chapter 4, no studies within ISR have examined effort and load in the

context of other judgement stages (e.g., *aboutness*, *novelty*, *usefulness*). Consequently, it remains unclear how ratings at these stages may affect user effort and load.

6.2.4 User Characteristics

Finally, **Studies 1** and **3** (plus **replication study)** examine how user characteristics may influence effort and load measurement. They seek to answer the corresponding research questions:

- **(e) How do perceptual speed and working memory influence effort and load? (Study 1)**
- **(e) How do topic knowledge and motivation influence effort and load measures? (Study 3)**

Several cognitive abilities, such as working memory and perceptual speed, have been investigated within the context of ISR, with findings suggesting that user characteristics can have significant effects. Arguello and Choi [102] examined the effects of aggregated search interface condition (interleaved vs. blocked) and three cognitive abilities (working memory, perceptual speed, and inhibition) on participants reported levels of workload. Perceptual speed and inhibition did not have a significant impact on participants' perceived workload or user engagement; however, they did significantly influence search behaviours. Specifically, in the interleaved interface condition, participants with lower perceptual speed experienced greater difficulty in locating relevant results on the search engine results page, while participants with lower inhibitory attention control exhibited a slower search pace. Working memory had a limited impact on participants' behaviours, but it significantly influenced the reported levels of workload and user engagement. Specifically, participants with lower working memory capacity reported higher levels of workload and lower levels of user engagement. Brennan and colleagues [29] investigated the impact of individuals' cognitive abilities on their search behaviours and perceptions of workload during search tasks of varying complexity. They found that high perceptual-speed participants experienced lower levels of workload while completing search tasks and they also completed their search activity in less

time. Finally, Gwizdka [68] found that individuals with high working memory exerted more search effort and were less likely to satisfice in demanding situations. While these studies highlight the importance of cognitive abilities on user effort and load in various ISR processes, none of these studies have explored how these abilities influence user effort and cognitive load in the context of a document judgement task. Furthermore, no study to date has simultaneously examined the interrelationships between effort, load, perceptual speed, and working memory within a single research framework. Therefore, **Study 1** examines how users' different cognitive abilities influence the document judgement process, specifically in relation to their perception of workload and effort. By exploring whether variations in cognitive abilities affect the document judgement process, the research may inform the design of documents and search interfaces that are tailored to the cognitive strengths or limitations of individual users.

In addition to cognitive abilities, there are other user characteristics to consider which may influence user effort and load. **Study 3**, examines how user topic knowledge and motivation relate to user effort and load. Drawing on research from Educational Psychology, individuals who have existing topic knowledge are considered to comprehend more text and have an increased reading speed compared to individuals who are less familiar with the topic [187]. This understanding of topic familiarity and text comprehension is grounded in schema theory, which suggests that for those with topic knowledge, the schemata for that topic is activated more quickly and the information is interpreted faster than for unfamiliar topics [187]. Schema theory suggests that a large amount of topic knowledge is held in an individuals long-term memory as schemata, yet can be processed as as single entity in working memory [53], and can therefore effectively reduce working memory load. As cognitive load is concerned with the limitations of working-memory capacity, it can be hypothesised that having prior knowledge of a topic will reduce the cognitive load experienced by the individual when exposed to a document relating to that topic. Prior research in ISR has demonstrated that users' knowledge of a topic can influence how they assess relevance. For example, Ruthven et al. [143] found that participants who rated themselves as having high topic knowledge were more likely to predict relevance than non-relevance, yet were less accu-

rate in anticipating their final relevance judgements. Similarly, Wen et al. [186] showed that topic knowledge can affect users' search strategies and perceptions of the type of information needed. These findings highlight the importance of accounting for individual differences, such as topic familiarity, when examining user effort and cognitive load, as such factors may shape both behaviour and judgement during information-seeking tasks. However, there have been no studies to date which have examined the influence of topic knowledge on user effort and load during a multi-stage relevance judgement task.

Finally, another user characteristic which may influence a users effort, load, and task behaviour is motivation. In this thesis, motivation is treated as *intrinsic motivation* - defined as the internal desire to engage in an activity for its own sake, such as personal interest or satisfaction. Previous research in ISR has shown that intrinsic motivation can influence users' engagement with tasks [185]. It has also been linked to greater enjoyment and a stronger willingness to learn, which may impact the amount of mental effort invested and the level of cognitive load experienced during information-seeking. Furthermore, user perceptions of search task difficulty are found to be affected by cognitive ability and motivation [70, 98]. As such, measuring motivation provides a way to control for a potential confounding factor when analysing effort and cognitive load.

Chapter 7

Study 1: Pilot Testing the Model and Measures

7.1 Motivation

This chapter presents the first user study, which investigates effort and load measures within the context of multi-stage relevance judgement and explores the influence of individual differences, specifically working memory and perceptual speed. Serving as a pilot, the study was designed to test the measures, procedures, and document judgement task that would be employed in the subsequent full-scale studies. This was considered an important first step in ensuring the following studies were well-designed, efficient, and likely to produce valid and reliable results. The study was designed in two phases: first, evaluating participants' perceptual speed and working memory, followed by a second phase in which participants made four relevance judgements per document. These four relevance judgements relate to the *aboutness*, *relevance*, *novelty*, and *usefulness* of the document as outlined in the multi-stage relevance judgement model conceptualised in Chapter 4. Effort and cognitive load were measured during the task and after its completion, allowing analysis of the relationships between these factors.

The user study presented in this chapter seeks to answer the third high-level research question outlined in Section 1.3, **How can effort and load be measured within the ISR sub-task of multi-stage document judgement, as defined by the**

theoretical model?. More specifically, this study aims to answer the following sub-questions:

- (c) What are the operational relationships between effort and load?
- (d) How do measures of effort and load change over time during the task?
- (e) To what extent do user characteristics (perceptual speed and working memory) influence measures of effort and load?

It is important to note why only three of the five research sub-questions were addressed in this study. The first study was deliberately narrower, as it was intended to pilot the methodology, measures, and task design. Once these have been evaluated and validated, later studies were intended to widen the scope to address all research questions. Therefore, although Study 1 did not address all five sub-questions, the research questions were set from the outset, with the understanding that they would be addressed progressively across iterative studies

7.2 Methodology

The experimental design used in this user study was narrow and controlled, consisting of two cognitive ability tests (working memory and perceptual speed) and a document judgement task. The study followed a **within-subjects** design, where participants answered **all** four document judgement questions corresponding to the four stages (*aboutness*, *relevance*, *novelty*, and *usefulness)* outlined in the multi-stage relevance judgement model (Chapter 4 - Figure 4.1). Before commencing the document judgement task, participants were asked to complete two cognitive ability tests to assess their working memory and perceptual speed ability. For the document judgement task, **one** topic (Wildlife Extinction) was selected from the TREC 2018 Common Core Track. Ten documents were selected from the document collection, five relevant and five non-relevant. Each document was manually edited to all have the same length, font size, and font to limit the effects of any confounding extraneous factors on the dependent variables. After each document judgement, participants rated the effort and difficulty of making

that specific judgement using two single-item scales. After completing judgements for all ten documents, participants were asked to complete the NASA-TLX in order to assess their workload for the judgement task overall. As participants judged the documents, Qualtrics software gathered log data (judgement time) and survey responses. Figure 7.1 shows an example of the document judgement page, Figure 7.2 shows an example of the effort and load question page, and Figure 7.3 shows an example of the NASA-TLX question page, used in the experimental system for study 1.

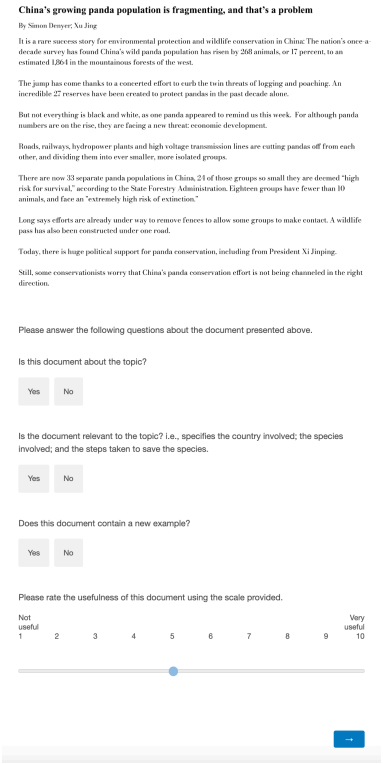

Figure 7.1: Screen shot of the experimental system used in Study 1. This figure illustrates the document judgement page.

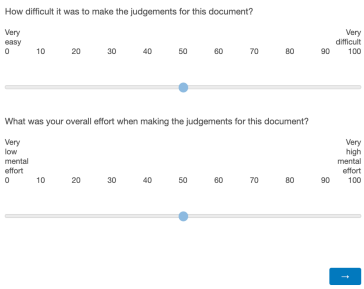

Figure 7.2: Screen shot of the experimental system used in Study 1. This figure illustrates the single-item effort and load question page.

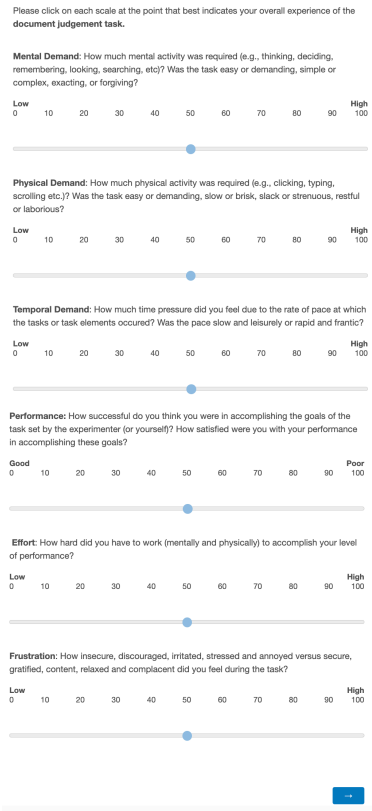

Figure 7.3: Screen shot of the experimental system used in Study 1. This figure illustrates the NASA-TLX question page.

7.3 Measures - IVs & DVs

The independent variables in this study were as follows:

- *Document Judgement Rating* (2 levels: no/yes)
- *Perceptual Speed Ability* (2 levels: high/low)
- *Working Memory Ability* (2 levels: high/low)

The dependent variables used in this user study were as follows:

- *Effort*: as measured by the single-item *Subjective Rating Scale*.
- *Difficulty*: as measured by the single-item *Perceived Difficulty Scale*.
- *Overall workload*: as measured by the NASA-TLX.
- *Judgement time*: gathered by Qualtrics data logs, and measured as the time spent on each document judgement page.

7.4 Procedure

This study obtained ethics approval from the University of Strathclyde's Department of Computer and Information Sciences (Approval No. 1661). Participants were recruited via the crowd-sourcing platform Prolific where participants were paid at a £7.20 hourly rate. The experimental procedure for this study is more generally outlined in Section 5 and in Figure 5.1, however, *note* that rather than using pre-task questionnaires, this study asked participants to complete two cognitive ability tests (working memory and perceptual speed) prior to commencing the document judgement task. The study took participants approximately 30 minutes to complete. Following the online experimental briefing and obtaining participants consent, the study proceeded as follows:

1. Demographic questionnaire (3 minutes)
2. Two cognitive ability tests (10 minutes)
3. Task description (1 minute)
4. Document judgement task - ten documents with in-task single-item effort and difficulty ratings (13 minutes)

5. Post-task workload questionnaire (3 minutes)

Prior to commencing the task, participants were introduced to the topic description. During the task, each full document was displayed at the top of the page with all four document questions positioned underneath. The order in which the ten documents were presented was randomised by the Qualtrics system for each participant to mitigate ordering effects. The four document judgement questions used were as follows:

- *Aboutness*: "Is the document about the topic...?"
- *Relevance*: "Is the document relevant to the topic?"
- *Novelty*: "Does the document contain a new example?"
- *Usefulness*: "Please rate the usefulness of this document using the scale provided"

The *aboutness, relevance* and *novelty* document judgement questions required a binary "yes" or "no" response. However, for the *usefulness* question, participants were required to respond using a sliding scale from 1 (not very useful) to 10 (very useful).

Following the judgement rating, participants were required to click an arrow to direct them to the following page which contained the follow-up difficulty and effort self report questions. This was repeated for all ten documents. Participants could not revisit previously viewed documents and there was no time limit imposed in relation to task completion, however participants were encouraged to perform the task as quickly and accurately as possible. After completing all ten documents, participants were directed to the NASA-Task Load Index, where they were asked to rate their perceived overall workload for the task.

Note that *usefulness* was measured on a 10-point scale, whereas effort, load, and overall workload were measured on a 100-point scale. A 10-point scale was used for usefulness to support straightforward comparative judgements about the documents. In contrast, effort and cognitive load are not document judgement questions but complex cognitive constructs that are known to vary incrementally across tasks and thus require higher precision to measure accurately. For this reason, effort and load were assessed

on a 100-point scale to provide greater sensitivity to subtle differences in participants' subjective workload. This approach is consistent with established workload-assessment practices (e.g., NASA-TLX), which recommend higher-resolution scales for capturing fine-grained variations in cognitive demands [129].

For more details about each of the measures used, please refer to Section 5.6.

7.5 Demographics

Overall, 55 participants completed the study: 24 females and 31 males. Participants were mostly aged between 18-39 (N= 50) with the remaining participants aged between 40-79 (N=5). The majority of participants held a Bachelor's level qualification or above (N=44). Prolific screening criteria was set to ensure participants were recruited from the UK and US. Participants who completed less than 90% of the study were removed from the final data set. A total of 479 annotations were used in the final analysis.

7.6 Results and Discussion

In this study, user effort and load were examined in the context of a relevance judgement task. The outcome of this study provided an enhanced insight as to how effort and load may vary across the duration of the task, as well as the highlighting relationships between different constructs, and how individual differences can further influence these constructs. Subsequently, this work adds to previous work which has mainly focused on examining effort and load in a *post-task* context. The results of this study help to answer the third high-level research question presented in section 1.3 of this thesis: **How can effort and load be measured within the ISR sub-task of multi-stage document judgement, as defined by the theoretical model?**, and more specifically, the following sub-questions:

- (c) What are the operational relationships between effort and load?
- (d) How do measures of effort and load change over time during the task?

- (e) To what extent do user characteristics (perceptual speed and working memory) influence measures of effort and load?

The following sections present the study results and discuss the implications of these in relation to the research questions outlined above. Note that the term "participants" is used in the methods sections to emphasise their role in the study procedures, consent, and experimental protocol, whereas "users" is used in the results sections to focus on their interactions and behaviours while interacting with the document judgement task.

7.6.1 Relationships between Measures

This section examines the operational relationships between the measures used **(HL-RQ3c)**. A Spearman correlation analysis was performed using the Bonferroni correction for multiple comparisons. Results of this analysis showed moderate correlations between effort and difficulty, rs = .54, $p<.05$, Figure 7.4 illustrates this relationship. Significant but relatively weak relationships were observed between difficulty and performance, rs = .29, $p<.05$; and temporal demand and judgement time, rs = .12, $p<.05$.

To examine the relationship between in-task and post-task measures, the minimum and maximum values of effort and difficulty were compared with the NASA-TLX dimensions and overall perceived workload. A significant, though relatively weak, correlation was found between maximum difficulty and performance, rs = .29, $p<.05$.

Table 7.1 shows the relationships between each of the dependent variables.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1. Minimum Effort	_													
2. Mean Effort	.93*	_												
3. Maximum Effort	.73*	.89*	_											
4. Minimum Difficulty	.43*	.36*	.18*	_										
5. Mean Difficulty	.45*	.54*	.48*	.80*	_									
6. Maximum Difficulty	.39*	.57*	.64*	.41*	.77*	_								
7. Mental Demand**	.05	.08	.09	.01	.06	.11	_							
8. Physical Demand**	.06	.02	-.01	-.04	-.12	-.22	.00	_						
9. Temporal Demand **	.05	.07	.02	.04	.06	.04	.40*	.35*	_					
10. Performance**	.12	.14	.13	.13	.29*	.29*	.23*	-.34	.07	_				
11. Frustration**	-.09	-.06	-.01	.02	-.02	-.04	.18*	.22*	.32*	-.20	_			
12. Effort**	-.01	.03	-.04	.05	.07	.01	.67*	.02	.32*	.22*	.27*	_		
13. Overall Workload**	.12	.15	.12	.12	.15	.08	.67*	.42*	.73*	.21*	.57*	.65*	_	
14. Judgement Time	-.01	-.01	-.04	-.12	-.12	-.03	.05	.14	.12*	-.10	-.03	.06	.05	_

Table 7.1: Means (M), Standard Deviations (SD), and Spearman correlation matrix for dependent variables (n=55) *p<.05 ** NASA-TLX Dimensions

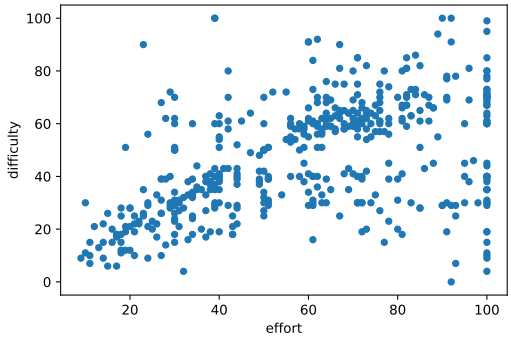

Figure 7.4: Relationship between effort and difficulty

Due to the wide variety of measures used to examine effort and load within ISR, this study incorporated several measures to gain a fuller understanding of the relationships between them and the extent to which they are measuring the *same* construct. In this study, single-item measures of effort and difficulty showed the strongest correlation, albeit a moderate one, this finding is consistent with previous work from outside ISR [188]. More surprisingly, the relationships between the remaining measures were either very weak or non-significant. Generally, these findings suggest that the different measures of effort and load which are widely incorporated within ISR studies may be measuring different constructs, and somewhat weakens support for the use of theoretically weaker measures such as judgement time as a proxy for cognitive load. Interestingly, the NASA-TLX scores did not correlate with the continuously recorded single-item measures of effort and difficulty. Given that the NASA-TLX was completed after the task, it is unclear what aspects of the experience users were basing their ratings on. Unlike the single-item measures, which captured moment-to-moment perceptions of effort and difficulty for each document, the NASA-TLX appears to reflect a more global, retrospective assessment of workload. This discrepancy suggests that post-task workload measures may not accurately capture the fine-grained cognitive and behavioural demands experienced during specific task stages. Consequently, care should be taken when interpreting NASA-TLX scores as indicators of task-specific effort or difficulty, particularly in studies where moment-to-moment variations are of interest. These findings underscore the importance of considering the timing and fo-

Document Order	Judgement Time (sec)			Effort			Difficulty		
	M	**Mdn**	**SD**	**M**	**Mdn**	**SD**	**M**	**Mdn**	**SD**
1st	66.89	48.05	51.47	61.22	61.0	26.55	45.76	45.0	23.31
2nd	55.68	40.90	40.85	61.57	60.5	23.98	47.20	42.0	19.94
3rd	60.56	44.53	49.69	61.72	67.5	25.76	48.46	56.0	22.38
4th	50.46	38.81	41.24	58.35	64.0	26.34	46.37	43.0	22.68
5th	58.61	40.91	67.54	59.98	60.0	24.84	48.00	45.0	20.80
6th	78.86	50.04	140.46	58.84	61.0	23.74	49.80	48.0	20.47
7th	58.61	36.87	60.20	58.04	60.0	25.27	45.88	40.0	21.04
8th	45.34	24.12	56.92	58.23	63.0	26.74	44.70	40.0	22.21
9th	43.85	30.29	41.11	56.96	58.0	26.83	41.17	39.0	18.93
10th	37.61	29.44	34.09	57.53	59.0	28.18	48.53	55.0	21.18

Table 7.2: Mean (M), Median (Mdn), and Standard Deviations (SD) for judgement time, effort, and difficulty by document order

cus of workload measures when designing experiments and interpreting user experience data.

7.6.2 Effort and Load across Task Duration

This section examines how effort and load measures vary across over time, or in other words, across the duration of the document judgement task **(HL-RQ3d)**. As Levene's Test showed non-significant differences and the dependent variables mostly followed a normal distribution, an ANOVA test was used to analyse differences in effort and load for document ordering. The ANOVA revealed no significant main effects for effort, difficulty, or judgement time across the duration of the task. Table 7.2 shows the descriptive statistics for all measures per document judgement ordering.

As noted in Chapter 4, many ISR studies assess effort and load only after task completion. Expecting these constructs to fluctuate, ratings were collected after each document was judged. Contrary to expectations, effort and load did not vary significantly across the task. This may reflect limitations in measure sensitivity, particularly for constructs like cognitive load, or that the task was not demanding enough to elicit changes. The results suggest that, for tasks of similar complexity, single-point post-task measures may reasonably capture overall effort and load, though they may miss

moment-to-moment variations.

7.6.3 Perceptual Speed and Working Memory

This section investigates the extent to which individual differences such as perceptual speed and working memory influence user effort and cognitive load **(HL-RQ3e)**. As Levene's test showed significance for homogeneity of variance for both working memory and perceptual speed, non-parametric Kruskal Wallis H tests were used, alongside Games-Howell post-hoc tests for pairwise comparisons. Before performing these tests, participants were grouped into two groups (high and low) for each cognitive ability using a median split of their score. This method of grouping participants into high/low groups has been demonstrated in previous studies which examine the effects of different cognitive abilities on user behaviour [102]. For perceptual speed, participants were divided at a median score of 39, where any score below or equal to 39 was considered a "low" perceptual speed ability, and anything equal to or above 39.1 was considered a "high" perceptual speed ability. For working memory, participants were divided at a median score of 12, where any score below or equal to 12 was considered a "low" working memory ability, and anything equal to or above 12.1 was considered a "high" working memory ability. Table 7.3 shows the descriptive statistics for each cognitive ability group.

For working memory, significant differences were observed between high and low ability groups for: effort (See Figure 7.5), $H(1) = 12.19$, $p<.05$; and judgement time (See Figure 7.6), $H(1) = 8.33$, $p<.05$; and for the following NASA TLX dimensions: overall workload (See Figure 7.7), $H(1) = 10.35$, $p<.05$; temporal demand (See Figure 7.7), $H(1) = 23.08$, $p<.05$; performance (See Figure 7.8), $H(1) = 14.40$, $p<.05$; and mental demand (See Figure 7.8), $H(1) = 9{,}07$, $p<.05$.

Post-hoc tests showed that although those with high working memory ability exerted greater effort and reported lower performance, users in this group report less overall workload, temporal demand, mental demand, and took less time to judge documents, than users with low working memory ability.

