A Systematic Review of Cost, Effort, and Load Research in Information Search and Retrieval, 1972-2020

MOLLY MCGREGOR, Department of Computer and Information Sciences, University of Strathclyde, UK
LEIF AZZOPARDI, Department of Computer and Information Sciences, University of Strathclyde, UK
MARTIN HALVEY, Department of Computer and Information Sciences, University of Strathclyde, UK

During the *Information Search and Retrieval* (ISR) process, user-system interactions such as submitting queries, examining results, and engaging with information, impose some degree of demand on the user's resources. Within ISR, these demands are well recognised, and numerous studies have demonstrated that the Cost, Effort, and Load (CEL) experienced during the search process are affected by a variety of factors. Despite this recognition, there is no universally accepted definition of the constructs of CEL within the field of ISR. Ultimately this has led to problems with how these constructs have been interpreted and subsequently measured. This systematic review contributes a synthesis of literature, summarising key findings relating to how researchers have been defining and measuring CEL within ISR over the past 50 years. After manually screening 1,109 articles, we detail and analyse 91 articles which examine CEL within ISR. The discussion focuses on comparing the similarities and differences between CEL definitions and measures before identifying the limitations of the current state of the nomenclature. Opportunities for future research are also identified. Going forward, we propose a CEL taxonomy that integrates the relationships between CEL and their related constructs, which will help focus and disambiguate future research in this important area.

CCS Concepts: • **Information systems → Users and interactive retrievals**.

Additional Key Words and Phrases: Cost, Effort, Cognitive Load, Workload

ACM Reference Format:
Molly McGregor, Leif Azzopardi, and Martin Halvey. 2023. A Systematic Review of Cost, Effort, and Load Research in Information Search and Retrieval, 1972-2020. 1, 1 (January 2023), 40 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

During the *Information Search and Retrieval* (ISR) process, the user submits queries, examines results, and engages with information to make a decision [8]. All of these user-system interactions impose some degree of demand on the user's internal (i.e., memory, attention) and external resources (i.e., time, money). In an ideal situation, the user has an abundant stream of resources to overcome these demands and find the perfect information to serve their information need. However, this is rarely the case. Rather, these internal and external resources are limited in capacity and the user must therefore make decisions during the ISR process based on these limitations. The demands imposed during

Authors' addresses: Molly McGregor, molly.mcgregor@strath.ac.uk, Department of Computer and Information Sciences, University of Strathclyde, 16 Richmond Street, Glasgow, UK, G1 1XQ; Leif Azzopardi, Department of Computer and Information Sciences, University of Strathclyde, 16 Richmond Street, Glasgow, UK, leif.azzopardi@strath.ac.uk; Martin Halvey, Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK.

the search process are well recognised, and numerous studies have demonstrated that the cost, effort, and load (CEL) experienced during the search process is affected by not only the search task [27, 91, 120], but also the system [21, 23, 37] and the individual characteristics [24, 57] of the user. Despite this recognition, our recent perspectives paper [97] highlighted two prominent challenges facing the ISR community in regard to CEL research. The first challenge relates to the general lack of consensus regarding cost, effort, and load definition, which has ultimately left the field with no universal definitions to describe these constructs. The second challenge relates to the scarcity of gold standard measures to examine these constructs. As conceptualisation precedes operationalisation, the latter challenge cannot be overcome without resolving the former. With these challenges at the forefront, a key aim of this systematic review is to build on our previous work, by providing a richer and more comprehensive analysis of the existing CEL literature within ISR. To this end, the scope of this review extends the literature analysis beyond studies which explicitly measure CEL constructs by also including studies which less overtly examine these constructs within an ISR context. In gaining a fuller understanding of CEL definition and measurement in ISR, the two principle goals of this review are to (1) work towards establishing clearer definitions of these constructs, which in turn, should (2) inspire critical evaluation of existing measures and their use within future CEL research. To our knowledge, this is the first systematic review which attempts to characterise and analyse how cost, effort, and load have been defined and measured within the field of ISR. This in turn raises a number of issues central to the use of CEL definitions and the effectiveness of measures in CEL assessment, which are key to understanding and supporting ISR.

2 BACKGROUND

2.1 Scope of Cost, Effort, and Load Research in ISR

For over 50 years, measures of cost and effort have been included within *Information Retrieval (IR)* evaluation frameworks, highlighting their importance among other evaluation measures such as precision and recall. Cost is frequently considered within ISR as the amount of time spent, or the number of interactions performed. This may indicate that users place great value on the time they spend searching, or that researchers consider time and the number of interactions performed as a relatively easy to collect measure of cost. As early as 1968, Miller [100] proposed the 2-second rule – which states that an interactive system should take no longer than two-seconds to respond in order to keep the users focus and attention on the current task. Similar observations were made in later studies, which found that delays in network and download speeds can have a detrimental effect on users' perception of usefulness and their interactions with web pages [32, 44, 103]. A study examining the effects of time delays on the Google search engine found that even a relatively small delay of 400ms in returning search results led to a reduction in the number of searches conducted by 0.59% over the course of 6-weeks [16]. Additional studies revealed similar findings, highlighting that as the cost of querying increases, the number of queries issued by the user will decrease [8, 96].

It seems that in high-cost situations, the user will attempt to preserve the effort they exert (i.e., by issuing less queries, viewing less pages, spending less time on a webpage) and are willing to accept fewer sources of information (i.e., webpages) to avoid exerting more effort. As more effort (i.e., cognitive actions such as reading a document; formulating a search query; examining search results; and also physical actions such as typing a query) is required, these search decisions are proposed to be based on satisficing - where the individual stops searching for information when the already acquired information is sufficient to satisfy their information need [94]. This decision-making process, often referred to as Zipf's "principle of least effort" Law [145], involves the assessment of the effort required to continue searching and find more and perhaps better information with respect to the expected utility of that information [94].

Studies have shown that as demands on available resources increase during the ISR process, users perform less actions over time and are likely to stop completely if these demands become too high [56]. For example, Azzopardi et al. [8] found that users searching on the structured interface (considered high cost) issued significantly less queries and examined significantly more documents per query