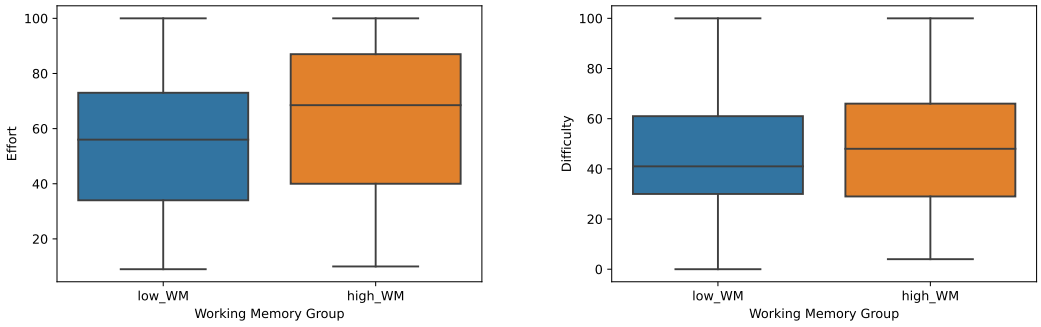

Figure 7.5: Effort and difficulty for low and high working memory groups

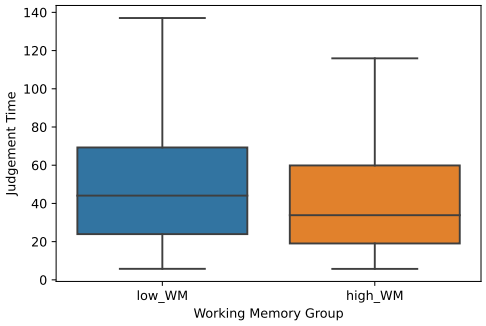

Figure 7.6: Judgement time for low and high working memory groups

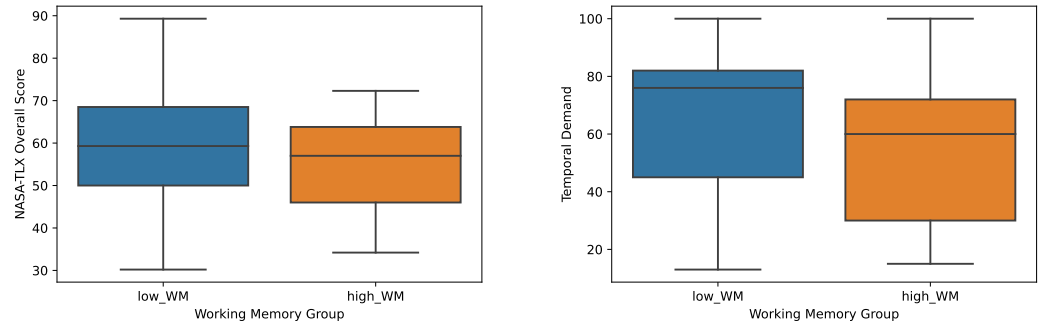

Figure 7.7: Overall workload and temporal demand for low and high working memory groups

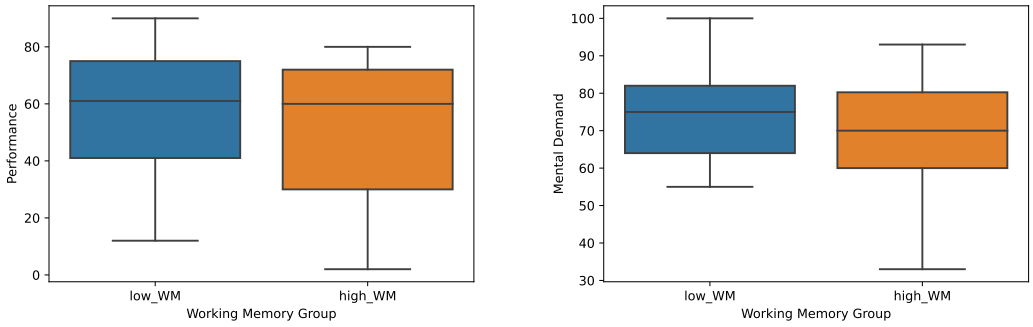

Figure 7.8: Performance and mental demand for low and high working memory groups

For perceptual speed, significant differences were observed between high and low ability groups for: effort (See Figure 7.9), $H(1) = 36.78$, $p<.05$; and for the following NASA-TLX dimensions: effort (See Figure 7.10), $H(1) = 7.93$, $p<.05$; mental demand (See Figure 7.10), $H(1) = 4.16$, $p<.05$; physical demand (See Figure 7.11), $H(1) = 19.47$, $p<.05$; and frustration (See Figure 7.11), $H(1) = 8.46$, $p<.05$.

Post-hoc tests showed that while those with high perceptual speed ability exerted greater effort during the task, and reported higher overall effort and mental demand post-task, these individuals experienced less physical demand and frustration compared to users with low perceptual speed ability.

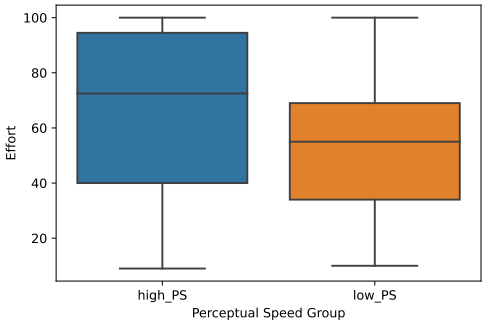

Figure 7.9: Effort for low and high perceptual speed groups

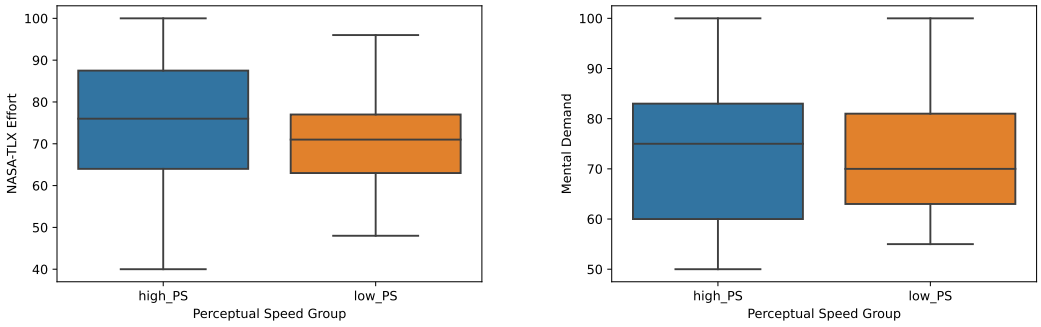

Figure 7.10: NASA-TLX effort and mental demand for low and high perceptual speed groups

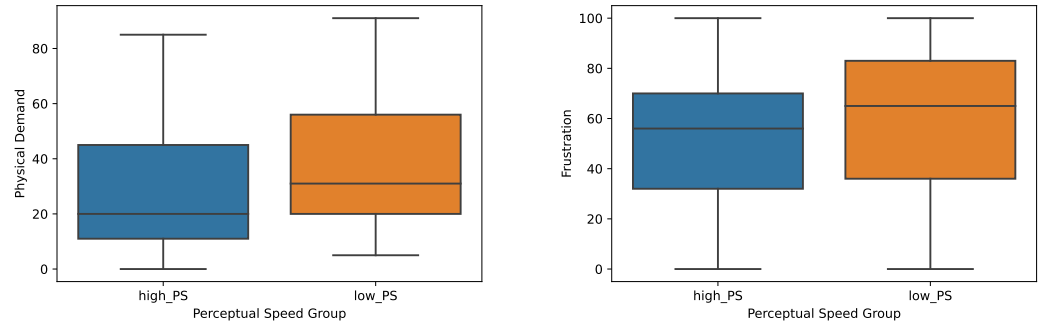

Figure 7.11: Physical demand and frustration for low and high perceptual speed groups

Group	High WM			Low WM			High PS			Low PS		
	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD
Judgement Time (sec)	49.58	33.84	50.61	59.99	44.09	73.37	50.28	38.19	46.57	60.43	39.61	77.78
Effort	63.94	68.5	27.17	55.75	56.0	23.99	66.69	72.5	28.73	52.32	55.0	20.22
Difficulty	48.40	48.0	23.46	45.23	41.0	19.43	48.35	47.0	23.55	44.94	41.0	18.82
Overall Workload*	54.91	57.0	11.37	59.32	59.3	14.12	56.85	57.0	11.62	57.54	58.80	14.32
Effort*	69.23	72.0	18.05	71.07	74.0	20.76	71.85	76.0	20.66	68.43	71.0	18.07
Mental Demand*	68.12	70.0	19.00	74.02	75.0	16.22	72.72	75.0	19.15	69.48	70.0	16.34
Physical Demand*	32.66	29.0	23.53	33.78	24.0	23.74	28.98	20.0	22.32	37.58	31.0	24.17
Temporal Demand*	54.55	60.0	24.91	63.94	76.0	26.72	59.54	65.0	26.51	59.10	68.0	26.02
Frustration*	54.94	63.0	28.35	54.39	56.0	26.74	52.40	56.0	25.59	56.98	65.0	29.23
Performance*	49.95	60.0	24.67	58.72	61.0	18.67	55.61	64.0	24.47	53.17	50.0	19.67

Table 7.3: Mean (M), Median (Mdn), and Standard Deviations (SD) for dependent variables for each working memory (WM) and perceptual speed (PS) group. *NASA-TLX dimensions

Previous work from ISR has suggested that individual differences between users can affect the amount of effort exerted or load experienced when completing a task [29, 80].

To investigate this area further, we examined whether an individual's working memory or perceptual speed abilities had the potential to influence their effort and load during a document judgement task.

The results indicate that individuals with high working memory (WM) capacity exerted greater effort during search tasks and demonstrated lower performance, yet reported reduced perceptions of overall workload, temporal demand, and mental demand. They also spent less time judging documents compared to their low WM counterparts. These findings align with Gwizdka [99], who observed that high WM users complete search tasks more quickly and perform a greater number of actions during sessions, suggesting a capacity for sustained cognitive engagement that may account for the increased effort observed. Similarly, Arguello and Choi [102] found that high WM users reported lower subjective workload, reinforcing the idea that cognitive resources may buffer the perceived strain of complex tasks. Gwizdka [68] also noted that high WM individuals exert more effort under demanding conditions, likely due to a reduced tendency to satisfice.

For perceptual speed (PS), users with higher PS exhibited greater in-task effort and post-task mental demand, yet reported lower physical demand and frustration. These mixed results suggest that while high PS users may engage more deeply with the task cognitively, their fluency in visually processing information reduces the affective or physical strain of the experience. Arguello and Choi [102] support this interpretation, showing that low PS users struggle more with visually complex layouts, likely contributing to their higher frustration levels. Conversely, Brennan and colleagues [29] found that high PS users experienced lower overall workload, raising the possibility that the relationship between PS and perceived task demand may be context or task-dependent.

Together, these findings and the findings from previous studies highlight the need to examine cognitive abilities in relation to specific task demands. The varying influence of WM and PS across different dimensions of effort and workload suggests that aggregated measures may mask important nuances. It also further reinforces the importance of measuring effort and load at varying task stages (i.e., during and post-hoc).

7.7 Conclusion

Overall, this study revealed that aside from the relationship between the single-item measures of effort and difficulty, the other measures such as judgement time and workload were mostly weakly correlated with one another. Moreover, as users progress through the document judgement task, it seems that the effort exerted and load experienced does not significantly change. The findings did however show, that individual differences such as working memory and perceptual speed, have a significant influence on user effort and load.

Chapter 8

Study 2: Between-Subjects Evaluation of Model and Measures

8.1 Motivation

The second user study conducted for this thesis aims to build on the insights gained from the first (pilot) study, addressing any identified limitations or gaps and refining the research design. By leveraging the lessons learned from the pilot study, Study 2 seeks to:

1. **Exclude Cognitive Ability Testing:** The perceptual speed and working memory tests employed in Study 1 were found to be mentally demanding, as indicated by participant feedback. Requiring participants to complete such a cognitively demanding task prior to the document judgement task may have affected the effort and cognitive load they experienced, given that factors such as fatigue and boredom can significantly influence these constructs. While the influence of cognitive abilities on user effort and cognitive load is acknowledged, it was determined that including such ability tests in the experimental design would no longer be appropriate, given their potential to confound the measurement of effort and load.

2. **Test Multi-Stage Model with New Approaches:** In the pilot study, effort and load were measured more granularly than in previous research, with ratings collected after each document rather than only at the end of the task (i.e., after all documents). However, because all judgement questions were presented on the same page, it remained difficult to determine whether specific questions imposed greater demands on participants, limiting the ability to detect subtle differences in cognitive load and effort across stages, as hypothesised in the multi-stage judgement model (Chapter 4). To more precisely test the hypothesis that intrinsic load increases as users progress through the model, it would be necessary to collect effort and load ratings at an even finer level of granularity, after each individual document judgement question, beyond the approach used in Study 1.

3. **Refine Methodology:** Participant feedback following the pilot study indicated that some of the wording in the document judgement questions was somewhat ambiguous and open to multiple interpretations. Additionally, the use of a different rating scale for the *usefulness* question caused confusion and complicated the analysis, particularly in determining which positions on the sliding scale should be classified as "useful" versus "not useful". As a result, the *usefulness* rating scale was modified in this study to align with the rating scales used for the other three questions.

The user study presented in this chapter seeks to address the limitations from Study 1 outlined above and also to answer the second high-level research question outlined in Section 1.3.

HL-RQ3 (Empirical): How can effort and load be measured within the ISR sub-task of multi-stage document judgement, as defined by the theoretical model?

- (a) How do different stages of document judgement vary in terms of user effort and load?
- (b) To what extent is effort and load influenced by the type of judgement rating ("no", "yes") between document judgement stages?

- (c) What are the operational relationships between effort and load?
- (d) How do measures of effort and load change over time during the task?

To summarise, Study 2 was designed as a methodological and theoretical extension of the pilot (Study 1). While the pilot demonstrated the feasibility of using the multi-stage relevance judgement task to capture effort and load, its design limited the ability to attribute differences in demand to specific judgement stages. In particular, effort and load ratings were collected at the document level, and all judgement questions were presented on a single page, making it difficult to isolate the contribution of each stage. To address these limitations, Study 2 increased the granularity of measurement by collecting effort and load ratings after each individual judgement question, and the task design was refined to separate the stages more clearly. These changes enabled a more rigorous test of the theoretical model outlined in Chapter 4, specifically the hypothesis that intrinsic load increases as users progress through the multi-stage judgement process. In this way, Study 2 not only refined the methods trialled in Study 1 but also broadened the scope of inquiry to address additional research questions, thereby providing a stronger empirical foundation for the subsequent studies.

8.2 Method

In this section, the methodology used in this user study complements the general user study methodology (Chapter 5). To help facilitate understanding of the methods used, reference will be made to the relevant section(s) in the general methodology.

The experimental design used in this user study was narrow and controlled, consisting of a document judgement task only. The study followed a **between-subjects** design, where participants were assigned to answer only **one** document judgement question for each of the ten documents. The document judgement questions correspond with the four judgement stages (*aboutness*, *relevance*, *novelty*, and *usefulness*) outlined in the model of multi-stage relevance judgement in Chapter 4 of this thesis. For this study, **one** topic (Wildlife Extinction) from the TREC 2018 Common Core Track was selected (the same topic was used in Study 1). Ten documents were selected from each topic

collection, five relevant and five non-relevant. Each document was manually edited to all reflect the same length, font size, and font to limit the effects of any confounding extraneous factors on the dependent variables. After each document judgement, participants were asked to rate their level of effort and difficulty on two-single item scales. After completing judgements for all ten documents, participants were asked to complete the NASA-TLX in order to assess their workload for the judgement task overall. As participants judged the documents, Qualtrics software gathered log data (judgement time) and survey responses. Figure 8.1 shows an example of the document judgement page (for the *novelty* condition), Figure 8.2 shows an example of the effort and load question page, Figure 8.3 shows an example of the NASA-TLX question page, for the experimental system used in study 2.

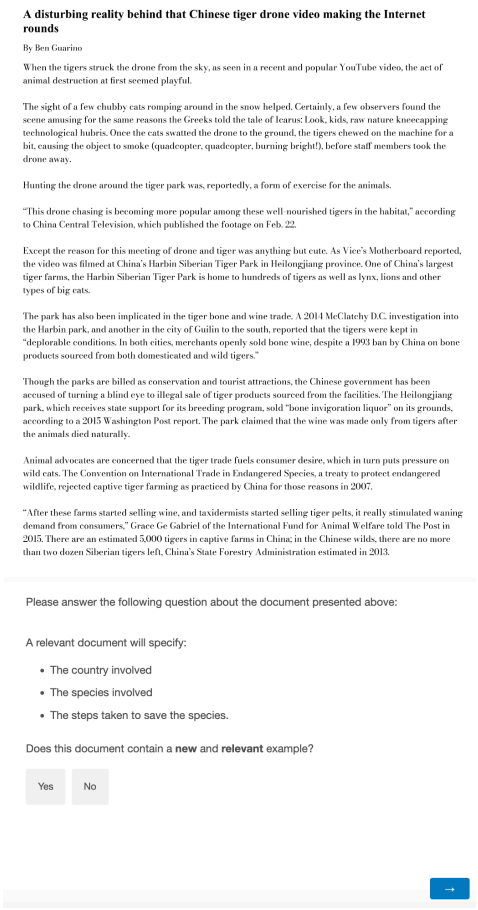

Figure 8.1: Screen shot of the experimental system used in study 2. This figure illustrates the document judgement page: *novelty* condition.

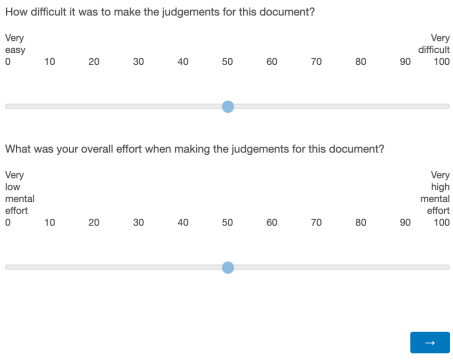

Figure 8.2: Screen shot of the experimental system used in study 2. This figure illustrates the effort and load question page.

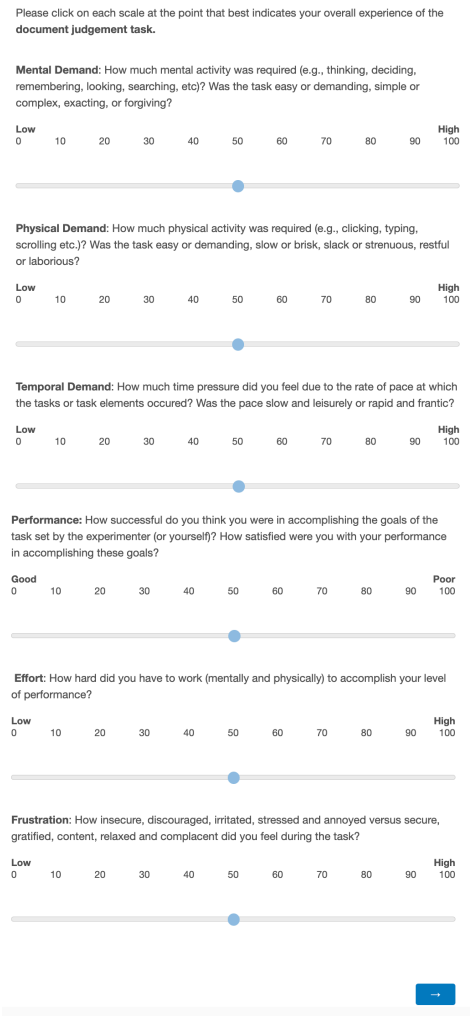

Figure 8.3: Screen shot of the experimental system used in study 2. This figure illustrates the NASA-TLX question page.

8.3 Measures - IVs & DVs

The independent variables used in this study were as follows:

- *Document Judgement Stage* (4 levels: *aboutness*/*relevance*/*novelty*/*usefulness*)
- *Document Judgement Rating* (2 levels: no/yes)

The dependent variables used in this study were as follows:

- *Effort*: as measured by the single-item *Subjective Rating Scale*.
- *Difficulty*: as measured by the single-item *Perceived Difficulty Scale*.
- *Overall workload*: as measured by the NASA-TLX.
- *Judgement time*: gathered by Qualtrics data logs, and measured as the time spent on each document judgement page.

8.4 Procedure

This study obtained ethics approval from the University of Strathclyde's Department of Computer and Information Sciences (Approval No. 1661). Participants were recruited via the crowd-sourcing platform Prolific where participants were paid at a £8.25 hourly rate. The experimental procedure for this study is generally outlined in Section 5.4 and in Figure 5.1 , however, *note* that pre-task questionnaires were not included in this study. The study took participants approximately 15 minutes to complete. Following the online experimental briefing and obtaining participants consent, the study proceeded as follows:

1. Demographic questionnaire (3 minutes)
2. Task description (1 minute)
3. Document judgement task - ten documents (8 minutes)
4. Post-task workload questionnaire (3 minutes)

Prior to commencing the task, participants were introduced to the topic description. During the task, each full document was displayed at the top of the page with one question underneath. The order in which the ten documents were presented was randomised by Qualtrics for each participant to mitigate ordering effects. Participants were also randomly allocated by Qualtrics to **one** judgement question condition from the following four judgement questions:

- *Aboutness*: "Is the document about the topic...?"

- *Relevance*: "Is the document relevant to the topic?"
- *Novelty*: "Does the document contain a new and relevant example?"
- *Usefulness*: "Does this document contain a relevant example that would be useful for your case study?"

Each document judgement question required a binary yes or no response. Following the judgement rating, participants were required to click an arrow to direct them to the following page which contained the follow-up difficulty and effort self report questions. This was repeated for all ten documents. Participants could not revisit previously viewed documents and there was no time limit imposed in relation to task completion, however participants were encouraged to perform the task as quickly and accurately as possible. After completing all ten documents, participants were directed to the NASA-Task Load Index, where they were asked to rate their perceived overall workload for the task.

For more details about each of the measures used, please refer to Section 5.6.

8.5 Demographics

Overall, 204 participants completed the study: 130 females, 72 males, and 2 non-binary/third gender. Participants were mostly aged between 18-39 (N=107) with the remaining participants aged between 40-79 (N=97). Prolific screening criteria was set to ensure participants were recruited from the UK and US, and 194 participants identified English as their first language. Participants were randomly allocated to one of the four document judgement question condition groups which consisted of the following participant numbers: ***aboutness***: (N=49); ***relevance***: (N=50); ***novelty***: (N=53); ***usefulness***: (N=52). Participants who completed less than 90% of the study were removed from the final data set. A total of 1933 annotations were used in the final analysis

8.6 Results and Discussion

In this second user study, user effort and cognitive load was investigated after *each document judgement* in the context of a relevance judgement task. This study built on Study 1 by investigating how the four document judgement stages differ in relation to effort and load, while also enhancing our understanding of commonly used measures of effort and load from within ISR. Furthermore, this work builds on previous research from ISR, which tends to focus on effort and load in the context of a single relevance judgement question. The results of this study further help to answer the third high-level research question: **HL-RQ3 (Empirical): How can effort and load be measured within the ISR sub-task of multi-stage document judgement, as defined by the theoretical model?** and more specifically, the following sub-questions:

- (a) How do different stages of document judgement vary in terms of user effort and load?
- (b) To what extent is effort and load influenced by the type of judgement rating ("no", "yes") between document judgement stages?
- (c) What are the operational relationships between effort and load?
- (d) How do measures of effort and load change over time during the task?

The next sections will present the results of Study 2 and discuss their implications in relation to the research questions above. Note that the term "participants" is used in the methods sections to emphasise their role in the study procedures, consent, and experimental protocol, whereas "users" is used in the results sections to focus on their interactions and behaviours while interacting with the document judgement task.

8.6.1 Effort and Load between Document Judgement Stages

This section examines how effort and load differed between each of the four document judgement stages **(HLRQ3a)**. As much of the data analysed showed significant differences for Levene's Test, the non-parametric statistical Kruskal-Wallis H-test was used,

with pairwise comparisons made using Games-Howell post-hoc tests. Results showed that there were significant differences between document judgement stages in relation to all of the dependent variables: effort, H (3) = 53.73, p<.05; document judgement time, H(3) = 84.92, p<.05; difficulty, H (3) = 53.16, p<.05; and overall workload, H (3) = 10.79, p<.05. Figures 8.4 and 8.5 illustrate the differences between stages for effort, difficulty, judgement time, and overall workload.

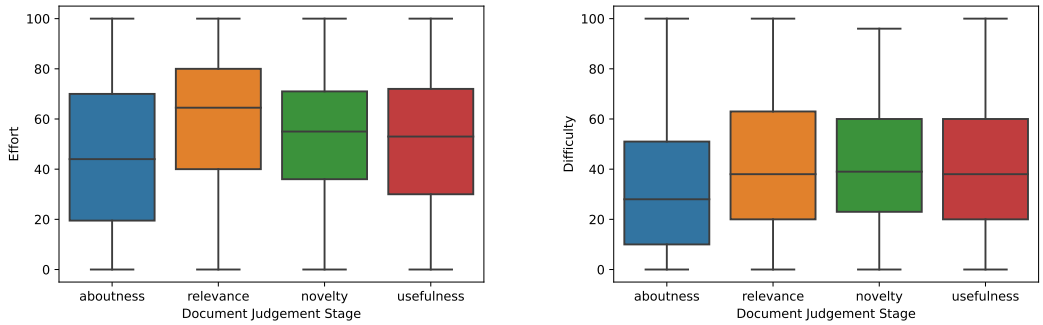

Figure 8.4: Effort and difficulty between document judgement stages

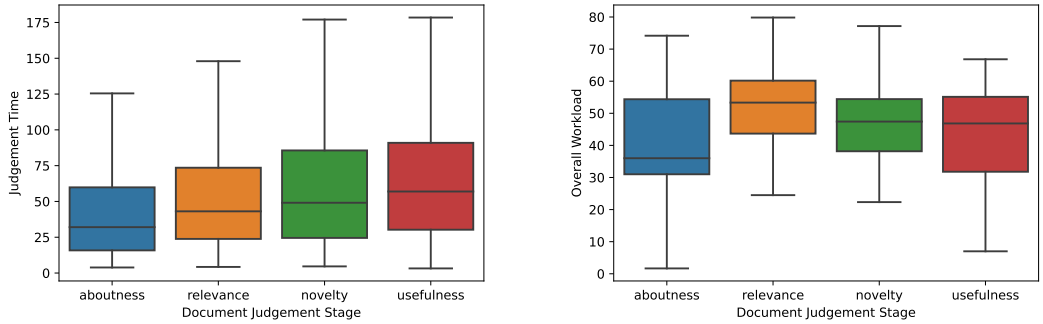

Figure 8.5: Document judgement time and workload between document judgement stages

Post-hoc comparisons showed that users exerted significantly more effort for *relevance* judgements compared to *aboutness*, *novelty*, and *usefulness*. Users took significantly less time to make *aboutness* judgements and found these judgements less difficult than the three remaining judgement stages. For overall workload, post-hoc tests showed no significant differences between each of the four judgement stages. Table 8.1 details the descriptive statistics for each of the dependent variables per stage of document

Judgement Stage	Judgement Time (s)			Effort			Difficulty			Overall Workload		
	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD
Aboutness	47.86	32.04	60.95	45.76	44.0	30.12	31.18	28.0	24.66	40.43	36.00	17.3
relevance	59.07	43.09	58.67	58.28	64.5	27.67	41.24	38.0	26.55	49.78	53.33	18.07
Novelty	65.53	49.08	69.68	52.80	55.0	22.73	40.67	39.0	22.68	48.12	47.42	13.42
Usefulness	77.18	56.94	97.29	50.39	53.0	27.03	39.99	38.0	25.23	42.56	46.84	15.81

Table 8.1: Mean (M), Median (Mdn), and Standard Deviations (SD) for judgement time, effort, difficulty, and overall workload by document judgement stage.

judgement.

The design of this study allowed exploration of how different stages of document judgement contribute to user effort and cognitive load. Building on the multi-stage relevance judgement model proposed in this thesis, it was hypothesised that cognitive demand would increase from *aboutness* to *relevance* to *novelty* to *usefulness* judgements, with *aboutness* requiring the least effort and *usefulness* the most. The findings provided partial support for this hypothesis. As expected, *aboutness* judgements were rated as the least difficult and took the least time, suggesting that early-stage, surface-level assessments impose relatively low cognitive demands. However, contrary to expectations, *relevance* judgements - an intermediate stage - elicited the greatest effort, rather than *usefulness* judgements. This unexpected pattern raises questions about cognitive processing during document evaluation. One possibility is that *relevance* judgements, which require aligning document content with relevance criteria, involve more active cognitive integration than previously assumed. In contrast, *usefulness* judgements may rely more on intuitive or affective assessments once relevance is established. These findings indicate that cognitive demands in document evaluation are more nuanced than a simple linear progression, highlighting the need to consider both judgement type and task-specific context when assessing effort.

8.6.2 Effort and Load between Document Judgement Ratings

This section examines the extent to which effort and load differ in relation to the type of judgement rating provided (i.e., *"yes"* or *"no"*) for each of the document judgement stages **(HL-RQ3b)**. *Analysis of Variance (ANOVA)* tests were used and main effects were examined with level of significance determined at $p< 0.05$. The Holm-Bonferroni

correction for multiple comparisons was used for post-hoc analysis to examine which stages showed significant differences.

Results showed when the user rated their judgement as *"yes"*, there were significant differences between stages for: effort, $F(3,688)=8.64$, $p<.05$; difficulty, $F(3,688)=7.14$, $p<.05$ and; judgement time, $F(3,688)=3.78$, $p<.05$. Post-hoc comparisons revealed that *"yes"* ratings for the *relevance* stage required significantly greater effort and were considered more difficult than *"yes"* ratings for the remaining three stages. For judgement time, post-hoc tests showed no significant differences between judgement stages in relation to *"yes"* ratings.

Results showed when the user rated their judgement as *"no"*, there were significant differences between stages for: effort, $F(3,1231)=8.61$, $p<.05$; difficulty, $F(3,1231)=14.13$, $p<.05$ and; judgement time, $F(3,1231)=9.96$, $p<.05$. Post-hoc comparisons revealed that *"no"* ratings for the *aboutness* stage required significantly less effort than *relevance* and *novelty* stages, and were considered as less difficult and took less time than all three remaining stages. Table 8.2 provides the descriptive statistics for each dependent variable per rating for each document judgement stage. Figures 8.6 and 8.7 provide a graphical depiction of each dependent variable per rating for each document judgement stage.

Judgement Stage	**Rating**	**Judgement Time (s)**			**Effort**			**Difficulty**		
		M	**Mdn**	**SD**	**M**	**Mdn**	**SD**	**M**	**Mdn**	**SD**
aboutness	**0**	52.75	35.03	81.45	51.70	52.0	28.83	43.23	42.0	24.95
	1	46.48	29.81	53.90	44.10	42.5	30.31	27.81	23.5	23.52
relevance	**0**	58.39	40.08	66.47	63.84	70.0	24.38	53.04	58.5	25.21
	1	59.50	43.28	53.44	54.87	62.0	29.02	33.99	30.0	24.73
Novelty	**0**	64.33	49.20	60.26	55.33	60.0	23.12	43.88	40.0	22.75
	1	66.35	49.03	75.52	51.08	51.0	22.33	38.49	35.0	22.40
Usefulness	**0**	82.10	63.86	116.73	51.94	56.0	25.66	42.75	40.0	24.78
	1	73.51	52.10	79.82	49.23	50.0	27.99	37.93	34.0	25.41

Table 8.2: Mean (M), Median (Mdn), and Standard Deviations (SD) for each dependent variable per judgement rating and document judgement stage
Note: 0 = "Yes"; 1 = "No"

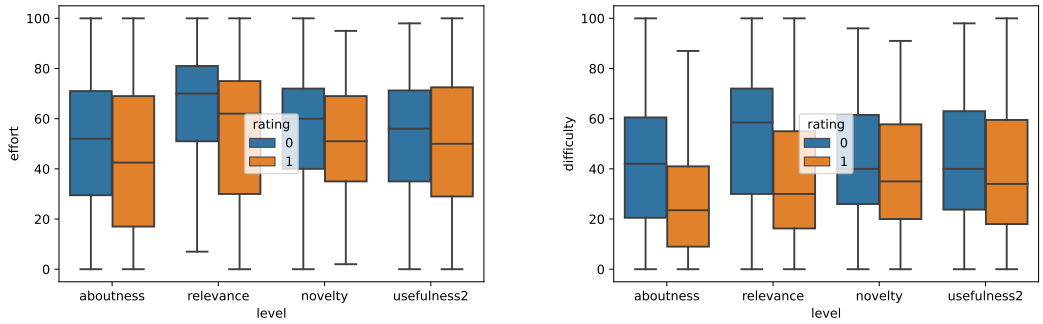

Figure 8.6: Effort and difficulty by document rating per document judgement stage

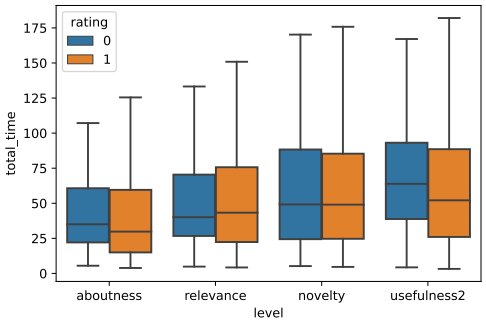

Figure 8.7: Judgement time by document rating per document judgement stage

Previous research has suggested that effort and load may be influenced by the users perception of the graded relevance of the document, i.e., whether the document is not relevant, relevant, or partially relevant. In this study, this hypothesis was tested by assessing whether there were differences in effort and load between the grade of rating assigned to the document. Users were found to exert the most effort and reported the highest difficulty for *"yes"* ratings assigned at the *relevance* stage compared to the other three stages. *"No"* ratings at the *aboutness* stage were the least difficult to make and took the least time compared to the remaining three stages.

These results offer valuable insight into how user effort and cognitive load vary not only across different stages of document judgement but also according to the type of rating assigned. When users judged a document as relevant (*"yes"*), the *relevance* stage consistently elicited significantly higher levels of effort and perceived difficulty compared

to the other stages. This finding suggests that confirming a document's relevance may demand more cognitive engagement than either initial topical assessments (*aboutness*) or more interpretive evaluations (*usefulness*, *novelty*). In contrast, when users rated (*"no"*), *aboutness* judgements required significantly less effort, were rated as easier, and were completed more quickly than *"no"* ratings at later stages. This supports the idea that early rejections are often based on surface-level cues and can be made with minimal cognitive investment.

8.6.3 Relationships between Measures

This section examines the operational relationships between the measures used **(HL-RQ3c)**. When all document stages were examined together, the results of the Spearman correlation analyses (see Table 8.3 for the correlation matrix of dependent variables) using the Bonferroni correction for multiple comparisons showed moderate correlations between effort and difficulty, $rs = .61$, $p<.05$ (see Figure 8.8). While a significant relationship was observed between judgement time and the mental demand dimension of the NASA-TLX, this correlation was very weak. No further relationships were observed between judgement time and the other measures.

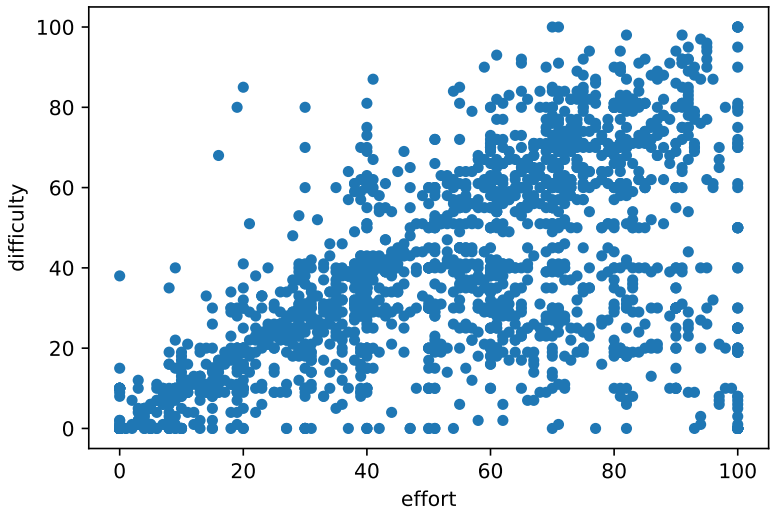

Figure 8.8: Relationship between effort and difficulty

To understand how in-task measures related to the post-task measures, a correlational analysis was performed to examine how the mean, minimum, and maximum, effort and difficulty, related to the NASA-TLX dimensions and perceived overall workload. In relation to NASA-TLX dimensions, significant moderate relationships were observed between: maximum effort and mental demand, rs = .56, p<.05; maximum difficulty and mental demand, rs = .46, p<.05; maximum effort and NASA-TLX effort, rs = .55, p<.05. Figures 8.9 and 8.10 show these relationships.

Weaker significant relationships were observed between: maximum effort and temporal demand, rs = .28, p<.05; maximum difficulty and temporal demand, rs = .34, p<.05; maximum effort and frustration, rs = .31, p<.05; maximum difficulty and frustration, rs = .45, p<.05; maximum difficulty and NASA-TLX effort, rs = .39, p<.05.

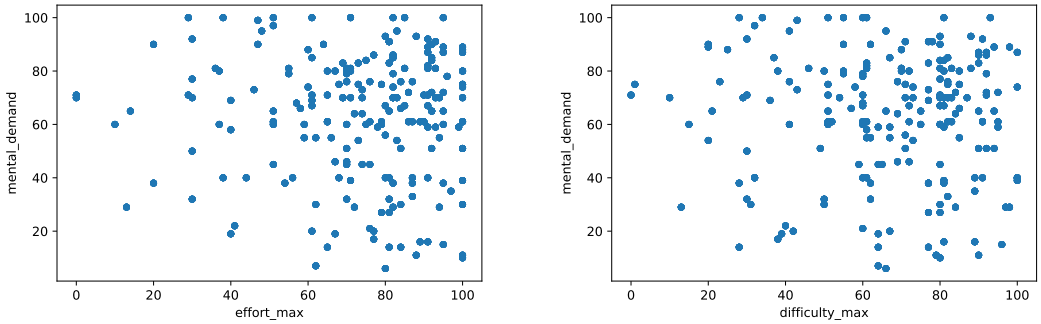

Figure 8.9: Relationships between maximum effort and maximum difficulty with mental demand

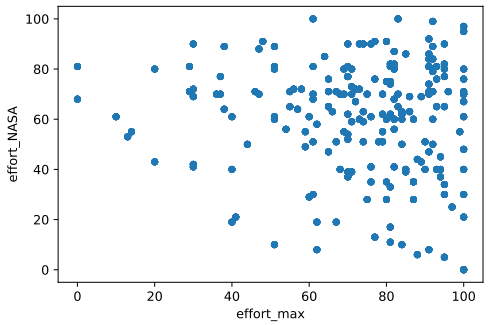

Figure 8.10: Relationships between maximum effort and NASA-TLX effort

Significant moderate correlations were also observed between: maximum effort and

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Minimum Effort	_													
2. Mean Effort	.87*	_												
3. Maximum Effort	.45*	.75*	_											
4. Minimum Difficulty	.58*	.45*	.12*	_										
5. Mean Difficulty	.50*	.61*	.46*	.75*	_									
6. Maximum Difficulty	.23*	.49*	.63*	.27*	.73*	_								
7. Mental Demand**	.08*	.03	.56*	-.04	-.02	.46*	_							
8. Physical Demand**	.02	.05	.23*	-.03	-.01	.21	.18*	_						
9. Temporal Demand **	.07	.01	.28*	-.03	-.06	.34*	.40*	.30*	_					
10. Performance**	-.04	-.04	.12	-.03	-.04	.17	.06	.12*	.11*	_				
11. Frustration**	.10*	.05	.31*	-.06	-.05	.45	.443	.23*	.50*	.15	_			
12. Effort**	.05	-.00	.55*	-.02	-.02	.39*	.73*	.23*	.38*	.21	.35*	_		
13. Overall Workload**	.09*	.54*	.50*	-.04	-.05	.50*	.72*	.48*	.73*	.40*	.70*	.71*	_	
14. Judgement Time	-.05	-.02	.03	-.08	-.04	-.00	.11*	.04	.05	.05	.03	.08	.08	_

Table 8.3: Spearman correlation matrix for dependent variables (n=204) *p<.05 ** NASA-TLX Dimensions

overall workload, rs = .50, p<.05; maximum difficulty and overall workload, rs = .50, p<.05. Figure 8.11 illustrates these relationships.

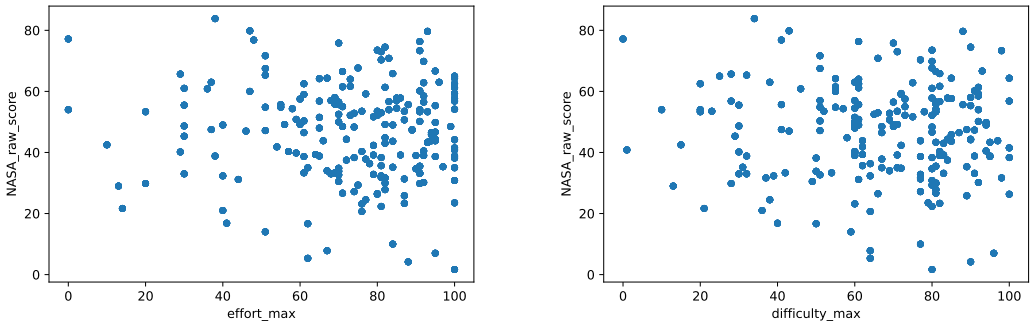

Figure 8.11: Relationships between maximum effort and maximum difficulty with overall workload

Similar to Study 1, this study showed a moderate correlation between the single-item measures of effort and difficulty, in alignment with previous research [188]. The moderate correlations between maximum in-task effort and difficulty and some of the post-task NASA-TLX ratings (mental demand and effort) suggest that users may anchor their retrospective assessments on the most demanding moments experienced during the task, rather than averaging across the entire task. This indicates that post-task workload measures may disproportionately reflect peak cognitive demands, potentially overlooking subtler fluctuations in effort and difficulty that occur at different stages. For

ISR research, this highlights a limitation of relying solely on post-task instruments like NASA-TLX to capture user workload. Combining continuous, in-task measures with retrospective assessments may therefore provide a more comprehensive understanding of cognitive demand, revealing both peak and sustained effort across task stages.

8.6.4 Effort and Load over Task Duration

This section examines how effort and load measures vary over time, or in other words, across the duration of the document judgement task (**HL-RQ3d**). As much of the data analysed showed significant differences for Levene's Test, the non-parametric Kruskal-Wallis H-Test and Games-Howell post-hoc tests were used for this analysis. When all document stages were analysed together, no significant differences were observed between user effort and difficulty over the duration of the task. However, significant differences were observed for document judgement time, $H(9) = 218.2$, $p<.05$. Table 8.4 shows the descriptive statistics for all measures per document judgement ordering. Post-hoc comparisons showed that the first document took significantly longer to judge than the remaining nine documents. Figure 8.12 illustrates how document judgement time varies across the duration of the task.

Document	Judgement time (sec)			Effort			Difficulty		
order	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD
1st	109.03	92.30	79.39	56.90	61.0	27.12	36.76	30.0	25.33
2nd	78.48	59.53	75.83	54.12	60.0	26.85	37.93	31.0	25.54
3rd	65.01	49.75	62.66	51.81	55.0	27.66	36.90	35.0	24.44
4th	57.47	45.74	57.61	52.97	60.0	26.34	39.07	35.0	24.84
5th	60.73	44.73	68.86	52.66	56.0	26.76	39.38	37.0	24.59
6th	54.89	40.40	51.23	52.58	56.0	26.20	41.17	39.0	24.70
7th	65.95	41.36	133.40	49.14	51.0	28.61	36.61	31.0	26.16
8th	49.59	34.18	50.10	48.95	46.0	27.97	38.74	31.0	26.13
9th	48.92	35.01	53.86	50.21	53.0	27.28	39.41	37.0	24.65
10th	43.27	30.30	43.56	47.84	51.0	27.98	37.04	34.0	24.69

Table 8.4: Mean (M), Median (Mdn), and Standard Deviations (SD) for judgement time, effort, and difficulty, by document order.

Although user effort and difficulty remained stable across the task, decision times decreased as the task progressed. This suggests that users became more efficient without

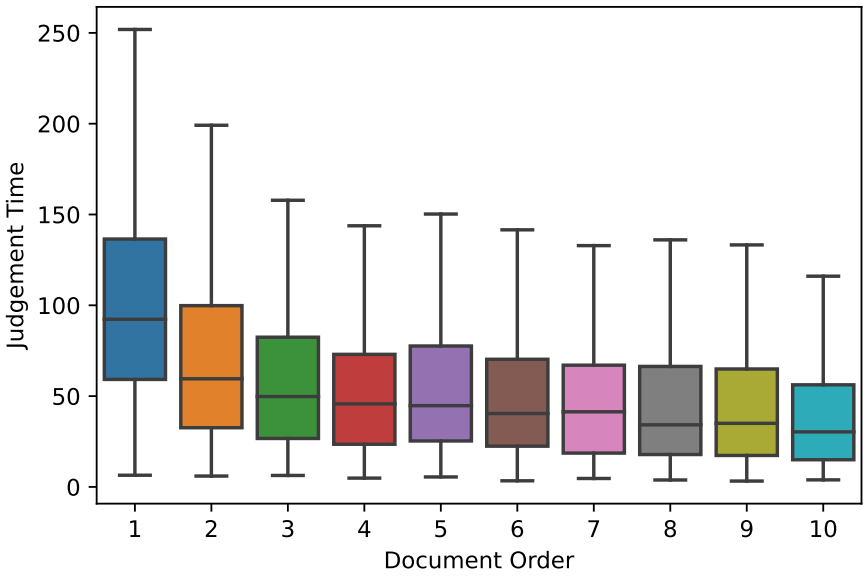

Figure 8.12: Document judgement time from first document judged to last document judged

experiencing reduced cognitive demand, potentially due to task familiarity or strategy development. For ISR research, these findings highlight that improvements in performance do not necessarily reflect reductions in perceived effort or load. Measuring both subjective workload and behavioural indicators, such as decision time, provides a more complete picture of user adaptation and efficiency during document judgement tasks.

8.7 Conclusion

Overall, Study 2 found that users exerted more effort for *relevance* judgements compared to the other three judgement stages (i.e., *aboutness*, *novelty*, and *usefulness*). This finding was also reflected for *relevance* judgements when the rating provided was *"yes"* rather than *"no"* or *"partially"*. When the relationships between measures were examined, moderate correlations were observed between effort and difficulty; maximum effort and mental demand; maximum difficulty and mental demand; and maximum effort and NASA-TLX effort, however the relationships between the remaining measures were shown to be weak or significant. Finally, when examining how effort and load

differed across the duration of the task, results showed that document judgement time decreased over time from the first document judged to the last document judged.

Chapter 9

Study 3: Within-Subjects Evaluation of Model and Measures

9.1 Motivation

The third study presented in this chapter further explores effort and load measurement within the context of multi-stage relevance judgement.

This study seeks to answer the third high-level research question outlined in Section 1.3, **HL-RQ3 (Empirical): How can effort and load be measured within the ISR sub-task of multi-stage document judgement, as defined by the theoretical model?**

- (a) How do different stages of document judgement vary in terms of user effort and load?
- (b) To what extent is effort and load influenced by the type of judgement rating ("no", "partially", "yes") between document judgement stages?
- (c) What are the operational relationships between effort and load?
- (d) How do measures of effort and load change over time during the task?

- (e) To what extent do user characteristics (topic knowledge and motivation) influence measures of effort and load?

In line with the progression from Study 1 to Study 2, this study can be regarded as a further iteration of the design, undertaken with the aim of refining the methodology to enable a more thorough assessment of user effort and load. By leveraging the lessons learned from the Study 1 and Study 2, Study 3 seeks to improve on the following:

1. **Within-Subjects Design:** Although Study 1 employed a within-subjects design, participants were required to answer all judgement questions for each document before providing a single effort and load rating. This made it difficult to determine which specific judgements were perceived as more effortful or cognitively demanding. Study 2 addressed this issue by adopting a between-subjects design, in which users answered only one type of judgement question, allowing the influence of each question on effort and load to be examined more directly. However, this design did not reflect a realistic document evaluation scenario, as users typically make multiple judgements when examining a document. Study 3 was therefore designed to overcome the limitations of both approaches by requiring users to complete all judgement questions for each document, more closely replicating real-world behaviour, while also collecting effort and load ratings after each individual judgement.

2. **Number of Topics:** Unlike Studies 1 and 2, which used documents from a single topic, Study 3 incorporated four different topics. This design reduces the influence of topic-specific factors, such as familiarity, complexity, or interest, on effort and load, providing a more reliable and generalisable assessment of how these constructs vary across judgement types.

3. **Readability:** In Study 3, readability scores were measured for each topic's documents, unlike in Studies 1 and 2. This allowed control for text complexity, ensuring that differences in effort and load could be attributed to judgement type rather than variations in document difficulty.

4. **Additional Measures:** Study 3 introduced **click count** as an additional objective measure of effort, providing triangulation with subjective ratings and complementing the measures used in Studies 1 and 2.

5. **Judgement Rating Options:** Study 3 expanded each judgement stage rating by adding a "partially" option, moving beyond the simple "yes/no" used in Studies 1 and 2 to capture more nuanced user judgements.

6. **Individual Differences:** Although Study 1 included cognitive ability measures (perceptual speed and working memory), these assessments added considerable time to the study and were reported by participants as cognitively demanding, potentially influencing subsequent effort and load ratings. Recognising the continued importance of individual differences, Study 3 instead incorporated two factors, *motivation* and *topic knowledge*, that could affect user effort and cognitive load. An advantage of these measures is that they can be assessed using single-item questions, imposing minimal additional cognitive demand and reducing the risk of confounding subsequent ratings.

Study 3 represents a natural progression from Study 2, building on the methodological refinements and insights gained from the earlier work. While Study 2 increased the granularity of effort and load measurement and began to disentangle stage-specific cognitive demands, it did not address all five research sub-questions. Study 3 was therefore designed to extend the investigation to comprehensively address the full set of research questions, ensuring that the theoretical model of multi-stage relevance judgement could be tested in its entirety. In addition, Study 3 incorporated further experimental variables and measurement types to examine how intrinsic cognitive load and effort vary across all stages, providing a more complete understanding of user behaviour during document judgement. This approach allowed the thesis to move from pilot and partial validation (Studies 1 and 2) toward a full-scale empirical evaluation, capturing the complex dynamics of effort and load across the multi-stage judgement process and providing the necessary evidence to answer all research questions.

9.2 Method

This section outlines the specific methodology used in this user study, building on the general methodology described in Chapter 5. To aid clarity, references are made to the relevant sections of the general methodology where appropriate.

The experimental design implemented in this user study was narrow and controlled, consisting of a document judgement task only. The study followed a **within-subjects** design, where participants were required to answer **four** document judgement questions corresponding with the four judgement stages (*aboutness*, *relevance*, *novelty*, and *usefulness*) outlined in the model of multi-stage relevance judgement (Figure 4.1). For this study, **four topics** (Wildlife Extinction; Airport Security; Transportation Tunnel Disasters; Tropical Storms) were selected from the TREC 2018 Common Core Track. Participants were randomly assigned to **one topic** only. Ten documents were selected from each topic collection, five relevant and five non-relevant. All 40 documents were selected in relation to their level of readability as assessed by the Flesch-Kincaid readability test. This test produces a Reading Ease score between 1 and 100, where 100 is the highest (i.e., most difficult) readability score. All documents selected for this study had an average Reading Ease score between 50-60, which corresponds to the "Fairly Difficult to Read" readability category. As outlined in Studies 1 and 2, documents were also manually edited to ensure they were of the same length, font type, and font size, to limit the effects of any confounding extraneous factors on the dependent variables.

After each document judgement, participants were asked to rate their level of effort and difficulty on two-single item scales. After completing judgements for all ten documents, participants were asked to complete the NASA-TLX in order to assess their workload for the judgement task overall. As participants judged the documents, Qualtrics software gathered log data (judgement time and number of clicks) and survey responses. Figure 9.1 shows an example of the document judgement page, Figure 9.2 shows the effort and load question page, and Figure 9.3 shows an example of the NASA-TLX question page, for the experimental system used in study 3.

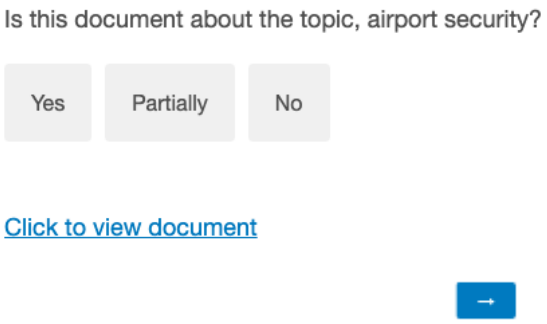

Figure 9.1: Screen shots of the experimental system used in study 3. This figure illustrates the document judgement page (for the *aboutness question* for the *airport security* topic condition).

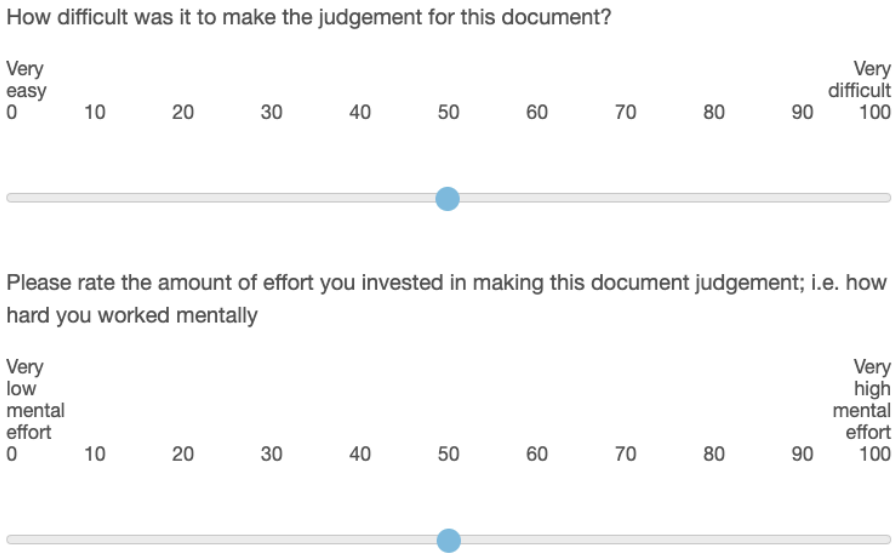

Figure 9.2: Screen shots of the experimental system used in study 3. This figure illustrates the effort and load question page.

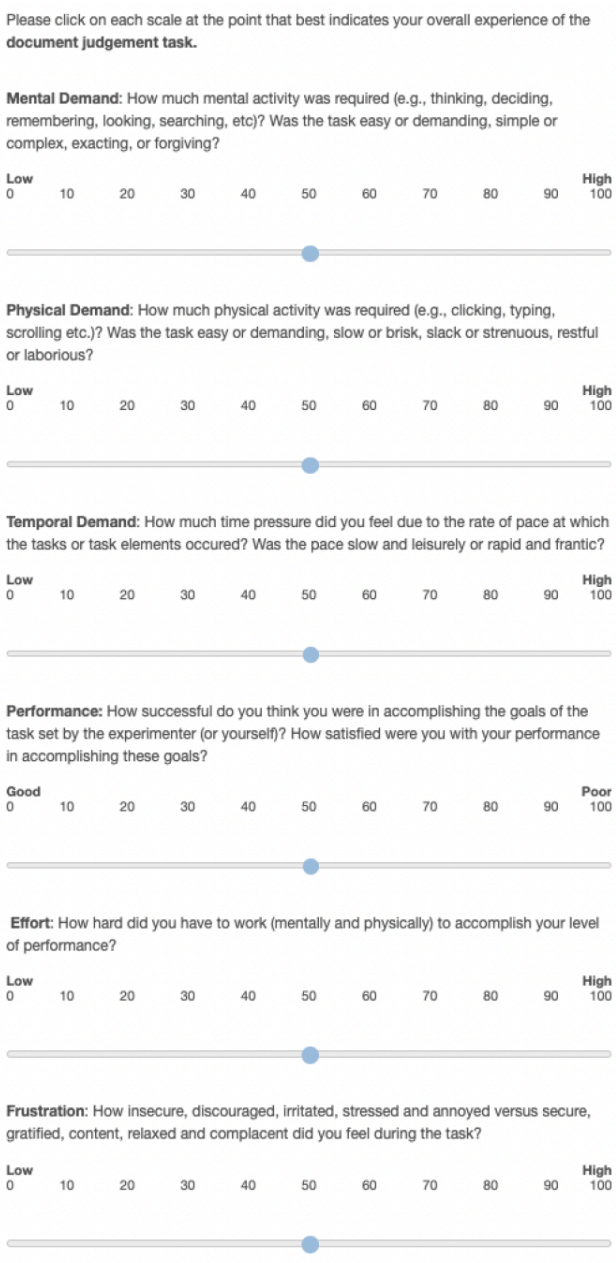

Figure 9.3: Screen shots of the experimental system used in study 3. This figure illustrates the NASA-TLX question page.

9.3 Measures - IVs & DVs

The independent variables used in this study were as follows:

- *Document Judgement Stage* (4 levels: *aboutness/*relevance*/novelty/usefulness*)
- *Document Judgement Rating* (3 levels: no/partially/yes)

The dependent variables used in this study were as follows:

- *Effort*: as measured by the single-item *Subjective Rating Scale*.
- *Difficulty*: as measured by the single-item *Perceived Difficulty Scale*.
- *Overall workload*: as measured by the NASA-TLX.
- *Judgement time*: gathered by Qualtrics data logs, and measured as the time spent on each document judgement page.
- *Click Count*: gathered by Qualtrics data logs, and measured as the number of clicks administered on each document judgement page.

Refer to Chapter 5 for a detailed overview of each measure.

9.4 Procedure

This study obtained ethics approval from the University of Strathclyde's Department of Computer and Information Sciences (Approval No. 2137). Participants were recruited via the crowd-sourcing platform Prolific where participants were paid at a £9.50 hourly rate. The experimental procedure for this study is outlined in Section 5.4, and in Figure 5.1. The study took participants approximately 30 minutes to complete. Following the online experimental briefing and obtaining participants consent, the study proceeded as follows:

1. Demographic questionnaire (3 minutes)
2. Task description, and pre-task questionnaires (topic knowledge and motivation) (2 minutes)

3. Document judgement task - ten documents (22 minutes)

4. Post-task workload questionnaire (3 minutes)

Prior to commencing the task, participants were introduced to the topic description and then asked to rate their topic knowledge and level of interest in that topic. During the task, the document judgement question was displayed at the top of the page, with a link to the document PDF below. To view the document, participants were required to click on the link. The order in which the ten documents were presented was randomised by the Qualtrics for each participant to mitigate ordering effects. For each document, participants were required to answer four judgement questions, which were as follows:

- *Aboutness*: "Is the document about the topic...?"
- *Relevance*: "Does the document fulfill the relevance criteria listed above?"
- *Novelty*: "Does this document contain a relevant example that was not included in the documents you have already seen?"
- *Usefulness*: "Does this document contain a relevant example that would be useful for your case study about..."

For each document judgement question, participants were required to rate their response on a graded scale of *"yes"*, *"partially"*, *"no"*. Following the judgement rating, participants were required to click an arrow to direct them to the following page which contained the follow-up difficulty and effort self report questions. This was repeated for all ten documents. Participants could not revisit previously viewed documents and there was no time limit imposed in relation to task completion, however participants were encouraged to perform the task as quickly and accurately as possible. After completing all ten documents, participants were directed to the NASA-Task Load Index, where they were asked to rate their perceived overall workload for the task.

For more details about each of the measures used, please refer to Section 5.6.

9.5 Demographics

A total of 204 participants completed the experiment: 115 males; 88 females; and one participant preferred not to say. The majority of participants were aged between 18-39 (N=120) with the remaining participants aged between 40-79 (N=84). The majority of participants held a Bachelor's level qualification or above (N=119). Screening criteria was specified on Prolific to only include participants from the UK or US, and those who had English as their first language. Participants who completed less than 90% of the study were removed from the final data-set, as were any extreme outliers as detected by the Interquartile Range (IQR) method. This left a total of 6819 annotations included in the final analysis.

9.6 Results and Discussion

This study focused on how document judgement decisions vary in relation to user effort and load, while also examining the relationships between different measurement approaches. The findings offer deeper insight into the patterns of effort and load that emerge across the relevance judgement process, extending beyond prior work that has largely considered effort and load within single-stage relevance judgements. This work sought to answer the second high-level research question outlined in Section 1.3, **HL-RQ3 (Empirical): How can effort and load be measured within the ISR sub-task of multi-stage document judgement, as defined by the theoretical model?**

- (a) How do different stages of document judgement vary in terms of user effort and load?
- (b) To what extent is effort and load influenced by the type of judgement rating ("no", "partially", "yes") between document judgement stages?
- (c) What are the operational relationships between effort and load?
- (d) How do measures of effort and load change over time during the task?

- (e) To what extent do user characteristics (topic knowledge and motivation) influence measures of effort and load?

The following sections first present the results of Study 3 and discuss the key findings. This is followed by the presentation and discussion of the replication study. Finally, a general overview is provided, synthesising the main findings across both studies. Note that the term "participants" is used in the methods sections to emphasise their role in the study procedures, consent, and experimental protocol, whereas "users" is used in the results sections to focus on their interactions and behaviours while interacting with the document judgement task.

9.6.1 Effort and Load between Document Judgement Stages

This section discusses the results on how effort and load varied across the four document judgement stages (**HL-RQ3a**). *Analysis of Variance (ANOVA)* tests were used and main effects were examined with level of significance determined at $p = 0.05$. The Holm-Bonferroni correction for multiple comparisons was used for post-hoc analysis to examine which stages showed significant differences.

Results showed statistically significant differences between document judgement stages for effort, $F(3,141)=3.09$, ($p<.05$), and document judgement time, $F(3,141)=2.9$, ($p<.05$). There were no significant differences between document judgement stages in relation to difficulty or click count. Table 9.1 shows the mean, median, and standard deviations for each dependent variable per document judgement stage. For the *aboutness* stage, post-hoc pairwise comparisons revealed that users reported significantly more effort than *novelty* or *usefulness* stages, and *usefulness* judgements required significantly more effort than *novelty* judgements. Users took significantly less time when making *relevance* judgements. Figure 9.4 illustrates these differences.

9.6.2 Effort and Load between Document Judgement Ratings

This section examines whether user effort and load differ in relation to the type of judgement rating provided (i.e., *"yes"*, *"partially"*, *"no"*) **(HL-RQ3b)**. *Analysis of Variance*

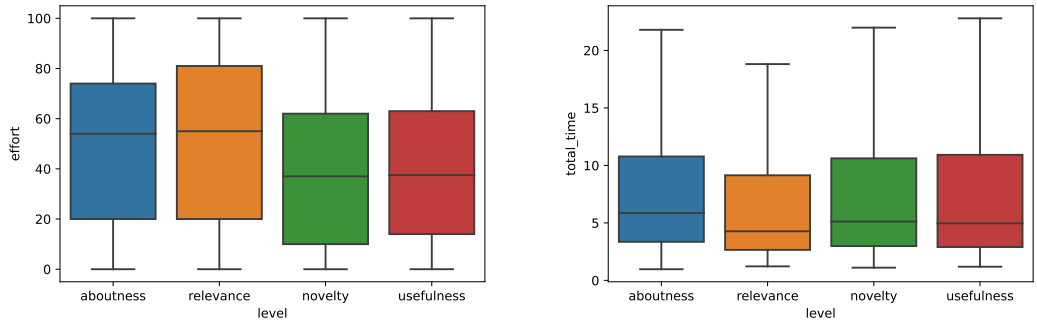

Figure 9.4: Effort and judgement time per document judgement stage

Stage	Judgement Time (s)			Effort			Difficulty			No. Clicks		
	M	**Mdn**	**Std**	**M**	**Mdn**	**Std**	**M**	**Mdn**	**Std**	**M**	**Mdn**	**Std**
Aboutness	11.47	5.86	18.32	48.64	54.0	30.50	29.51	21.0	25.76	1.39	1.0	0.92
Relevance	9.61	4.27	15.89	51.47	55.0	33.76	26.78	20.0	26.44	1.38	1.0	0.73
Novelty	11.20	5.12	21.55	37.91	37.0	28.49	23.13	16.0	22.58	1.43	1.0	1.16
Usefulness	11.75	4.96	21.91	40.17	37.5	29.48	28.18	22.0	24.92	1.39	1.0	0.85

Table 9.1: Mean (M), Median (Mdn), and Standard Deviations (SD) for each dependent variable per document judgement stage

Stage	Rating	Judgement Time (s)			Effort			Difficulty			No. Clicks		
		M	**Mdn**	**SD**	**M**	**Mdn**	**SD**	**M**	**Mdn**	**SD**	**M**	**Mdn**	**SD**
Aboutness	**0**	13.96	5.94	21.44	46.05	45.0	31.68	28.92	18.0	27.35	1.50	1.0	1.08
	1	14.93	7.75	22.22	56.59	60.0	26.92	42.12	40.0	24.83	1.52	1.0	1.03
	2	10.72	5.26	24.24	45.29	41.0	31.10	24.30	15.0	23.92	1.30	1.0	0.72
Relevance	**0**	9.37	4.19	14.72	41.22	35.0	35.04	20.12	10.0	24.56	1.42	1.0	0.76
	1	12.19	6.16	20.26	65.29	67.0	23.16	45.96	46.0	24.58	1.39	1.0	0.68
	2	8.91	4.16	14.27	55.59	60.0	33.64	23.64	17.0	24.15	1.35	1.0	0.72
Novelty	**0**	12.33	5.00	24.69	31.67	24.0	28.66	19.48	10.0	23.30	1.49	1.0	0.91
	1	12.58	6.80	19.46	48.80	53.0	25.29	37.19	33.0	22.71	1.50	1.0	0.91
	2	10.89	4.75	22.80	35.15	30.0	27.91	17.73	10.0	18.11	1.41	1.0	1.36
Usefulness	**0**	10.10	4.49	16.42	34.51	29.0	29.93	22.36	14.0	23.96	1.38	1.0	0.96
	1	15.44	6.42	28.25	50.61	52.5	25.86	40.49	39.0	24.20	1.48	1.0	0.93
	2	13.38	5.57	23.20	39.38	35.0	29.38	26.53	20.0	23.72	1.44	1.0	0.88

Table 9.2: Mean (M), Median (Mdn), and Standard Deviations (SD) for each dependent variable per judgement rating and document judgement stage
Note: 0 = "No"; 1 = "Partially"; 2="Yes"

(ANOVA) tests were used and main effects were examined with level of significance determined at p= 0.05. The Holm-Bonferroni correction for multiple comparisons was used for post-hoc analysis to examine which stages showed significant differences.

Results showed when the user rated their judgement as *"yes"*, there were significant differences between stages for: effort, $F(3,141)=2.77$, $p<.05$; and judgement time,

$F(3,141)=3.44$, $p<.05$. Post-hoc comparisons revealed that *"yes"* ratings for the *aboutness* stage required less effort than *"yes"* ratings for the other three stages, and while *"yes"* ratings for the *relevance* stage were performed faster than for *novelty* and *usefulness*, they required greater user effort.

Results further showed, that when the user rated their judgement as *"partially"*, there were significant differences between stages for effort only, $F(3,135)=4.39$, $p<.05$. Post-hoc comparisons showed that for *"partially"* judgements made in the *aboutness* stage, users exerted significantly less effort than for the other three stages. For *"partially"* judgements made in the *relevance* stage, users exerted significantly greater effort than the *novelty* and *usefulness* stages.

Finally, results showed that when the user rated their judgement as *"no"*, there were significant differences between stages for: effort, $F(3,141)=5.31$, $p<.05$; difficulty, $F(3,141)=9.54$, $p<.05$; and document judgement time, $F(3,141)=6.20$, $p<.05$. Post-hoc comparisons showed that when users rated their judgement as *"no"* for the *aboutness* stage, they experienced greater difficulty and exerted more effort than for the other three judgement stages. *"No"* ratings provided in the *aboutness* stage took longer than those made in the *relevance* and *usefulness* stages.

Table 9.2 shows the mean, median, and standard deviations for each of the dependent variables per judgement rating and document judgement stage. Figures 9.5 and 9.6 illustrate how each measure differs in relation to the judgement rating and document judgement stage.

9.6.3 Relationships between Measures

This section examines the operational relationships between the effort and load measures used both during and post-task **(HL-RQ3c)**. The Spearman's rank correlation found moderately significant correlations between effort and difficulty, rs = .56, $p=<.05$; and click count and judgement time, rs = .40, $p=<.05$, Figure 9.7 illustrates relationships.

No significant correlations were observed between the NASA-TLX factors and the other measures. Table 9.3 shows the descriptive statistics and correlation coefficients

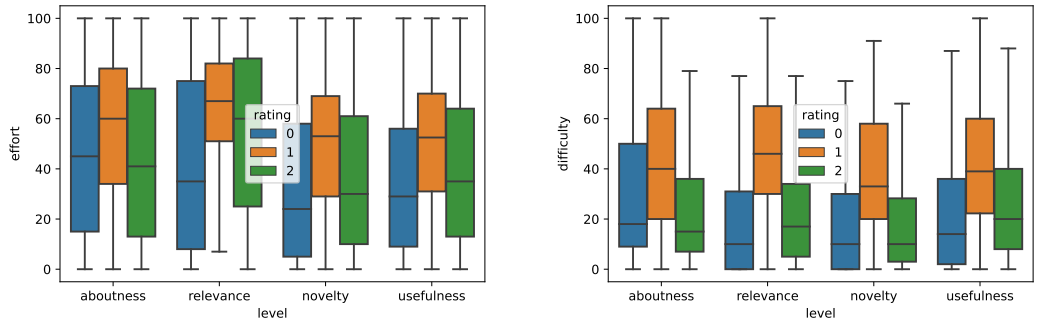

Figure 9.5: Effort and difficulty for document judgement ratings for each document judgement stage
Note: 0 = "No"; 1 = "Partially"; 2="Yes"

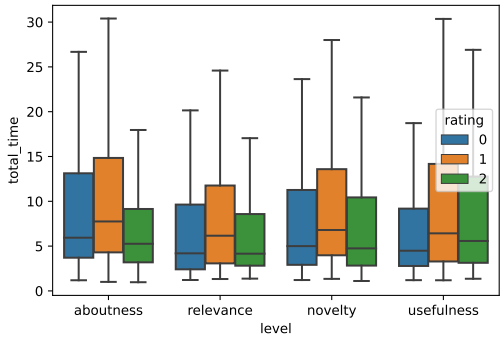

Figure 9.6: Judgement time for document judgement ratings for each document judgement stage
Note: 0 = "No"; 1 = "Partially"; 2="Yes"

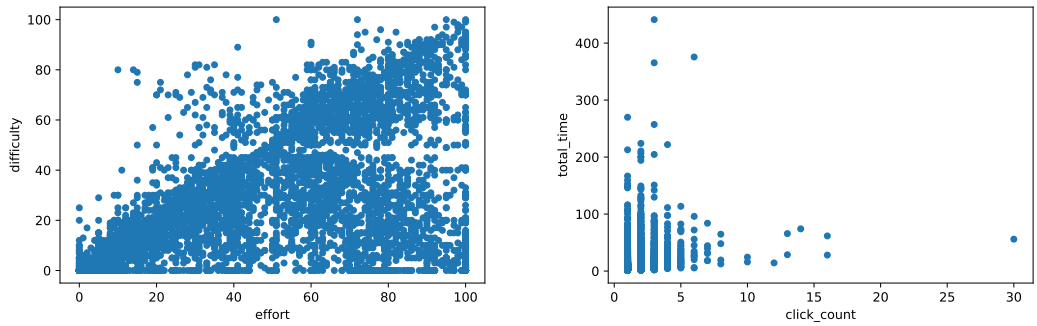

Figure 9.7: Relationship between effort and difficulty, and judgement time and number of clicks

for each dependent variable. Correlations between maximum and minimum effort and difficulty scores, and the NASA-TLX dimensions revealed weak significant relationships between: maximum effort and NASA-TLX effort, rs =.17, p<.05; maximum difficulty and NASA-TLX effort, rs =.26, p<.05, and maximum difficulty and NASA-TLX overall score, rs =.16, p<.05. Table 9.4 shows the descriptive statistics and correlation coefficients for the maximum and minimum effort and difficulty scores, and the NASA-TLX dimensions.

	M	**SD**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1. Effort	44.52	31.04	_										
2. Difficulty	27.18	25.11	.56*	_									
3. Judgement Time	11.67	21.27	.12*	.20*	_								
4. No. Clicks	1.42	.94	.03	.10*	.40*	_							
5. Mental Demand**	59.12	25.12	-.05	.02	-.06	-.07	_						
6. Physical Demand**	28.99	23.90	-.04	-.00	-.05	-.05	.40*	_					
7. Temporal Demand **	38.77	23.94	.00	-.04	-.07	-.06	.46*	.50*	_				
8. Performance**	52.40	25.17	-.00	-.03	-.00	-.06	.29*	.10*	.28*	_			
9. Frustration**	36.15	26.23	-.05	-.03	-.05	-.05	.39*	.33*	.57*	.18*	_		
10. Effort**	62.02	22.21	.01	.00	-.04	-.02	.64*	0.35*	.35*	.29*	.30*	_	
11. Overall Workload**	46.24	16.78	-.02	.00	-.06	-.06	.75*	.64*	.77*	.48*	.70*	.67*	_

Table 9.3: Means (M), Standard Deviations (SD), and Spearman correlation matrix for dependent variables (n=204).
*p <.05
** NASA-TLX measures

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1. Minimum Effort	_												
2. Mean Effort	.60*	_											
3. Maximum Effort	-.10	.39*	_										
4. Minimum Difficulty	.46*	.24*	-.17	_									
5. Mean Difficulty	.47*	.48*	.05	.28*	_								
6. Maximum Difficulty	.13*	.23*	.25*	.23*	.60*	_							
7. Mental Demand**	-.12	.06	.18*	.04	.14*	.22*	_						
8. Physical Demand**	.09	.02	-.00	.06	.06	.12*	.43*	_					
9. Temporal Demand **	.04	.07	.14	-.06	-.08	-.04	.47*	.52*	_				
10. Performance**	.02	-.08	.06	-.06	-.06	.12*	.29*	.10*	.28*	_			
11. Frustration**	-.03	-.08	.07	-.03	-.01	.13*	.40*	.33*	.57*	.18*	_		
12. Effort**	-.05	.03	.18*	.02	.08	.29*	.65*	.37*	.35*	.29*	.32*	_	
13. Overall Workload**	.04	.04	.11*	-.01	.07	.19*	.76*	.65*	.77*	.47	.70*	.67*	_

Table 9.4: Spearman correlation matrix for minimum, Mean, maximum dependent variable values and NASA-TLX dimensions (n=204).
*p <.05
** NASA-TLX measures

9.6.4 Effort and Load across Task Duration

This section of results addresses how effort and load vary over time, or in other words, across the duration of the document judgement task from the first document judged to the last document judged **(HL-RQ3d)**. The one-way ANOVA revealed significant differences for all measures over time; effort, $F(9,6809)=9.13$, $p<.05$; difficulty, $F(9,6809)=5.34$,$p<.05$; judgement time, $F(9,6809)=119$, $p<.05$; click count,$F(9,6809)=48.7$, $p<.05$. Figures 9.8 and 9.9 illustrate these differences.

Post-hoc pairwise comparisons showed that the first document judgement differed significantly from the subsequent nine documents in relation to mean effort, click count, and judgement time. The first document judged also differed significantly from the 4th-10th documents judged in relation to user difficulty. For document judgement time, the first and second documents judged differed significantly from the 3rd-10th documents judged. Number of clicks differed significantly for the first document compared to the 2nd-10th documents. Table 9.5 shows the mean, median, and standard deviations for each dependent variable from the 1st document judged to the 10th document judged.

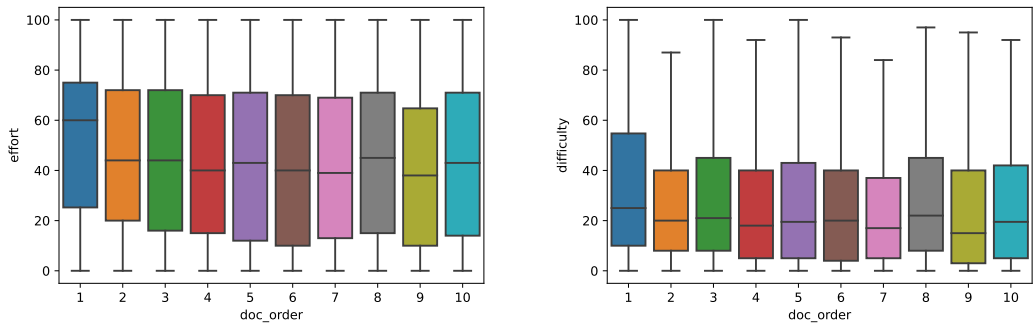

Figure 9.8: Effort and difficulty from first document judged to last document judged

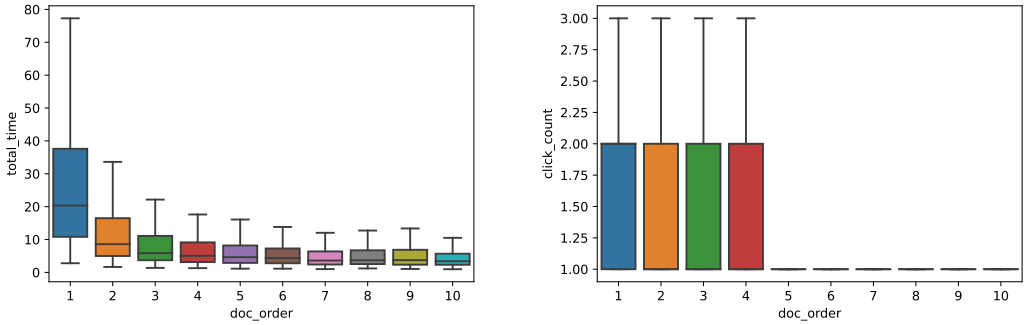

Figure 9.9: Judgement time and click count from first document judged to last document judged

Document order	Judgement time (sec) mean	median	SD	Effort mean	median	SD	Difficulty mean	median	SD	No. Clicks mean	median	SD
1st	30.36	20.32	34.64	52.91	60.0	29.28	31.33	25.0	26.32	2.04	2.0	1.62
2nd	16.02	8.58	25.39	46.64	44.0	30.16	27.51	20.0	24.59	1.50	1.0	1.43
3rd	10.62	5.85	15.54	44.88	44.0	31.12	28.38	21.0	25.56	1.36	1.0	0.74
4th	9.31	5.05	14.15	43.85	40.0	30.75	24.86	18.0	23.74	1.33	1.0	0.66
5th	9.39	4.63	17.31	44.51	43.0	32.22	26.85	19.5	25.30	1.29	1.0	0.62
6th	7.46	4.36	11.82	42.36	40.0	31.86	25.77	20.0	24.84	1.29	1.0	0.62
7th	5.93	3.61	41.76	41.76	39.0	31.59	24.01	17.0	23.94	1.27	1.0	0.56
8th	6.86	3.69	45.02	45.02	45.0	30.92	28.63	22.0	25.36	1.30	1.0	0.69
9th	6.47	3.73	39.55	39.55	38.0	31.11	24.92	15.0	25.50	1.29	1.0	0.61
10th	6.28	3.39	44.01	44.01	43.0	30.91	26.22	19.5	24.74	1.29	1.0	0.62

Table 9.5: Mean (M), Median (Mdn), and Standard Deviations (SD) for judgement time, effort, difficulty, and number of clicks by document order.

9.6.5 Topic Knowledge and Motivation

This section of results examines how individual differences between users such as topic knowledge and motivation may influence the level of effort and load they experience during the *relevance* judgement task **(HL-RQ3e)**. Table 9.6 provides an overview of the number of users for each level of motivation and topic knowledge.

	Motivation Level	***N***	**Topic Knowledge Level**	***N***
0	"Not motivated"	21	"Almost nothing"	62
1	"Somewhat motivated"	85	"Same as most people"	141
2	"Motivated"	98	"Expert"	1

Table 9.6: Number of users (N) by each level of motivation and topic knowledge

For motivation, a series of one-way ANOVAs found that there were no significant

differences in effort, judgement time, and click count for all three levels of motivation ("not motivated", "somewhat motivated", "motivated"). However, significant differences were observed for difficulty and the three levels of motivation, $F(2,7837)=7.9$, $p<.05$. There were also significant differences observed for all NASA-TLX factors for each level of motivation: overall workload, $F(2,7837)=17.22$, $p<.05$; frustration, $F(2,7837)=11.96$, $p<.05$; effort, $F(2,7837)=52.45$, $p<.05$; mental demand $F(2,7837)=53.72$, $p<.05$; physical demand, $F(2,7837)=336.78$, $p<.05$; temporal demand, $F(2,7837)=21.76$, $p<.05$, and performance, $F(2,7837)=262.42$, $p<.05$. Significant Tukey post-hoc tests revealed that "motivated" users report the highest difficulty, overall workload, NASA-TLX effort, mental demand, and physical demand. Users who were "not motivated" reported the highest frustration and temporal demand, and "somewhat motivated users" reported the highest NASA-TLX performance.

For topic knowledge, significant differences were revealed for: effort, $F(2,7837)=29.87$, $p<.05$; difficulty, $F(2,7837)=5.48$, $p<.05$, and the following NASA-TLX dimensions: overall workload, $F(2,7837)=141.6$, $p<.05$, effort, $F(2,7837)=66.63$, $p<.05$, frustration, $F(2,7837)=90.21$, $p<.05$, mental demand, $F(2,7837)=151.67$, $p<.05.05$, physical demand, $F(2,7837) = 719.25$, $p<.05$, temporal demand, $F(2,7837)=120.39$, $pp<.05$, and performance $F(2,7837)=122.84$, $p<.05$. Significant post-hoc comparisons showed that users with "almost no" topic knowledge reported higher effort but lower difficulty than those with "some" topic knowledge. Users with "some" topic knowledge reported higher overall workload, mental demand, NASA-TLX effort, physical demand, temporal demand, and performance. Finally, users with "almost no" topic knowledge reported the highest frustration. It is important to note, that as the "expert" category comprised only a single participant ($n = 1$), it was excluded from statistical analyses and is therefore not discussed in the results. Table 9.7 shows the descriptive statistics for each dependent variable for the three levels of motivation and two levels of topic knowledge.

		Motivation			Topic Knowledge	
Dependent Variable		**Not motivated**	**Somewhat motivated**	**Motivated**	**Almost nothing**	**Same as most people**
	M	43.86	45.23	44.12	48.04	42.83
Effort	**Mdn**	42.0	41.0	44.0	50.0	40.0
	SD	31.50	31.61	30.30	32.27	30.39
	M	26.24	26.20	28.49	25.84	27.65
Difficulty	**Mdn**	20.0	20.0	21.0	19.0	20.0
	SD	25.13	24.52	25.60	24.96	25.17
Judgement Time	**M**	11.97	11.20	11.98	12.24	11.40
	Mdn	5.20	5.44	5.11	5.43	5.19
	SD	21.46	18.25	23.68	19.76	21.58
Click Count	**M**	1.43	1.40	1.43	1.44	1.41
	Mdn	1.0	1.0	1.0	1.0	1.0
	SD	0.85	0.99	0.92	0.90	0.95
Mental Demand*	**M**	53.69	58.64	61.86	52.71	61.21
	Mdn	55.0	65.0	70.0	63.0	70.0
	SD	27.03	21.70	26.83	27.55	23.64
Physical Demand*	**M**	22.49	23.23	37.14	22.84	29.95
	Mdn	24.0	17.0	32.0	15.0	28.0
	SD	16.66	17.82	28.62	23.76	21.45
Temporal Demand*	**M**	40.69	36.60	40.00	35.62	39.39
	Mdn	40.0	35.0	39.0	33.5	39.0
	SD	29.78	20.55	24.00	22.36	24.32
Performance*	**M**	52.17	59.64	45.72	51.50	53.54
	Mdn	62.0	60.0	42.0	51.5	57.0
	SD	26.85	22.79	42.0	59.82	62.45
Effort*	**M**	56.63	63.71	62.70	59.82	62.45
	Mdn	58.0	65.0	69.0	60.0	65.0
	SD	22.50	20.38	23.37	24.04	21.40
Frustration*	**M**	39.09	34.97	36.05	39.86	34.13
	Mdn	40.0	30.0	32.0	35.0	32.0
	SD	25.60	24.78	27.69	26.88	25.82
Overall Workload*	**M**	44.12	46.13	47.24	43.73	46.78
	Mdn	50.33	49.17	47.33	44.00	47.33
	SD	16.26	12.02	20.37	16.99	16.46

Table 9.7: Mean (M), Median (Mdn), and Standard Deviations (SD) for dependent variables by level of motivation and topic knowledge. Note that as the "expert" topic knowledge category contained only one participant ($n = 1$), descriptive statistics are not provided, as such a limited sample does not permit meaningful interpretation.
* NASA-TLX measures

.

9.7 Replication Study

Following the completion of Study 3 outlined above, a replication study was conducted. Few replication studies are conducted in ISR, despite their importance for maintaining scientific rigour, reliability, and enhanced external validity [189]. Therefore, to

strengthen and justify the claims regarding the validity of the multi-stage relevance judgement model and user effort and load, it seemed appropriate to examine whether a second study would produce the same results.

The replication focused on Study 3 because it represented the most comprehensive and refined iteration of the experimental design. Unlike Studies 1 and 2, which were primarily exploratory and iterative in nature, Study 3 addressed all five sub-research questions and incorporated methodological improvements informed by the earlier studies. Specifically, Study 3 included more granular measures of effort and load, the additional *"partially"* response option, and the integration of measures such as click count. Replicating this design therefore provided the strongest basis for assessing the robustness and generalisability of the findings. In contrast, replicating Study 1 or 2 would have limited theoretical and methodological value, as these earlier studies did not yet fully operationalise the multi-stage relevance judgement model or encompass the complete set of research questions.

The replication study followed the exact same experimental design as the first study, with no variations. Users were again recruited via Prolific and paid at a £9.50 hourly rate. Specific screening criteria was set on Prolific to ensure that those who participated in the first study could not participate in the second study. For the replication study, 199 participants were crowd-sourced via Prolific and completed the experiment: 116 males, 82 females, and 1 non-binary/third gender. Participants were mostly between the ages of 18-49 (N=145) with the remaining participants aged between 50-79 (N=54), and the majority held a Bachelors level qualification or above (N=129). A total of 6619 annotations were used in the analysis.

The following section will provide a comparative overview and key findings from Study 3 and the follow-up replication study.

The results from the replication study are provided in more detail in the Appendix.

9.8 Overview and Discussion of Study 3 and Replication Study Findings

- **HL-RQ3a: Effort and Load between Document Judgement Stages:** The results for both studies were similar, both showing that the *relevance* stage judgements required significantly more user effort compared to the other three judgement stages, and that *novelty* and *usefulness* judgements took significantly longer than *relevance* judgements. The finding that *relevance* judgements required the highest user effort, while *novelty* and *usefulness* judgements took longer, has several important implications. First, it suggests that cognitive demand and task duration are not necessarily aligned, judgements that feel effortful may not always take the most time, and longer decision times do not automatically indicate greater perceived effort. Second, the consistently higher effort associated with *relevance* judgements suggests that these decisions may involve more active cognitive integration, such as aligning document content with relevance criteria, than other judgement types. Finally, these results underscore the importance of considering multiple measures when evaluating task design and user experience, as relying on either metric alone may provide an incomplete understanding of the cognitive demands involved in document evaluation.

- **HL-RQ3b: Effort and Load between Document Judgement Ratings:** "Yes" ratings were easiest at the *aboutness* stage but required more effort for *novelty* and *usefulness* judgements. "Partially" ratings highlighted increased effort, particularly for *relevance* judgements, potentially reflecting the cognitive load of resolving uncertainty. Interestingly, "no" ratings at the *aboutness* stage demanded the most effort, suggesting that rejecting documents based on surface-level criteria may require closer scrutiny. These results demonstrate that both judgement type and outcome influence user effort and load, highlighting the importance of developing fine-grained models of cognitive demand in ISR tasks.

- **HL-RQ3c: Relationships between Measures:** The results from both stud-

ies were very similar. Both studies showed moderate significant relationships between effort and difficulty, and click count and judgement time. Weak significant relationships were also observed between effort and judgement time; difficulty and judgement time; and minimum and maximum effort and difficulty with the NASA-TLX dimensions. No significant correlations emerged for mean values of effort, difficulty, time, click-count and the NASA-TLX dimensions, suggesting that post-task workload assessments may not fully reflect users' moment-to-moment experience.

- **HL-RQ3d: Effort and Load across Task Duration:** Both studies showed that effort, difficulty, judgement time, and click count all differed significantly over time, generally showing a downwards trend as the user judges more documents. Results for both also further demonstrated that the first document judged required more user effort, longer judgement time, and a higher number of clicks compared to the remaining documents, while not significant, users reported difficulty was the highest for the first document judged. These patterns indicate that cognitive demand is higher at the beginning of a task and gradually declines, highlighting the importance of considering temporal dynamics when measuring user effort and load in ISR tasks.

- **HL-RQ3e: Topic Knowledge and Motivation:** Both studies revealed that motivation and topic knowledge significantly influenced post-task workload measures. Highly motivated users reported the highest overall workload, effort, and mental demand, while less motivated users experienced the most frustration, and moderately motivated users reported the highest performance. Differences between the studies were mainly observed in the in-task measures: Study 3 showed significant effects of motivation on difficulty only, whereas the replication study found significant differences across effort, difficulty, judgement time, and click count. Regarding topic knowledge, users with minimal knowledge reported higher frustration, while those with some knowledge experienced higher mental demand and performance. Although the studies showed variation in results, the findings

generally suggest that motivation and topic knowledge significantly shape user experience during ISR tasks. Highly motivated users appear to allocate more cognitive resources, reporting greater effort and mental demand, while less motivated users experience higher frustration. Similarly, users with some topic knowledge demonstrate higher mental demand but also better performance, indicating that prior knowledge can influence both cognitive load and task efficiency.

9.9 Conclusion

Study 3 and the replication study show that user effort and cognitive load in ISR tasks are shaped by a combination of judgement type, rating outcome, temporal dynamics, and user characteristics. *Relevance* judgements required the most effort, while *novelty* and *usefulness* judgements took the longest, demonstrating that perceived effort and task duration do not always align. Effort also varied by rating: "yes" ratings were generally easiest at the *aboutness* stage, "partially" ratings increased cognitive demand, and "no" ratings at *aboutness* unexpectedly required more effort, highlighting the influence of decision outcome. Across the task, all measures declined, with the first document demanding the most resources, illustrating temporal effects on effort and load. Motivation and topic knowledge further modulated user experience, with highly motivated users allocating more cognitive resources, less motivated users experiencing greater frustration, and prior knowledge influencing both mental demand and performance. These findings underscore the complex interplay between task structure, judgement stage, and user factors, and emphasise the importance of using multiple measures and individual factors to capture the nuances of user effort and load. Finally, the replication study produced very similar results to Study 3, providing strong validation of these patterns. This consistency reinforces the reliability of the multi-stage relevance judgement model and suggests that the observed effects are robust across participant samples, supporting the broader generalisability of the findings for understanding user effort and load in ISR tasks.

Part IV

Final Discussion and Conclusion

Chapter 10

General Discussion

The final chapter of this thesis will provide a summary and discussion of the results reported in both the theoretical and empirical work. More specifically, this chapter will highlight the implications of the findings on the Information Seeking and Retrieval (ISR) research landscape, discuss the limitations of this thesis work, and propose several suggestions/recommendations for future work in this important area.

10.1 Discussion

As observed in Chapter 3, cost, effort, and load (CEL) definition and measurement have encountered key challenges within ISR, namely in relation to lack of definitions, weak theoretical underpinnings, and operational overlap between constructs and measures. Subsequently, the field has struggled to advance our cohesive understanding of these constructs within an ISR context. In this thesis, three user studies and a replication study (presented in Chapters 7, 8, and 9) were conducted using a multi-stage document relevance judgement task outlined in Chapter 4. In this task, the primary focus was on the examination of a variety of commonly used effort and load measures. This involved investigating the relationships between them, how the measures vary over time, and the influence of individual differences, to gain a broader insight into the characteristics of these constructs within an ISR context. While the key focus of the empirical work was on the examination of effort and load measures, the nature of the task also allowed

the investigation of how effort and load differed between document judgement stages and between document judgement ratings.

Both the theoretical and empirical work conducted for this thesis revealed a number of novel and important areas of discussion. In this section, these findings are discussed and guided by the high-level research questions (Section 1.3) which are reiterated below.

HL-RQ1 (Theoretical): How has cost, effort, and load (CEL) been defined and measured within Information Seeking and Retrieval (ISR)?

- (a) How have CEL, and their related constructs been defined within ISR?
- (b) Which methods have been used/proposed to measure CEL, within ISR?
- (c) What are the relationships between the different definitions of CEL and the methods used to measure them?

HL-RQ2 (Theoretical Model): How can effort, and load be integrated into a multi-stage model of relevance judgement?

- (a) What are the key stages involved in relevance judgement, based on existing theory?
- (b) How do effort, and load theoretically influence each stage of relevance judgement?
- (c) How can these theoretical relationships be structured to form a coherent multi-stage model?

HL-RQ3 (Empirical): How can effort and load be measured within the ISR sub-task of multi-stage document judgement, as defined by the theoretical model?

- (a) How do different stages of document judgement vary in terms of user effort and load?
- (b) To what extent is effort and load influenced by the type of judgement rating (i.e.,"no", "partially", "yes") between document judgement stages?

- (c) What are the operational relationships between effort and load?
- (d) How do measures of effort and load change over time during the task?
- (e) To what extent do user characteristics influence measures of effort and load?

10.1.1 HL-RQ1: Development of a Conceptual Framework of CEL and Measures

This thesis has addressed HL-RQ1 by critically examining how CEL have been defined and measured within ISR, and by proposing ways to advance conceptual and methodological clarity. The systematic review showed that while CEL has been investigated for decades, progress in developing theoretical foundations has been limited. Many studies rely on intuitive or face-valid understandings of CEL, often without explicit definitions, leading to inconsistent use of measures and ambiguity in what is being captured. Search interaction metrics such as clicks, dwell time, and task duration remain the dominant approach, yet these are frequently used interchangeably across constructs (e.g., cost, effort, satisfaction, interest), undermining construct validity. Self-report measures of effort and workload have become common but are often applied in a "quick and dirty" way, providing only general insights rather than identifying the specific points of demand within the search process. Cognitive load, although theoretically rich, has been the least examined construct and has seen a decline in attention despite its potential to pinpoint moments of heightened demand that impact performance and system usability. In response to these challenges, the thesis has advanced a set of working definitions and a conceptual framework for CEL, integrating both cognitive and physical resource demands, thereby offering a more comprehensive and theoretically grounded basis for future research.

Taken together, these contributions demonstrate that progress in defining and measuring CEL requires not only methodological refinement (e.g., triangulating across multiple methods) but also stronger theoretical grounding to guide measure selection, interpretation, and comparability across studies. The iterations of the user studies conducted as part of this thesis further highlighted the challenges of measuring effort

and load in practice. They showed that "quick and dirty" methods are not sufficient; valid measurement of effort and load requires rigorous design, careful operationalisation, and sensitivity to the dynamics of the task. Without such planning, measures risk oversimplifying or misrepresenting the demands that effort and load are intended to capture.

To address these issues, the three (plus replication study) empirical studies presented in this thesis (Chapters 7, 8, and 9) built directly on the systematic review's insights by combining multiple measures, including interaction metrics and validated self-report scales, within controlled experimental contexts. This approach enabled a more fine-grained examination of effort and load across different document judgement stages, demonstrating both the promise and the methodological complexity of measuring these constructs rigorously in ISR research.

10.1.2 HL-RQ2: Development of a Multi-Stage Model of Relevance Judgement

The multi-stage relevance judgement model developed in this thesis represents a key theoretical contribution to understanding how effort and load operate within one of the most fundamental processes of ISR: the evaluation of document relevance. Drawing on existing relevance frameworks, the proposed model distinguishes between topical, cognitive, and situational dimensions of relevance, mapping these onto four judgement stages, *aboutness*, *relevance*, *novelty*, and *usefulness*. This structured approach provides a more fine-grained account of the user's evaluative process than traditional "single-stage" views of relevance judgement, which often collapse complex decision-making into a binary relevant/not-relevant outcome.

The value of the model lies not only in clarifying the different cognitive demands involved across stages, but also in highlighting how effort and load may vary dynamically within a single judgement task. By situating the model within cognitive load theory, this thesis demonstrates that the demands placed on users are not static, but differ between surface-level assessments (*aboutness*) to more integrative and task-oriented evaluations (*usefulness*). This perspective advances ISR research beyond earlier work

that tended to associate effort and load with either broad search tasks or isolated document features (e.g., readability), without accounting for the internal dynamics of cognitive demand within the judgement process itself.

To summarise, the multi-stage relevance judgement model provides a theoretically grounded framework for examining how effort and load shape document evaluation in ISR. By offering a controlled context for analysing staged cognitive demand, it enables more precise measurement while also underscoring the need for future refinements to account for the flexibility and non-linearity of real-world relevance judgements.

10.1.3 HL-RQ3: Measuring Effort and Load within an ISR Task

(a) Effort and load differences between document judgement stages

Studies 2 and 3, including the replication, examined whether effort and load varied across the stages of document judgement. The multi-stage relevance judgement model proposed in Chapter 4 hypothesised a progressive increase in demand from *aboutness* to *usefulness*. This was only partially supported. Across both designs, *aboutness* judgements consistently required the least effort, while *relevance* judgements emerged as the most demanding stage. The finding that *relevance* judgements carried the greatest cognitive burden aligns with Yilmaz et al., [25], who suggest that users front-load cognitive resources to determine whether a document warrants further consideration, conserving effort once a threshold of futility is reached, a behaviour described as "tolerance to irrelevance" [190]. This interpretation could help explain the Study 3 pattern, where effort declined at the later *novelty* and *usefulness* stages. In Study 2, the between-subjects design meant that each stage was judged in isolation, so patterns of resource reallocation could not be observed directly. Nonetheless, the higher effort ratings for *relevance* judgements in this design remain consistent with the idea that users invest the most resources at this critical decision point. Taken together, these findings may suggest that the "*relevance*" stage represents a pivotal moment in the multi-stage model, where intrinsic cognitive load peaks before demands diminish in later stages.

Another consideration is that the increased demand observed at the *relevance* stage may have been influenced by a potential confounding factor in the task design. In the

relevance condition, participants were required to assess three relevance criteria before giving their overall relevance judgement, whereas the *aboutness*, *novelty*, and *usefulness*, stages each involved only a single judgement question. This additional demand may have inflated effort and load ratings for the relevance stage. Importantly, this does not undermine the theoretical expectation that higher-level judgements, *novelty* and *usefulness* should impose greater intrinsic cognitive load due to the more complex cognitive processing they require. Rather, it suggests that the observed peak in effort at the *relevance* stage may reflect an interaction between the intrinsic cognitive demands of the judgement process and the confounding influence of increased task requirements in that condition.

(b) Effort and load differences between ratings

Importantly, both studies also revealed that effort and load were influenced not only by the type of judgement but also by the specific rating outcome. In Study 2, "*yes*" ratings for *relevance* were judged as most effortful and difficult, while "*no*" ratings for *aboutness* required the least effort. In Study 3, however, *"yes"* ratings for *novelty* and *usefulness* judgements placed the highest overall demands, and "*no*" ratings for *aboutness* required more effort than in other stages. Together, these findings suggest that cognitive demand does not rise in a straightforward linear fashion but instead reflects a more complex interaction between judgement type, rating outcome, and experimental design. This has important implications for both theory and practice: the multi-stage model should be refined to account for peaks of effort at specific stages (notably *relevance*), while future research and system design should consider how task structure and judgement framing shape the way users allocate their cognitive resources during document evaluation.

(c) Relationships between effort and load measures

A key challenge facing CEL measurement within ISR is the lack of standardised or gold standard methods to measure constructs like effort and load. Rather, a wide range of different methods are used to measure the same construct, or the same methods are used to measure different constructs. Consequently, it is unclear which construct is ac-

tually being measured. To investigate the *content* validity of widely used measures, the empirical work undertaken in this thesis employed multiple measures of effort and load within a single-study in order to assess the relationships between them. In alignment with previous research, all three studies (in addition to the replication study) found a moderate relationship between the single-item measures of effort and difficulty. The positive relationship between effort and difficulty aligns with well-established theory from the field of Psychology, where the amount of effort expended during a task will increase proportionally alongside the level of perceived difficulty [188].

In Study 3 (and the replication study), a moderate relationship was observed between click count and judgement time. As highlighted in the systematic review discussed in Chapter 3, click count and judgement time are often used within ISR as a proxy for effort and cognitive load, respectively, and if we were to draw inferences about the relationship between them independently, it would seem that in Study 3 these measures may reflect effort and cognitive load. However, due to the triangulation of methods used in this study, we can see that click count and judgement time are not significantly correlated with the more robust measures of effort and load (i.e., validated self-report scales). If we consider these measures in the wider context of the study rather than independently, users were required to click on the document link in order to view it, therefore it would be anticipated that as the number of clicks increased, judgement time will also likely increase due to the user reading the document. These findings may support the argument that time as a measure of cognitive processing is "too simplistic" in nature, and if it is to be used, it is also important to consider the influence of other contextual variables such as the task, topic, and the individual characteristics of the user [132]. To get a greater understanding of the relationship between judgement time and the other measures used, it would have been useful to separate "viewing time" and response time, as each is likely to require different levels of cognitive processing. For example, the time taken to read the document vs. preparing and providing a response/judgement after reading the document are likely to differ, with the latter likely to require additional processing. These different time variables are also hypothesised to relate differently to the users self-reported effort - highlighting

the importance of selecting which time variables we use and for which purpose [191].

Another challenge which was highlighted in Chapter 3 relates to the administration of effort and workload measures *after* task completion. Several issues were raised with this within the systematic review discussion, such as the reliance on memory of the task and the vulnerability of these measures to recall biases such as primacy/recency effects or peaks/deviations in demand throughout the task [130]. To investigate this issue further, the three studies conducted for this thesis examined the extent to which the post-task measures (i.e., NASA-TLX) reflected the during-task measures (effort, difficulty, judgement time). Maximum and minimum values of the during-task measures were also included in the analysis to examine whether the post-task measure was reflective of peaks or troughs in demand during the task. In Study 1, no significant relationships were observed between the during-task measures of effort and difficulty and overall workload. This finding was surprising considering previous research which suggests that self-report measures of effort and workload are essentially different measures of the same construct [43]. This finding highlights the risk of using post-task measures of workload, such as the NASA-TLX. Load is considered to fluctuate on a moment-to-moment basis in response to the demands imposed by the task. Administering the NASA-TLX at the end of the task requires users to retrospect and integrate their memories of the task to form an overall interpretation of their experienced workload, and the integration of these memories are likely to be influenced by peaks or deviations of demands rather than an overall account [130]- leading to to the question of how user workload is actually being assessed post-task? Study 2 however, revealed moderate relationships between during-task maximum effort and maximum difficulty with post-task ratings of mental demand and overall workload. These findings may suggest that users are basing their post-task assessments of load on their memory of peaks in task demand.

Taken together, these findings underscore the complexity of measuring effort and load within ISR and highlight the limitations of relying solely on either behavioural proxies (e.g., time, clicks) or post-task assessments (e.g., NASA-TLX). While behavioural measures may capture aspects of task interaction, they cannot be assumed

to directly reflect cognitive effort or load without consideration of contextual factors. Similarly, post-task workload measures appear to capture users' memory of peak task demands rather than the full spectrum of their experience, raising questions about their validity for evaluating dynamic task processes. These insights emphasise the need for multi-method, context-sensitive approaches to effort and load measurement in ISR, and point toward the importance of refining how and when effort and load are assessed if we are to achieve meaningful, standardised evaluation practices in the field.

(d) Effort and load across task duration

When examining how in-task measures of effort and load varied over time (i.e., from the first document judged to the last), Studies 1 and 2 found no significant differences across the duration of the task. In Study 1, users provided ratings of effort and load only once per document, after making four judgements, which may have averaged out any within-document fluctuations and masked changes across the task. In Study 2, the between-subjects design meant that users made only one type of judgement per document, limiting opportunities to observe variation across multiple stages or time points. In contrast, Study 3 (and the replication study) used a within-subjects design where users completed all four judgements per document, with effort and load ratings collected after *every* judgement. This design provided finer-grained data, and revealed significant downward trends in effort, difficulty, click count, and judgement time as users progressed through the task.

These findings raise important considerations for how user effort and load are conceptualised within ISR. The downward trend observed in Study 3 may reflect a form of task fatigue or habituation, whereby users gradually reduce their investment of cognitive resources as the task continues. Alternatively, it could indicate increasing efficiency, with users learning how to navigate the task more effectively and thereby requiring fewer resources. Distinguishing between these interpretations is crucial, as the former would suggest a decline in task engagement that could undermine evaluation outcomes, whereas the latter would suggest an adaptive process that reflects realistic user behaviour in prolonged search contexts.

Importantly, these temporal effects also have direct implications for the multi-stage relevance judgement model proposed in Chapter 4. If users' effort and cognitive load decline as the task unfolds, this may interact with the complexity of later judgement stages (e.g., *novelty* and *usefulness*). What might appear as lower effort in these stages may in fact reflect task-level fatigue or disengagement, rather than stage-specific complexity. Conversely, if the downward trend reflects increasing efficiency, then the effort/load attributed to each judgement stage may be partially confounded with learning effects across the task. This highlights the need for future work to disentangle stage-specific cognitive demand from broader temporal dynamics, ensuring that interpretations of the model are not conflated with task progression effects.

(e) Influence of individual differences

Studies 1 and 3 (including the replication study) incorporated measures of individual differences into the experimental design. Study 1 examined the influence of cognitive abilities (i.e., working memory and perceptual speed); and Study 3 investigated the influence of user motivation and topic knowledge on user effort and load.

In Study 1, users in the "high" working memory category exerted greater effort and reported lower performance, but also reported lower overall workload, temporal demand, mental demand, and took less time to judge documents than their "low" working memory counterparts. Previous studies within ISR have also shown that high-working memory users exert greater effort during the search process, more specifically, they were less likely to satisfice during periods of high demand [68]. In this case, users with low-working memory were found to alter their behaviour by decreasing the number of result documents they opened. The results from Study 1 also support previous work which found that high-working memory users work at a faster pace during the search process than users from the low-working memory group [9]. Users with "high" perceptual speed exerted greater effort and reported higher overall workload and mental demand, but experienced less physical demand and frustration compared to those with "low" perceptual speed. Previous research has shown that low-perceptual speed users may encounter more difficulty scanning for relevant information, which could explain

why low-perceptual speed users in Study 1 experienced higher frustration and physical demand [102]. High-perceptual speed users have been found to engage with results with greater intensity (more clicks, more documents viewed, type longer queries) than low-perceptual speed users [29].

In Study 3 (and the replication study), users categorised as "motivated" reported experiencing the highest overall workload, effort, and mental demand. This finding was unsurprising, considering motivation is commonly defined as the "driver" behind human actions, where individuals exert sustained effort to achieve their goals and fulfill their needs [192]. Conversely, those labelled as "not motivated" reported the highest levels of frustration. Frustration typically arises when perceived effort outweighs perceived reward. Individuals with high levels of motivation are more likely to find high-effort situations enjoyable and engaging, as they are intrinsically driven by the task itself, focusing on personal achievement and goal attainment rather than external rewards like monetary gain. Since Study 3 did not provide any tangible extrinsic rewards or outcomes within the task, individuals with low intrinsic motivation may perceive high-effort situations as frustrating rather than engaging.

Users who indicated having "no topic knowledge" also reported the highest frustration levels and lowest performance. Previous research has indicated a correlation between assessors topic knowledge and their ability to evaluate documents on a given topic, with significantly higher precision observed among assessors with extensive topic knowledge compared to those with limited knowledge [143]. However, individuals with "some" topic knowledge reported experiencing higher mental demand. Those with greater expertise are likely to persist more in their quest for relevant information [83], which could explain why individuals with higher levels of topic knowledge perceived greater mental demand. Other constructs, such as *interest*, have also been proposed as mediators of this relationship, with suggestions that interest may influence relevance judgements by enhancing a user's ability to perceive connections with the topic [193].

The results relating to individual differences derived from this thesis work may have important implications for ISR research. Understanding of user's specific cognitive abilities, motivation, or topic knowledge may lead to better inferences about effort and

load derived from search interaction data. However, it is worth noting that there is likely a much wider collection of individual factors at play which may influence user effort and load during the search process. Nonetheless, the findings highlight a number of potential individual factors that can influence user effort and load, and should therefore be taken into consideration as potential confounding factors when interpreting effort and load results in the context of document judgement.

10.1.4 Limitations

Theoretical limitations

The systematic literature review (Chapter 3) has several limitations which would be beneficial to highlight, particularly for future replicability purposes. Firstly, it is important to note that the review does not claim to encompass every article pertaining to CEL within ISR. Instead, it focuses solely on CEL studies that met the inclusion criteria. Despite best efforts, there remains the possibility that certain articles were overlooked due to limitations inherent in the database search process or human error. Secondly, coding the articles, particularly concerning definitions and measures, was not always straight-forward. Some definitions were not explicitly stated, making them challenging to identify and extract compared to studies with clear definitions. Similarly, certain studies did not specify the units of measurement used to indicate the CEL construct, particularly in the results section of the article. In such cases, the qualitative nature of coding may imply that the reported results are less reproducible than those derived from studies with more explicit definitions and measures. Lastly, it is worth noting that the review did not address other constructs related to CEL, such as difficulty and complexity, in the working definitions. As noted by Wildemuth and colleagues [169], difficulty and complexity face similar challenges to CEL in ISR, including inconsistent definitions and ambiguity in levels within tasks. Since these constructs are closely related, clarifying their definitions is a necessary step toward strengthening the measurement of CEL.

It is important to acknowledge limitations associated with the novel multi-stage relevance judgement model introduced in Chapter 4 this thesis. The model was designed

to examine measures of effort and load within a theoretically robust framework, manipulating intrinsic load while controlling for other load types, such as extraneous load. Although hypotheses regarding document judgement stages and their relative complexity were grounded in previous research, the empirical studies were effectively testing both the model and the sensitivity of the effort and load measures simultaneously. This dual focus complicates interpretation, as it is difficult to determine whether observed effects reflect the structure of the model, the properties of the measures, or an interaction between the two. Consequently, caution is needed when generalising these findings beyond the controlled experimental context. This limitation underscores the importance of further validation with independent measures and alternative experimental designs to isolate intrinsic load effects. To address this, a replication study was conducted, which validated the findings from Study 3 and provided additional confidence in the robustness of the model and the reliability of the effort and load measures.

Empirical limitations

The studies conducted for this thesis were subject to several limitations. Firstly, a significant constraint relates to the recruitment of participants. As a result of the Covid pandemic, conducting laboratory-based studies was unfeasible for the majority of the duration of my doctoral work. Therefore online recruitment of participants through crowd-sourcing worked as an alternative. Consequently, online recruitment of participants through crowd-sourcing served as a viable alternative. While utilising crowd-sourcing for participant recruitment can alleviate certain challenges linked with laboratory-based studies, such as temporal and monetary costs, as well as the absence of diversity among the user population, it also presents potential drawbacks [172]. For instance, the artificial nature of the task implies that participants lack a genuine real-life information need driving their query. Instead, they're instructed to envision themselves as hypothetical users, which might lead to varied motivations influencing their performance and potentially impacting the results.

The studies used in this thesis were very narrow and controlled in their design. While this allowed for a more robust assessment of effort and load measures, users

were afforded minimal control over the systems functionality or document interactions. To enhance comprehension of how effort and load may impact user behaviour during the document judgement process and, consequently, to construct a predictive model of user relevance judgement behaviour, it may have been beneficial to offer users greater functionalities during the judgement process. This might have included options such as the ability to abandon a document or revisit previously examined documents.

A potential limitation of the present studies is that the *relevance* stage involved more criterion requirements than the other judgement stages, as participants considered three relevance criteria before providing their overall judgement. This may have artificially elevated effort and load ratings, complicating the isolation of intrinsic cognitive demands. Future research could address this by standardising the number of criteria across stages or separating criterion evaluations from the primary judgement, enabling a clearer assessment of how cognitive load and effort varies with judgement complexity.

Finally, it is important to consider other individual variables that may influence user effort and load beyond those examined in this thesis. For instance, constructs such as *interest* has shown positive relationship with effort within ISR studies, particularly in relation to an individuals persistence and time on task [193].

10.1.5 Recent Developments in the Field

Since the commencement of this research, a number of studies have emerged that contribute new insights to the understanding of CEL within ISR. Notably, some of these recent publications have cited earlier work presented in this thesis, highlighting its contribution to ongoing discussions in the field. Several recent studies [194–198] have drawn on the definitions of CEL proposed in the systematic review, reinforcing the value of these definitions in establishing a common language and conceptual framework for discussing and investigating these constructs within ISR. Notably, these studies focus primarily on exploring the cost-benefit dynamics of user-system interactions. As highlighted in the systematic review, earlier literature frequently used the terms effort and cost interchangeably. However, the adoption of clearly delineated definitions in recent

research has enabled cost to be treated as a distinct construct, thereby contributing to a more nuanced understanding. Although this thesis work did not empirically examine cost, these recent contributions offer important insights into its operational characteristics and enrich the field's conceptual clarity. Other recent studies that have cited this thesis work have extended its scope in meaningful ways. For example, Gwizdka [199] provides a critical review of the terminology and operationalisation of mental workload, arriving at conclusions that closely align with those presented in this thesis. Specifically, Gwizdka underscores the importance of clearly defining key constructs and systematically evaluating their measurement within the field.

It is also noteworthy that several recent publications in the field have further reinforced the findings and conclusions presented in this thesis. Most notably, Babaei and colleagues [200], through a systematic review, critically examine the application of the NASA-TLX within the domain of Human-Computer Interaction. Their analysis highlights significant issues related to the definition of mental workload and the use of the NASA-TLX as a measurement tool. Consistent with the conclusions drawn in this thesis, the authors advocate for a more cautious application of the NASA-TLX and emphasise the need for clearer, more precise definitions of workload within the field.

These recent works have expanded the theoretical and methodological approaches in the field, offering complementary or, in some cases, novel perspectives to those presented in this thesis. While these developments post-date the initial design and data collection phases of this study, they underscore the continued relevance of the research questions addressed here and highlight opportunities for further investigation. Incorporating these perspectives into future work could enhance the generalisability, timeliness, and theoretical grounding of research in this domain.

10.1.6 Recommendations and Future Research

Further development of working definition framework

The working definition framework presented in this thesis serves as an initial foundation for establishing shared terminology and a common understanding of CEL within the context of ISR. As research in this domain progresses, particularly with the adoption

of more precise measurement techniques, it is anticipated that our understanding of these constructs and their characteristics will continue to evolve. It is expected that subsequent research will refine and build upon these definitions, potentially producing more concise iterations. Additionally, it is important to note that constructs related to CEL, such as difficulty and complexity, were excluded from the framework outlined in this thesis. Such constructs are a crucial aspect of CEL research within ISR, particularly with regard to the search task itself. Therefore, investigating their conceptual and operational characteristics would significantly enhance our understanding of CEL within the context of ISR.

Triangulation of measures

The theoretical work conducted in this thesis underscores key challenges associated with the use of single data collection methods to measure CEL. This issue is particularly prominent in methods like search interaction logging, where it is inherently difficult to determine what the unit of measurement actually reflects. Similarly, measures such as time on task, which lack contextual information, present comparable challenges. Methodological triangulation is a strategy can help overcome this issue - aimed at increasing the credibility and validity of research findings [30]. It involves the use of multiple data collection methods within a single study, thereby reducing the potential for bias associated with relying on a single method. For example, self-report measures could be combined with search interaction data to provide a more comprehensive understanding of the contextual factors involved.

Use of physiological measures

Given the challenges associated with the use of indirect measures of effort and load such as search interaction log metrics and self-report tools, it may be beneficial for the field to explore alternative approaches to effort and load assessment. Fields like Human Factors have increasingly advocated the use of physiological measures as a more robust alternative (refer to Charles and Nixon [48] for a review of physiological measurements). For instance, signals obtained from electrocardiography, respiratory, dermal, and blood

pressure measures have been shown to differentiate between various aspects of mental workload, such as task type, task demand, and task difficulty [48]. Similarly, Cognitive Neuroscience, have showcased the effectiveness of highly sensitive and precise physiological measures to gauge cognitive load. For example, Functional Magnetic Resonance Imaging (fMRI) allows for the direct measurement of brain resource consumption, even presenting the potential to differentiate between the three types of cognitive load [43]. While these measures hold promise for future assessment of constructs such as cognitive load, their current use in situational or naturalistic experiments remains fairly limited.

Development of ISR-specific self-report measures

The research conducted for this thesis reveals that effort and workload are frequently measured in ISR through self-report instruments. These questionnaires often rely on self-developed items with limited theoretical justification, or, in the case of workload, predominantly utilise the NASA-TLX, an instrument designed for a different domain. ISR research would significantly benefit from the development of a tailored self-report tool specifically intended to measure CEL constructs within the context of ISR. A useful self-report tool for ISR would ideally possess the ability to measure effort and load at various stages of the search task and across different levels of granularity. Additionally, it would be advantageous if such a tool could differentiate between different types of load. This capability would enable system developers to more easily identify specific components of the system or task that contribute to increased cognitive load.

Chapter 11

Conclusion

This thesis set out to address fundamental questions about how cost, effort, and load (CEL) are conceptualised and measured within the field of Information Seeking and Retrieval (ISR), and to develop empirically grounded methods to improve understanding of these constructs. Through a combination of theoretical work, and empirical studies, the research has advanced both conceptual clarity and methodological rigour in examining user effort and cognitive load during document evaluation tasks.

The systematic review revealed that despite decades of ISR research, CEL constructs have been inconsistently defined, often relying on intuitive or face-valid interpretations rather than theory-driven operationalisations. Measures followed a similar pattern, with interaction metrics, time on task, and self-report scales frequently used interchangeably across different constructs. These findings highlight a critical challenge: without clear definitions and validated measures, interpretations of CEL remain ambiguous, limiting comparability across studies and hindering cumulative knowledge building in the field.

To address this, the thesis developed a working definition framework for CEL, distinguishing between external and internal resources. Within this framework, cost is conceptualised as the demand on external resources, such as time, money, or human resource that the user spends or pays to fulfil the demands of the task, system, and context, making it less user-centric and therefore difficult to operationalise through behavioural user evaluations. Load, in contrast, refers to the quantity of resources be-

ing consumed at any given moment, while effort is defined as the actual expenditure of these internal resources. Given this distinction, the empirical studies focused on effort and cognitive load, which could be directly measured through behavioural and self-report metrics, while cost remained a conceptual construct. The framework thus provides both theoretical clarity and practical guidance for empirical investigation.

Empirical studies conducted across three experiments (including a replication study) demonstrated the nuanced dynamics of effort and load during document judgements. Findings revealed that *aboutness* judgements were consistently the least demanding, while *relevance* judgements required disproportionately higher effort, challenging assumptions of a simple linear increase in cognitive demand. Differences in observed effort and load were further influenced by judgement rating outcomes and experimental design, highlighting the complex interaction between task structure, user strategy, and cognitive resource allocation. Moreover, triangulating in-task measures (effort, difficulty, click count, and judgement time) with post-task workload assessments (NASA-TLX) revealed limitations of relying solely on post-task or log-based metrics, emphasising the importance of multi-method approaches for capturing dynamic constructs like cognitive load.

Taken together, this thesis demonstrates that advancing understanding of CEL in ISR requires a combination of clear conceptualisation, rigorous measurement, and careful attention to task and experimental design. The work contributes a validated framework for defining CEL, theoretical and empirical evidence on the distribution and determinants of effort and load across judgement stages, and insights into the strengths and limitations of common measurement approaches. These contributions provide both theoretical guidance for conceptual development and practical recommendations for measuring user effort and cognitive load in ISR tasks.

Looking forward, the findings suggest several avenues for future research. Greater attention should be given to cognitive load, including moment-to-moment fluctuations, as well as the integration of more robust physiological and behavioural measures. Experimental designs that allow users to interact with documents in more ecologically valid ways could further clarify how effort is allocated across stages and tasks. Fi-

nally, the working definitions framework developed in this thesis offers a foundation for building a more unified and cumulative body of ISR research, supporting both theory development and practical system design.

In sum, this thesis has taken important steps toward untangling the complex constructs of cost, effort, and load in ISR, providing both conceptual clarity and empirical evidence to guide future research and improve our understanding of how users engage with information.

Appendix A

Replication Study

A.1 Replication Study Results

A.1.1 Effort and Load between Document Judgement Stages:

The ANOVA revealed there were statistically significant differences between document judgement stages for document judgement time, $F(3,138)=3.1$, $p<.05$. Post-hoc comparisons show that as the user progresses up the document stages (from *aboutness* to *usefulness)*, the time taken to make a judgement increases.

While there were no significant differences observed for effort, difficulty or click count, Table A.1 shows that *relevance* judgements required the greater effort and number of clicks compared to the other three stages. This aligns with the findings from Study 3, which shows that *relevance* judgements require greater user effort compared to *aboutness*, *novelty*, and *usefulness*.

Stage	Judgement Time (s)			Effort			Difficulty			No. Clicks		
	M	**Mdn**	**SD**	**M**	**Mdn**	**SD**	**M**	**Mdn**	**SD**	**M**	**Mdn**	**SD**
Aboutness	9.58	5.27	14.25	44.24	44.0	28.46	29.35	25.0	23.95	1.34	1.0	0.73
Relevance	11.15	5.25	22.67	50.0	50.0	33.13	23.31	15.0	24.25	1.51	1.0	1.19
Novelty	11.26	4.88	22.17	44.26	42.0	28.85	29.71	25.0	24.81	1.37	1.0	0.77
Usefulness	13.32	5.60	27.85	45.59	43.0	27.84	29.71	24.0	24.85	1.44	1.0	0.87

Table A.1: Mean (M), Median (Mdn), and Standard Deviations (SD) for effort, difficulty, and number of clicks by document judgement stage.

A.1.2 Effort and Load between Document Judgement Ratings

The ANOVA showed that for"*yes*" ratings and "*partially*" ratings, there were no significant differences between stages for effort, difficulty, judgement time, or click count. However, table A.2, shows that "*yes*" ratings required the most effort compared to the other three stages, this aligns with the findings from Study 3. For "*partially*" ratings, judgements took longer as the user moves up the judgement stages, and *relevance* and *usefulness* judgements require the most effort. Finally, for "*no*" ratings there were significant differences between stages, but for difficulty only, $F(3,132)=9.79$, $p<.05$., where the *aboutness* stage was considered the most difficult document judgement compared to the other three stages. This finding was also observed in Study 3.

Stage	Rating	Judgement Time (s)			Effort			Difficulty			No. Clicks		
		M	Mdn	SD	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD
Aboutness	**0**	12.80	6.47	19.27	50.64	50.5	28.27	36.71	31.0	26.54	1.60	1.0	1.21
	1	10.59	5.74	16.05	53.16	57.0	23.56	41.70	40.0	22.37	1.39	1.0	0.77
	2	10.90	5.09	20.99	38.73	32.0	29.30	22.35	17.0	21.70	1.39	1.0	0.91
Relevance	**0**	12.61	5.39	23.47	40.98	35.0	33.82	19.59	10.0	24.43	1.67	1.0	1.42
	1	11.54	5.86	19.54	57.41	60.0	25.60	39.77	39.0	24.22	1.59	1.0	1.43
	2	13.16	5.45	30.77	51.39	58.0	33.63	20.99	12.0	22.60	1.64	1.0	1.61
Novelty	**0**	11.76	4.68	22.00	37.13	29.0	32.25	23.03	10.0	27.19	1.39	1.0	0.75
	1	11.99	6.26	15.10	53.32	60.0	25.96	40.72	40.0	22.89	1.46	1.0	0.90
	2	12.08	4.85	26.37	43.98	40.0	27.81	27.89	22.0	23.50	1.36	1.0	0.77
Usefulness	**0**	12.33	5.15	26.40	38.86	35.0	26.77	22.62	15.5	23.44	1.51	1.0	1.02
	1	15.57	6.64	25.77	57.41	61.0	24.10	43.04	41.0	23.48	1.57	1.0	1.32
	2	13.43	5.81	29.95	43.13	40.0	28.59	27.24	20.0	24.12	1.46	1.0	1.01

Table A.2: Mean (M), Median (Mdn), and Standard Deviations (SD) for judgement time, effort, difficulty, and number of clicks for document judgement rating by document judgement stage.
Note: 0 = "No"; 1 = "Partially"; 2="Yes"

A.1.3 Relationships between Measures

The Spearman correlation analysis showed moderate statistically significant correlations between: effort and difficulty, rs=.53, $p=<.05$; and click count and judgement time, rs=.36,$p=<.05$.

Weak statistically significant correlations were observed between effort and judgement time, rs= .18, $p=<.05$; and difficulty and judgement time, rs= .17, $p=<.05$. There were no significant correlations between the NASA-TLX factors and the other

measures (judgement time, effort, difficulty, click count). When the relationships between minimum and maximum effort and difficulty, with the NASA TLX dimensions and overall workload score were examined, only a weak correlation was observed between maximum difficulty and NASA-TLX effort, rs= .19,p=<.05. These findings reflect observations in Study 3.

	M	**SD**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1. Effort	45.61	29.69	_										
2. Difficulty	28.00	24.60	.53*	_									
3. Judgement time	11.29	22.20	.17*	.17*	_								
4. No. Clicks	1.42	.91	.13*	.09*	.36*	_							
5. Mental Demand**	60.45	24.76	-.07	-.02	-.09	-.07	_						
6. Physical Demand**	24.93	24.70	-.05	-.00	-.00	-.04	.41*	_					
7. Temporal Demand **	35.35	24.47	-.09	-.04	-.05	-.00	.60*	.42*	_				
8. Performance**	45.80	27.26	.08	.07	-.00	-.00	.25*	.05*	.21*	_			
9. Frustration**	27.70	25.75	-.05	-.02	.02	.02	.65*	.44*	.48*	-.03	_		
10. Effort**	62.76	23.75	-.01	-.05	-.14	-.05	.20*	.17*	.43*	.33*	.01	_	
11. Overall Workload**	42.83	16.69	-.06	-.00	-.06	-.03	.76*	.62	.82*	.44*	.55*	.63*	_

Table A.3: Means (M), Standard deviations (SD), and Spearman correlation matrix for dependent variables (n=199) *p<.05 ** NASA-TLX Dimensions

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1. Minimum Effort	_												
2. Mean Effort	.53*	_											
3. Maximum Effort	.18*	.53*	_										
4. Minimum Difficulty	.00	.00	.00	_									
5. Mean Difficulty	.13*	.21*	-.28	.00	_								
6. Maximum Difficulty	.20*	.37*	.43*	.00	.24*	_							
7. Mental Demand**	-.13	-.09	.05	.00	.05	.10*	_						
8. Physical Demand**	.01	-.10	-.03	.00	-.00	-.05	.41*	_					
9. Temporal Demand **	-.27	-.19	.00	.00	-.10	.00	.61*	.41*	_				
10. Performance**	-.02	.07	-.03	.00	.10*	.09	.25*	.04	.22*	_			
11. Frustration**	-.15	-.15	.00	.00	-.09	.04	.20*	.44*	.48*	-.02	_		
12. Effort**	-.11	-.02	.02	.00	-.00	.14*	.65*	.16*	.43*	.33*	.01	_	
13. Overall Workload**	-.17	-.14	-.02	.00	-.03	.05	.77*	.60*	.82*	.45*	.55*	.63*	_

Table A.4: Spearman correlation matrix for minimum, Mean, maximum, dependent variable values and NASA-TLX dimensions (n=199).
*p<.05
** NASA-TLX measures

A.1.4 Effort and Load across Task Duration

When the measures were examined over the duration of the task, the findings reflected those of Study 3 - with statistically significant differences observed over the duration of the task for all measures: effort, F(9,6609)=10.44, p<.05; difficulty, F(9,6609)=1.92,

$p<.05$; judgement time, $F(9,6609)=53.91$, $p<.05$; click count, $F(9,6609)=31.10$, $p<.05$. Post-hoc comparisons showed that the first document differed from the subsequent nine documents in relation to effort, click count, and judgement time, but not difficulty as was observed in Study 3. Table A.5 shows that as users progress through the task, they exert less effort, experience less difficulty, perform less clicks, and become faster at making judgements.

Document Order	Judgement Time (sec)			Effort			Difficulty			No. Clicks		
	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD
1st	30.65	19.90	34.30	54.59	60.0	27.96	30.30	23.0	25.68	1.92	2.0	1.47
2nd	15.08	8.62	19.88	48.33	50.0	28.99	27.11	20.0	24.36	1.47	1.0	0.78
3rd	12.34	6.09	30.05	46.55	44.0	29.84	27.98	20.0	24.62	1.40	1.0	0.80
4th	9.52	4.93	16.65	44.80	41.5	29.94	26.90	20.0	24.88	1.41	1.0	1.14
5th	9.19	4.61	25.68	43.41	40.0	29.03	27.92	20.0	24.17	1.35	1.0	0.84
6th	7.98	4.47	16.09	43.71	41.0	29.71	27.75	22.0	23.89	1.31	1.0	0.66
7th	6.81	4.11	11.34	41.95	40.0	30.36	25.58	19.0	23.77	1.30	1.0	0.66
8th	7.08	4.12	13.21	43.31	40.0	30.39	28.69	24.0	25.45	1.37	1.0	0.71
9th	7.06	3.82	16.56	44.30	42.0	29.06	28.36	22.0	24.09	1.30	1.0	0.71
10th	5.83	3.65	9.81	44.42	43.0	29.75	29.33	24.0	24.82	1.30	1.0	0.69

Table A.5: Mean (M), Median (Mdn), and Standard Deviations (SD) for judgement time, effort, difficulty, and click count by document order

A.1.5 Topic Knowledge and Motivation

The series of ANOVA tests revealed that for the three levels of motivation (not motivated, somewhat motivated, motivated), there were significant differences observed for effort, $F(2,7637)=6.26$, $p<.05$, difficulty, $F(2,7637)=28.60$, $p<.05$ judgement time, $F(2,7637)=16.24$, $p<.05$, and click count, $F(2,7637)=12.37$, $p<.05$. For the three levels of motivation, there were also significant differences observed for all dimensions of the NASA-TLX: overall workload, $F(2,7637)=19.66$, $p<.05$; frustration, $F(2,7637)=198.18$), $p<.05$; effort, $F(2,7637)=347.83$, $p<.05$; mental demand, $F(2,7637)=111.54$, $p<.05$; physical demand, $F(2,7637)=25.42$, $p<.05$; temporal demand, $F(2,7637)=41.03$, $p<.05$; and performance, $F(2,7637)=56.96$, $p<.05$.

Tukey post-hoc tests revealed that motivated users reported the highest effort but not motivated users reported the highest difficulty and took more time to make their document judgements. Motivated users report the highest workload, effort, and mental

Dependent Variable		Motivation			Topic Knowledge	
		Not motivated	Somewhat motivated	Motivated	Almost nothing	Same as most people
Effort	**M**	49.27	46.14	45.09	40.30	46.77
	Mdn	51.0	48.0	42.0	38.0	47.0
	SD	29.65	30.01	29.21	27.31	29.87
Difficulty	**M**	33.01	30.09	26.58	28.01	28.95
	Mdn	30.0	25.0	20.0	23.0	22.0
	SD	25.99	25.04	24.80	23.38	25.35
Judgement Time	**M**	13.20	13.81	10.58	11.59	12.43
	Mdn	13.81	5.78	4.93	5.61	5.47
	SD	10.58	26.79	21.17	17.45	24.73
Click Count	**M**	1.38	1.56	1.45	1.38	1.51
	Mdn	1.0	1.0	1.0	1.0	1.0
	SD	0.93	1.14	1.14	0.72	1.16
Mental Demand*	**M**	49.74	59.40	63.87	57.65	60.78
	Mdn	58.0	69.0	69.0	63.0	65.5
	SD	23.86	22.47	25.91	25.88	24.31
Physical Demand*	**M**	28.89	25.76	22.70	25.75	24.34
	Mdn	18.0	20.0	10.5	10.0	15.0
	SD	26.80	26.14	22.29	30.97	23.65
Temporal Demand*	**M**	31.11	37.90	33.50	34.52	35.39
	Mdn	20.0	35.0	26.0	20.0	29.5
	SD	23.48	26.10	22.18	24.07	24.35
Performance*	**M**	36.58	47.74	47.36	41.17	47.18
	Mdn	26.0	50.0	42.0	40.0	50.0
	SD	21.84	26.02	28.85	28.34	26.89
Effort*	**M**	55.79	56.68	70.25	63.74	62.40
	Mdn	60.0	60.0	70.0	67.0	67.5
	SD	25.95	21.26	23.14	26.10	23.23
Frustration*	**M**	33.63	32.44	21.19	34.39	26.68
	Mdn	30.0	25.0	10.0	20.0	17.5
	SD	15.49	26.71	24.94	30.92	24.72
Overall Workload*	**M**	39.29	43.32	43.15	43.17	42.80
	Mdn	41.83	45.58	44.50	36.00	44.75
	SD	18.22	44.50	16.18	19.20	16.14

Table A.6: Mean (M), Median (Mdn), and Standard Deviations (SD) for dependent variables by level of motivation and topic knowledge

demand. Not motivated users report the highest frustration and physical demand, and the lowest performance. Somewhat motivated users report the highest temporal demand and performance. Table A.6 shows the descriptive statistics for each dependent variable by level of motivation and topic knowledge.

Bibliography

[1] L. Azzopardi, D. Kelly, and K. Brennan, "How Query Cost Affects Search Behavior Categories and Subject Descriptors," in *SIGIR 2013 - Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013, pp. 23–32.

[2] C. Speier and M. G. Morris, "The influence of query interface design on decision-making performance," *MIS Quarterly: Management Information Systems*, vol. 27, no. 3, pp. 397–423, 2003.

[3] M. J. Cole, J. Gwizdka, C. Liu, and N. J. Belkin, "Dynamic assessment of information acquisition effort during interactive search," in *Proceedings of the ASIST Annual Meeting*, vol. 48, 2011.

[4] C. Luo, Y. Liu, T. Sakai, K. Zhou, F. Zhang, X. Li, and S. Ma, "Does Document Relevance Affect the Searcher's Perception of Time?" in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 141–150.

[5] H. Chen, H. Fan, M. Chau, and D. Zeng, "Testing a Cancer Meta Spider," *International Journal of Human Computer Studies*, vol. 59, no. 5, pp. 755–776, 2003.

[6] A. Chevalier and M. Kicka, "Web designers and web users: Influence of the ergonomic quality of the web site on the information search," *International Journal of Human Computer Studies*, vol. 64, no. 10, pp. 1031–1048, 2006.

[7] A. Edwards, D. Kelly, and L. Azzopardi, "The impact of query interface design on stress, workload and performance," in *European Conference of Information Retrieval (ECIR)*, vol. 9022, 2015, pp. 691–702.

[8] J. Gwizdka and I. Lopatovska, "The role of subjective factors in the information search process," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 12, pp. 2452–2464, 2009.

[9] B. Choi, R. Capra, and J. Arguello, "The Effects of Working Memory during Search Tasks of Varying Complexity," in *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, ser. CHIIR '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 261–265.

[10] M. McGregor, L. Azzopardi, and M. Halvey, "Untangling Cost, Effort, and Load in Information Seeking and Retrieval," in *CHIIR 2021 - Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 2021, pp. 151–161.

[11] M. McGregor, , L. Azzopardi, , and M. Halvey, "A Systematic Review of Cost, Effort, and Load Research in Information Search and Retrieval , 1972-2020," *ACM Transactions on Information Systems*, vol. 42, no. 1, p. 39, 2023.

[12] M. McGregor, "Defining and Measuring Cost, Effort, and Load in Information Retrieval," *SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, no. 2014, p. 3490, 2023.

[13] M. D. Cooper, "A cost model for evaluating information retrieval systems," *Journal of the American Society for Information Science*, vol. 23, no. 5, pp. 306–312, 1972.

[14] C. C. Kwok, O. Etzioni, and D. S. Weld, "Scaling question answering to the web," in *Proceedings of the 10th International Conference on World Wide Web, WWW 2001*, 2001, pp. 150–161.

[15] L. Azzopardi, "The economics in interactive information retrieval," in *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011, pp. 15–24.

[16] L. Azzopardi, P. Thomas, and N. Craswell, "Measuring the utility of search engine result pages: An information foraging based measure," in *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, 2018, pp. 605–614.

[17] R. B. Miller, "Response time in man-computer conversational transactions. Introductions and major concepts," *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, pp. 267–277, 1968.

[18] A. R. Dennis and N. J. Taylor, "Information foraging on the web: The effects of "acceptable" Internet delays on multi-page information search behavior," *Decision Support Systems*, vol. 42, no. 2, pp. 810–824, 2006.

[19] D. Galletta, R. Henry, S. McCoy, and P. Polak, "Web Site Delays: How Tolerant are Users?" *Journal of the Association for Information Systems*, vol. 5, no. 1, pp. 1–28, 2004.

[20] J. Brutlag, "Speed matters for google web search," *Google. June*, 2009.

[21] D. Maxwell and L. Azzopardi, "Stuck in traffic: How temporal delays affect search behaviour," pp. 155–164, 2014.

[22] Y. Mansourian and N. Ford, "Search persistence and failure on the web: A "bounded rationality" and "satisficing" analysis," *Journal of Documentation*, vol. 63, no. 5, pp. 680–701, 2007.

[23] G. Zipf, *Human Behaviour and the principle of least effort.* Reading,MA: Addison-Wesley, 1949.

[24] R. Capra, J. Arguello, and Y. Zhang, "The Effects of Search Task Determinability on Search Behaviour," in *ECIR 2017*, 2017, pp. 108–121.

[25] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey, "Relevance and effort: An analysis of document utility," in *CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management*, 2014, pp. 91–100.

[26] J. Gwizdka, "Distribution of cognitive load in Web search," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 11, pp. 2167–2187, 2010.

[27] R. González-IbáñTez, V. Proaño-Ríos, G. Fuenzalida, and G. Martinez-Ramirez, "Effects of a visual representation of search engine results on performance, user experience and effort," in *Proceedings of the Association for Information Science and Technology*, vol. 54, no. 1, 2017, pp. 128–138.

[28] H. H. Choi, J. J. van Merriënboer, and F. Paas, "Effects of the Physical Environment on Cognitive Load and Learning: Towards a New Model of Cognitive Load," *Educational Psychology Review*, vol. 26, no. 2, pp. 225–244, 2014.

[29] K. Brennan, D. Kelly, and J. Arguello, "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*, ser. IIiX '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 165–174.

[30] P. Price, R. Jhangiani, and I.-C. Chiang, "Research Methods in Psychology," in *Research Methods in Psychology*, 2nd ed. Pressbooks.com, 2015, p. 322.

[31] D. Kelly, "Methods for evaluating interactive information retrieval systems with users," *Foundations and Trends in Information Retrieval*, vol. 3, no. 1-2, pp. 1–224, 2009.

[32] D. F. Feldon, G. Callan, S. Juth, and S. Jeong, "Cognitive Load as Motivational Cost," *Educational Psychology Review*, vol. 31, no. 2, pp. 319–337, 2019.

[33] A. Westbrook and T. S. Braver, "Cognitive effort: A neuroeconomic approach," *Cognitive, Affective and Behavioral Neuroscience*, vol. 15, no. 2, pp. 395–415, 2015.

[34] A. Westbrook, D. Kester, and T. S. Braver, "What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference," *PLoS ONE*, vol. 8, no. 7, pp. 1–8, 2013.

[35] A. R. Otto and N. D. Daw, "The opportunity cost of time modulates cognitive effort," *Neuropsychologia*, vol. 123, no. May 2018, pp. 92–105, 2019.

[36] A. Shenhav, S. Musslick, F. Lieder, W. Kool, T. L. Griffiths, J. D. Cohen, and M. M. Botvinick, "Toward a Rational and Mechanistic Account of Mental Effort," *Annual Review of Neuroscience*, vol. 40, no. December 2016, pp. 99–124, 2017.

[37] H. Egeth and D. Kahneman, "Attention and Effort," *The American Journal of Psychology*, vol. 88, no. 2, p. 339, 1975.

[38] M. Inzlicht, A. Shenhav, and C. Y. Olivola, "The Effort Paradox: Effort Is Both Costly and Valued," *Trends in Cognitive Sciences*, vol. 22, no. 4, pp. 337–349, 2018.

[39] B. B. Van Acker, D. D. Parmentier, P. Vlerick, and J. Saldien, "Understanding mental workload: from a clarifying concept analysis toward an implementable framework," *Cognition, Technology and Work*, vol. 20, no. 3, pp. 351–365, 2018.

[40] J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," *Learning and Instruction*, vol. 4, no. 4, pp. 295–312, 1994.

[41] C. D. Wickens, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, no. 2, pp. 159–177, 2002.

[42] J. Sweller, "Element interactivity and intrinsic, extraneous, and germane cognitive load," *Educational Psychology Review*, vol. 22, no. 2, pp. 123–138, 2010.

[43] R. Brunken, T. Seufert, and F. G. W. C. Paas, "Measuring cognitive load," *Perspectives on Medical Education*, vol. 7, no. 1, 2018.

[44] G. Miller, "The magical number seven plus minus two." *Psych. Rev.*, vol. 63, pp. 81–97, 1956.

[45] T. de Jong, "Cognitive load theory, educational research, and instructional design: Some food for thought," in *Instructional Science*, vol. 38, no. 2, 2010, pp. 105–134.

[46] F. G. Paas, J. J. Van Merriënboer, and J. J. Adam, "Measurement of cognitive load in instructional research." *Perceptual and Motor Skills*, vol. 79, no. 1 Pt 2, pp. 419–430, 1994.

[47] M. S. Young, K. A. Brookhuis, C. D. Wickens, and P. A. Hancock, "State of science: mental workload in ergonomics," *Ergonomics*, vol. 58, no. 1, pp. 1–17, 2015.

[48] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," *Applied Ergonomics*, vol. 74, no. September 2016, pp. 221–232, 2019.

[49] M. A. Boksem and M. Tops, "Mental fatigue: Costs and benefits," *Brain Research Reviews*, vol. 59, no. 1, pp. 125–139, 2008.

[50] S. Martin, "Measuring cognitive load and cognition: metrics for technology-enhanced learning," *Educational Research and Evaluation*, vol. 20, pp. 592–621, 2014.

[51] J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning - Sweller - 2010 - Cognitive Science - Wiley Online Library," *Cognitive science*, vol. 285, pp. 257–285, 1988.

[52] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational Psychologist*, vol. 38, no. 1, pp. 63–71, 2003.

[53] P. A. Kirschner, "Cognitive load theory: Implications of cognitive load theory on the design of learning," *Learning and Instruction*, vol. 12, no. 1, pp. 1–10, 2002.

[54] R. Brünken, J. L. Plass, and D. Leutner, "Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects," *Instructional Science*, vol. 32, no. 1-2, pp. 115–132, 2004.

[55] M. Krnic Martinic, D. Pieper, A. Glatt, and L. Puljak, "Definition of a systematic review used in overviews of systematic reviews, meta-epidemiological studies and textbooks," *BMC Medical Research Methodology*, vol. 19, no. 1, pp. 1–12, 2019.

[56] C. Mulrow, D. Cook, M. O. Meade, and W. S. Richardson, "Systematic Review Series Series Editors: Selecting and Appraising Studies for a Systematic Review Selecting Studies for Systematic Reviews," *Annals of Internal Medicine*, vol. 127, pp. 531–537, 1997.

[57] D. Kelly and C. R. Sugimoto, "A systematic review of interactive information retrieval evaluation studies, 1967-2006," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 4, pp. 745–770, 2013.

[58] D. O'Connor, S. E. Green, and J. P. T. Higgins, "Defining the review question and developing criteria for including studies," in *Cochrane Handbook for Systematic Reviews of Interventions*, 1st ed., J. P. Higgins and S. Green, Eds. United States of America: John Wiley & Sons, 2008, pp. 8–94.

[59] M. J. Grant and A. Booth, "A typology of reviews: An analysis of 14 review types and associated methodologies," *Health Information and Libraries Journal*, vol. 26, no. 2, pp. 91–108, 2009.

[60] Y. Chen, Y. Liu, K. Zhou, M. Wang, M. Zhang, and S. Ma, "Does vertical bring more satisfaction? Predicting search satisfaction in a heterogeneous environment," in *International Conference on Information and Knowledge Management, Proceedings*, vol. 19-23-Oct-, 2015, pp. 1581–1590.

[61] L. Azzopardi and G. Zuccon, "An Analysis of the Cost and Benefit of Search Interactions," in *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ser. ICTIR '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 59–68.

[62] T. Vuong, M. Saastamoinen, G. Jacucci, and T. Ruotsalo, "Understanding user behavior in naturalistic information search tasks," *Journal of the Association for Information Science and Technology*, vol. 70, no. 11, pp. 1248–1261, 2019.

[63] Y. Zhang and J. Gwizdka, "Rethinking the cost of information search behavior," in *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016, pp. 969–972.

[64] J. Gwizdka and M. Cole, "Least effort? Not if I can search more," in *In Proceedings of the 5th Workshop on Human-Computer Interaction and Information Retrieval.*, vol. 2, no. L, 2011, p. 2012.

[65] G. Singer, U. Norbisrath, and D. Lewandowski, "Ordinary search engine users assessing difficulty, effort, and outcome for simple and complex search tasks," in *IIiX 2012 - Proceedings 4th Information Interaction in Context Symposium: Behaviors, Interactions, Interfaces, Systems*, 2012, pp. 110–119.

[66] J. He, M. Bron, A. De Vries, L. Azzopardi, and M. De Rijke, "Untangling result list refinement and ranking quality: A framework for evaluation and prediction," in *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, no. i, 2015, pp. 293–302.

[67] J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White, "Understanding and Predicting Graded Search Satisfaction," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 57–66.

[68] J. Gwizdka, "I Can and So I Search More: Effects Of Memory Span On Search Behavior," in *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, ser. CHIIR '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 341–344. [Online]. Available: https://doi.org/10.1145/3020165.3022148

[69] C. H. Fenichel, "Online searching: Measures that discriminate among users with different types of experiences," *Journal of the American Society for Information Science*, vol. 32, no. 1, pp. 23–32, 1981.

[70] Y.-M. Kim and S. Y. Rieh, "Dual-task performance as a measure of mental effort in searching a library system and the Web," in *Proceedings of the American Society for Information Science and Technology*, vol. 42, no. 1, 2006.

[71] L. L. Di Stasi, A. Antolí, M. Gea, and J. J. Cañas, "A neuroergonomic approach to evaluating mental workload in hypermedia interactions," *International Journal of Industrial Ergonomics*, vol. 41, no. 3, pp. 298–304, 2011.

[72] R. Nordlie Oslo and N. Pharo, "Search transition as a measure of effort in information retrieval interaction," in *Proceedings of the ASIST Annual Meeting*, vol. 50, no. 1, 2013, pp. 1–7.

[73] M. Halvey and R. Villa, "Evaluating the effort involved in relevance assessments for images," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 887–890.

[74] S. E. Crudge and F. C. Johnson, "Using the repertory grid and laddering technique to determine the user's evaluative model of search engines," *Journal of Documentation*, vol. 63, no. 2, pp. 259–280, 1 2007.

[75] P. Schmutz, S. Heinz, Y. Métrailler, and K. Opwis, "Cognitive Load in eCommerce Applications—Measurement and Effects on User Satisfaction," *Advances in Human-Computer Interaction*, vol. 2009, pp. 1–9, 2009.

[76] F. Ariza, D. Kalra, and H. W. Potts, "How do clinical information systems affect the cognitive demands of general practitioners? Usability study with a focus on cognitive workload," *Journal of Innovation in Health Informatics*, vol. 22, no. 4, pp. 379–390, 2015.

[77] L. Longo and P. Dondio, "On the relationship between perception of usability and subjective mental workload of web interfaces," in *Proceedings - 2015*

IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, vol. 1, no. December, 2016, pp. 345–352.

[78] A. Jimenez-Molina, C. Retamal, and H. Lira, "Using psychophysiological sensors to assess mental workload during web browsing," *Sensors (Switzerland)*, vol. 18, no. 2, 2018.

[79] S. Y. Rieh, Y. M. Kim, and K. Markey, "Amount of invested mental effort (AIME) in online searching," *Information Processing and Management*, vol. 48, no. 6, pp. 1136–1150, 2012.

[80] J. Gwizdka, "Effects of working memory capacity on users' search effort," in *ACM International Conference Proceeding Series*, 2013.

[81] M. L. Wilson, "Evaluating the cognitive impact of search user interface design decisions," in *Euro HCIR Workshop Proceedings*, vol. 763, 2011, pp. 27–30.

[82] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, "An eye-tracking study of website complexity from cognitive load perspective," *Decision Support Systems*, vol. 62, pp. 1–10, 2014.

[83] E. Bailey and D. Kelly, "Is amount of effort a better predictor of search success than use of specific search tactics?" *Proceedings of the ASIST Annual Meeting*, vol. 48, 2011.

[84] I. C. Wu and P. Vakkari, "Supporting navigation in Wikipedia by information visualization: Extended evaluation measures," *Journal of Documentation*, vol. 70, no. 3, pp. 392–424, 2014.

[85] M. Rath, S. Ghosh, and C. Shah, "Exploring Online and Offline Search Behavior Based on the Varying Task Complexity," in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, ser. CHIIR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 285–288.

[86] K. Ong, M. Sanderson, K. Järvelin, and F. Scholer, "QWERTY: The effects of typing on web search behavior," in *CHIIR 2018 - Proceedings of the 2018*

Conference on Human Information Interaction and Retrieval, vol. 2018-March, 2018, pp. 281–284.

[87] A. Kittur, A. M. Peters, A. Diriye, T. Telang, and M. R. Bove, "Costs and Benefits of Structured Information Foraging," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 2989–2998.

[88] M. Verma and E. Yilmaz, "Search costs vs. User satisfaction on mobile," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10193 LNCS, pp. 698–704, 2017.

[89] T. Flynn, P. A. Holohan, M. S. Magson, and J. D. Munro, "Cost effectiveness comparison of online and manual bibliographic information retrieval," *Journal of Information Science*, vol. 1, no. 2, pp. 77–84, 1979.

[90] B. Amento, W. Hill, D. Hix, R. Schulman, and L. Terveen, "Experiments in Social Data Mining," in *ACM Transactions on Computer-Human Interaction*, vol. 10, no. 1, 2003, pp. 54–85.

[91] R. Villa and J. M. Jose, "A study of awareness in multimedia search," *Information Processing and Management*, vol. 48, no. 1, pp. 32–46, 2012.

[92] S. Oksanen and P. Vakkari, "Emphasis on Examining Results in Fiction Searches Contributes to Finding Good Novels," in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: Association for Computing Machinery, 2012, p. 199–202.

[93] A. Mikkonen, "Books ' Interest Grading and Fiction Readers ' Search Actions During Query Reformulation Intervals Categories and Subject Descriptors," in *In Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2015, pp. 27–36.

[94] S. Pothirattanachaikul, Y. Yamamoto, T. Yamamoto, and M. Yoshikawa, "Analyzing the effects of document's opinion and credibility on search behaviors and belief dynamics," in *International Conference on Information and Knowledge Management, Proceedings*, 2019, pp. 1653–1662.

[95] R. Capra, J. Arguello, H. O'Brien, Y. Li, and B. Choi, "The effects of manipulating task determinability on search behaviors and outcomes," in *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, 2018, pp. 445–454.

[96] J. Kiseleva, K. Williams, J. Jiang, A. Hassan Awadallah, A. C. Crook, I. Zitouni, and T. Anastasakos, "Understanding User Satisfaction with Intelligent Assistants," in *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, ser. CHIIR '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 121–130.

[97] J. Gwizdka, "Revisiting search task difficulty: Behavioral and individual difference measures," in *Proceedings of the ASIST Annual Meeting*, vol. 45, 2008.

[98] J. Gwizdka and I. Spence, "What can searching behavior tell us about the difficulty of information tasks? A study of web navigation," in *Proceedings of the ASIST Annual Meeting*, vol. 43, 2006.

[99] J. Gwizdka, "What a difference a tag cloud makes: effects of tasks and cognitive abilities on search results interface use," *Information Research*, no. Fallows 2008, pp. 1–27, 2009.

[100] N. J. Belkin, M. Cole, and R. Bierig, "Is relevance the right criterion for evaluating interactive information retrieval," in *Proceedings of the ACM SIGIR 2008 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments.*, 2008.

[101] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu, "Different users, different opinions: Predicting search satisfaction with mouse movement

information," in *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 493–502.

[102] J. Arguello and B. Choi, "The effects of working memory, perceptual speed, and inhibition in aggregated search," *ACM Transactions on Information Systems*, vol. 37, no. 3, 2019.

[103] M. J. Cole, J. Gwizdka, C. Liu, N. J. Belkin, and X. Zhang, "Inferring user knowledge level from eye movement patterns," *Information Processing and Management*, vol. 49, no. 5, pp. 1075–1091, 2013.

[104] J. Gwizdka, "Characterizing Relevance with Eye-Tracking Measures," in *Proceedings of the 5th Information Interaction in Context Symposium*, ser. IIiX '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 58–67.

[105] J. Jiang, D. He, and J. Allan, "Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time," *SIGIR 2014 - Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 607–616, 2014.

[106] P. R. Mosaly, L. Mazur, and L. B. Marks, "Usability evaluation of electronic health record system (EHRs) using subjective and objective measures," *CHIIR 2016 - Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval*, pp. 313–316, 2016.

[107] S. Dennis, P. Bruza, and R. McArthur, "Web searching: A process-oriented experimental study of three interactive search paradigms," *Journal of the American Society for Information Science and Technology*, vol. 53, no. 2, pp. 120–133, 2002.

[108] Y. Zhang and J. Gwizdka, "Effects of tasks at similar and different complexity levels," *Proceedings of the ASIST Annual Meeting*, vol. 51, no. 1, 2014.

[109] H. A. Maior, M. Pike, M. L. Wilson, and S. Sharples, "Directly evaluating the cognitive impact of search user interfaces: A two-pronged approach with fNIRS," in *Euro HCIR Workshop Proceedings*, vol. 1033, 2013, pp. 43–46.

[110] R. Villa and M. Halvey, "Is relevance hard work? Evaluating the effort of making relevant assessments," in *SIGIR 2013 - Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013, pp. 765–768.

[111] S. Avula, J. Arguello, R. Capra, J. Dodson, Y. Huang, and F. Radlinski, "Embedding search into a conversational platform to support collaborative search," in *CHIIR 2019 - Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 2019, pp. 15–23.

[112] C. Shah and R. González-Ibáñez, "Evaluating the synergic effect of collaboration in information seeking," in *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, no. January, 2011, pp. 913–922.

[113] I. Arapakis, L. A. Leiva, and B. B. Cambazoglu, "Know your onions: Understanding the user experience with the knowledge module in web search," in *International Conference on Information and Knowledge Management, Proceedings*, vol. 19-23-Oct-, 2015, pp. 1695–1698.

[114] D. Kelly and L. Azzopardi, "How many results per page?" in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 183–192.

[115] H. Bota, K. Zhou, and J. M. Jose, "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*, ser. CHIIR '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 131–140.

[116] R. González-Ibáñez, J. L. Varela-Otárola, and C. Barrera-Pulgar, "Evaluating Body-Centered Interactions in an Image Search Task," in *Proceedings of the*

2016 ACM on Conference on Human Information Interaction and Retrieval, ser. CHIIR '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 269–272.

[117] M. Kamvar and S. Baluja, "Query Suggestions for Mobile Search: Understanding Usage Patterns," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 1013–1016.

[118] M. Dubiel, M. Halvey, L. Azzopardi, and S. Daronnat, "Interactive Evaluation of Conversational Agents: Reflections on the Impact of Search Task Design," in *ICTIR 2020 - Proceedings of the 2020 ACM SIGIR International Conference on Theory of Information Retrieval*, 2020, pp. 85–88.

[119] P. Gerwe and C. L. Viles, "User Effort in Query Construction and Interface Selection," in *Proceedings of the Fifth ACM Conference on Digital Libraries*, ser. DL '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 246–247.

[120] W. Ke, C. R. Sugimoto, and J. Mostafa, "Dynamicity vs. effectiveness: Studying online clustering for scatter/gather," in *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, 2009, pp. 19–26.

[121] J. Jiang, D. He, D. Kelly, and J. Allan, "Understanding ephemeral state of relevance," *CHIIR 2017 - Proceedings of the 2017 Conference Human Information Interaction and Retrieval*, pp. 137–146, 2017.

[122] J. Jiang, D. He, and J. Allan, "Comparing in situ and multidimensional relevance judgments," in *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 405–414.

[123] R. Capra, G. Marchionini, J. S. Oh, F. Stutzman, and Y. Zhang, "Effects of structure and interaction style on distinct search tasks," in *Proceedings of the*

ACM International Conference on Digital Libraries, no. May 2014, 2007, pp. 442–451.

[124] P. Vakkari and S. Huuskonen, "Search effort degrades search output but improves task outcome," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 4, pp. 657–670, 2012.

[125] Q. Liu, Y. Liu, and E. Agichtein, "Exploring Web Browsing Context for Collaborative Question Answering," in *Proceedings of the Third Symposium on Information Interaction in Context*, ser. IIiX '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 305–310.

[126] A. R. Ward and R. Capra, "Immersive Search: Using Virtual Reality to Examine How a Third Dimension Impacts the Searching Process," in *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1621–1624.

[127] R. Brünken, J. L. Plass, and D. Leutner, "Direct measurement of cognitive load in multimedia learning," *Educational Psychologist*, vol. 38, no. 1, pp. 53–61, 2003.

[128] J. C. de Winter, "Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective," *Cognition, Technology and Work*, vol. 16, no. 3, pp. 289–297, 2014.

[129] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Human Mental Workload*, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds. North-Holland, 1988, vol. 52, pp. 139–183.

[130] G. Matthews, J. De Winter, and P. A. Hancock, "What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures," *Theoretical Issues in Ergonomics Science*, vol. 21, no. 4, pp. 369–396, 2020.

[131] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," *Proceedings of the Human Factors and Ergonomics Society*, pp. 904–908, 2006.

[132] D. Kelly and N. J. Belkin, "Display time as implicit feedback: Understanding task effects," in *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 377–384.

[133] M. Claypool, P. Le, M. Wased, and D. Brown, "Implicit interest indicators," *International Conference on Intelligent User Interfaces, Proceedings IUI*, pp. 33–40, 2001.

[134] B. Yang and G. Jeh, "Retroactive answering of search queries," in *Proceedings of the 15th International Conference on World Wide Web*, no. 1, 2006, pp. 457–466.

[135] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White, "Evaluating Implicit Measures to Improve Web Search," *ACM Trans. Inf. Syst.*, vol. 23, no. 2, p. 147–168, 4 2005.

[136] Q. Guo and E. Agichtein, "Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior," in *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*, 2012, pp. 569–578.

[137] C. L. Barry and L. Schamber, "Users' criteria for relevance evaluation: A cross-situational comparison," *Information Processing and Management*, vol. 34, no. 2-3, pp. 219–236, 1998.

[138] P. Borlund, "The concept of relevance in IR," *Journal of the American Society for Information Science and Technology*, vol. 54, no. 10, pp. 913–925, 2003.

[139] T. Saracevic, *The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?*, 2016, vol. 8, no. 3.

[140] I. Ruthven, "Resonance and the experience of relevance," *Journal of the Association for Information Science and Technology*, vol. 72, no. 5, pp. 554–569, 2021.

[141] T. Saracevic, "Relevance reconsidered," in *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)*, 1996, pp. 201–218.

[142] Y. Xu and Z. Chen, "Relevance judgment: What do information users consider beyond topicality?" *Journal of the American Society for Information Science and Technology*, vol. 57, no. 7, pp. 961–973, 2006.

[143] I. Ruthven, M. Baillie, and D. Elsweiler, "The relative effects of knowledge, interest and confidence in assessing relevance," *Journal of Documentation*, vol. 63, no. 4, pp. 482–504, 2007.

[144] A. L. Al-Harbi and M. D. Smucker, "A qualitative exploration of secondary assessor relevance judging behavior," in *Proceedings of the 5th Information Interaction in Context Symposium, IIiX 2014*, 2014, pp. 195–204.

[145] M. D. Smucker, X. S. Guo, and A. Toulis, "Mouse movement during relevance judging," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 979–982.

[146] H. Chu, "Factors affecting relevance judgment: A report from TREC Legal track," *Journal of Documentation*, vol. 67, no. 2, pp. 264–278, 2011.

[147] T. T. Damessie, F. Scholer, and J. S. Culpepper, "The influence of topic difficulty, relevance level, and document ordering on relevance judging," in *ACM International Conference Proceeding Series*, 2016, pp. 41–48.

[148] L. Tang and S. Clematide, "Searching for legal documents at paragraph level: Automating label generation and use of an Extended Attention Mask for boosting neural models of semantic similarity," in *Natural Legal Language Processing, NLLP 2021 - Proceedings of the 2021 Workshop*, 2021, pp. 114–122.

[149] M. van Opijnen and C. Santos, "On the concept of relevance in legal information retrieval," *Artificial Intelligence and Law*, vol. 25, no. 1, pp. 65–87, 2017.

[150] J. Wang, "Accuracy , Agreement , Speed , and Perceived Difficulty of Users ' Relevance Judgments for E-Discovery," in *Proceedings of SIGIR information retrieval for e-discovery workshop*, 2011.

[151] S. Sivarajkumar, H. A. Mohammad, D. Oniani, K. Roberts, W. Hersh, H. Liu, D. He, S. Visweswaran, and Y. Wang, *Clinical Information Retrieval: A Literature Review.* Springer International Publishing, 2024, vol. 8, no. 2.

[152] B. Koopman and G. Zuccon, "Why assessing relevance in medical IR is demanding," in *CEUR Workshop Proceedings*, vol. 1276, no. July, 2014, pp. 16–19.

[153] A. A. Tawfik, K. M. Kochendorfer, D. Saparova, S. Al Ghenaimi, and J. L. Moore, "Using semantic search to reduce cognitive load in an electronic health record," in *2011 IEEE 13th International Conference on e-Health Networking, Applications and Services*, no. June. IEEE, 2011, pp. 181–184.

[154] H. Saitwal, X. Feng, M. Walji, V. Patel, and J. Zhang, "Assessing performance of an Electronic Health Record (EHR) using Cognitive Task Analysis," *International Journal of Medical Informatics*, vol. 79, no. 7, pp. 501–506, 2010.

[155] E. Asgari, J. Kaur, G. Nuredini, J. Balloch, A. M. Taylor, N. Sebire, R. Robinson, C. Peters, S. Sridharan, and D. Pimenta, "Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review." *JMIR medical informatics*, vol. 12, 2024.

[156] A. R. Taylor, C. Cool, N. J. Belkin, and W. J. Amadio, "Relationships between categories of relevance criteria and stage in task completion," *Information Processing and Management*, vol. 43, no. 4, pp. 1071–1084, 2007.

[157] A. Taylor, "Examination of work task and criteria choices for the relevance judgment process," *Journal of Documentation*, vol. 69, no. 4, pp. 523–544, 2013.

[158] Y. Shao, Y. Wu, Y. Liu, J. Mao, and S. Ma, "Understanding Relevance Judgments in Legal Case Retrieval," *ACM Transactions on Information Systems*, vol. 41, no. 3, 2023.

[159] T. Saracevic, "The stratified model of information retrieval interaction : Extension and applications," in *Proceedings of the Annual Meeting-American Society for Information Science*, vol. 34, 1997, pp. 313–327.

[160] E. Cosijn and P. Ingwersen, "Dimensions of relevance," *Information Processing and Management*, vol. 36, no. 4, pp. 533–550, 2000.

[161] C. L. Barry, "User-defined relevance criteria: An exploratory study," *Journal of the American Society for Information Science*, vol. 45, no. 3, pp. 149–159, 1994.

[162] T. K. Park, "The Nature of Relevance in Information Retrieval: An Empirical Study," *Library Quarterly*, vol. 63, no. 3, pp. 318–351, 1993.

[163] C. C. Kuhlthau, "Inside the search process: Information seeking from the user's perspective," *Journal of the American Society for Information Science*, vol. 42, no. 5, pp. 361–371, 1991.

[164] P. Wang and D. Soergel, "A cognitive model of document use during a research project. Study I. Document selection," *Journal of the American Society for Information Science*, vol. 49, no. 2, pp. 115–133, 1998.

[165] H. Greisdorf, "Relevance thresholds: A multi-stage predictive model of how users evaluate information," *Information Processing and Management*, vol. 39, no. 3, pp. 403–423, 2003.

[166] R. Tang and P. Solomon, "Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior," *Information Processing and Management*, vol. 34, no. 2-3, pp. 237–256, 1998.

[167] P. Vakkari, "Changes in Search Tactics and Relevance Judgements when Preparing a Research Proposal a Summary of the Findings of a Longitudinal Study," *Information Retrieval*, vol. 4, no. 3-4, pp. 295–310, 2001.

[168] S. G. Hirsh, "Children's relevance criteria and information seeking on electronic resources," *Journal of the American Society for Information Science*, vol. 50, no. 14, pp. 1265–1283, 1999.

[169] B. Wildemuth, L. Freund, and E. G. Toms, "Untangling search task complexity and difficulty in the context of interactive information retrieval studies," *Journal of Documentation*, vol. 70, no. 6, pp. 1118–1140, 2014.

[170] D. Jiang and S. Kalyuga, "Confirmatory Factor Analysis of Cognitive Load Ratings Supports a Two-Factor Model," *The Quantitative Methods for Psychology*, vol. 16, no. 3, pp. 216–225, 2020.

[171] A. Bondarenko, C. Biemann, M. Völske, B. Stein, A. Panchenko, and M. Hagen, "Webis at TREC 2018: Common Core Track," *27th Text REtrieval Conference, TREC 2018 - Proceedings*, pp. 2–4, 2018.

[172] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J. M. Jose, and L. Azzopardi, *Crowdsourcing interactions: Using crowdsourcing for evaluating interactive information retrieval systems*, 2013, vol. 16, no. 2.

[173] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," in *ACM SIGIR Forum*, vol. 42, no. 2, 2008, pp. 9–15.

[174] L. Han, E. Maddalena, A. Checco, C. Sarasua, U. Gadiraju, K. Roitero, and G. Demartini, "Crowd worker strategies in relevance judgment tasks," in *WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 241–249.

[175] P. Thomas, G. Kazai, R. W. White, and N. Craswell, "The Crowd is Made of People Observations from Large-Scale Crowd Labelling," in *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, no. April 2021, 2022, pp. 25–35.

[176] O. Alonso and S. Mizzaro, "Using crowdsourcing for TREC relevance assessment," *Information Processing and Management*, vol. 48, no. 6, pp. 1053–1066, 2012.

[177] E. Maddalena, M. Basaldella, D. De Nart, D. Degl'Innocenti, S. Mizzaro, and G. Demartini, "Crowdsourcing Relevance Assessments: The Unexpected Benefits

of Limiting the Time to Judge," in *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016*, 2016, pp. 129–138.

[178] B. J. Jansen, D. Booth, and B. Smith, "Using the taxonomy of cognitive learning to model online searching," *Information Processing and Management*, vol. 45, no. 6, pp. 643–663, 2009.

[179] S. Kalyuga, P. Chandler, and J. Sweller, "Managing split-attention and redundancy in multimedia instruction," *Applied Cognitive Psychology*, vol. 25, pp. 351–371, 1999.

[180] K. Ouwehand, A. v. d. Kroef, J. Wong, and F. Paas, "Measuring Cognitive Load: Are There More Valid Alternatives to Likert Rating Scales?" *Frontiers in Education*, vol. 6, no. September, pp. 1–13, 2021.

[181] F. Paas, "Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach," *Journal of Educational Psychology*, vol. 84, no. 4, pp. 429–434, 1992.

[182] P. W. Van Gerven, F. Paas, J. J. Van Merriënboer, and H. G. Schmidt, "Memory load and the cognitive pupillary response in aging," *Psychophysiology*, vol. 41, no. 2, pp. 167–174, 2004.

[183] D. Wechsler, "Wechsler adult intelligence scale," *Archives of Clinical Neuropsychology*, 1955.

[184] R. B. Ekstrom, "Kit of factor-referenced cognitive tests," *Educational Testing Service*, 1976.

[185] P. I. Santosa, K. K. Wei, and H. C. Chan, "User involvement and user satisfaction with information-seeking activity," *European Journal of Information Systems*, vol. 14, no. 4, pp. 361–370, 2005.

[186] L. Wen, I. Ruthven, and P. Borlund, "The effects on topic familiarity on online search behaviour and use of relevance criteria," *Lecture Notes in Computer Science*, vol. 3936, no. July, pp. 456–459, 2006.

[187] T. A. Shimoda, "The effects of interesting examples and topic familiarity on text comprehension, attention, and reading speed," *Journal of Experimental Education*, vol. 61, no. 2, pp. 93–103, 1993.

[188] K. Schuessler, V. Fischer, M. Walpuski, and D. Leutner, "The Moderating Role of Interest in the Relationship between Perceived Task Difficulty and Invested Mental Effort," *Education Sciences*, vol. 14, no. 10, 2024.

[189] N. Ferro, N. Fuhr, K. Järvelin, N. Kando, M. Lippold, and J. Zobel, "Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science"," in *SIGIR Forum*, vol. 50, no. 1, 2016, p. 68–82.

[190] A. P. de Vries, G. Kazai, and M. Lalmas, "Tolerance to Irrelevance: A User-effort Oriented Evaluation of Retrieval Systems without Predefined Retrieval Unit," in *RIAO 2004 Conference Proceedings*, 2004, pp. 463–473.

[191] J. Leppink and P. Pérez-Fuster, "Mental Effort, Workload, Time on Task, and Certainty: Beyond Linear Models," *Educational Psychology Review*, vol. 31, no. 2, pp. 421–438, 2019.

[192] K. Seddon, N. C. Skinner, and K. C. Postlethwaite, "Creating a model to examine motivation for sustained engagement in online communities," *Education and Information Technologies*, vol. 13, no. 1, pp. 17–34, 2008.

[193] L. Sinnamon, L. Tamim, S. Dodson, and H. L. O'Brien, "Rethinking Interest in Studies of Interactive Information Retrieval," in *CHIIR 2021 - Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 2021, pp. 39–49.

[194] J. Qi, Z. Han, and P. Hansen, "Information search process model based on costs and benefits: a behavioural economics perspective," *Journal of Documentation*, 2024.

[195] B. Wang and J. Liu, "Understanding users' dynamic perceptions of search gain and cost in sessions: An expectation confirmation model," *Journal of the Association for Information Science and Technology*, vol. 75, no. 9, pp. 937–956, 2024.

[196] G. Wiggers, S. Verberne, W. van Loon, and G. J. Zwenne, "Bibliometric-enhanced legal information retrieval: Combining usage and citations as flavors of impact relevance," *Journal of the Association for Information Science and Technology*, vol. 74, no. 8, pp. 1010–1025, 2023.

[197] N. Roy, A. Câmara, D. Maxwell, and C. Hauff, *Incorporating Widget Positioning in Interaction Models of Search Behaviour*. Association for Computing Machinery, 2021, vol. 1, no. 1.

[198] B. Wang and J. Liu, "Investigating the role of in-situ user expectations in Web search," *Information Processing and Management*, vol. 60, no. 3, 2023.

[199] J. Gwizdka, ""Overloading" Cognitive (Work)Load: What are We Really Measuring?"," in *Proceedings NeuroIS Retreat*, 2023, pp. 1–272.

[200] E. Babaei, T. Dingler, B. Tag, and E. Velloso, "Should we use the NASA-TLX in HCI? A review of theoretical and methodological issues around Mental Workload Measurement." *International Journal of Human Computer Studies*, vol. 201, 2025.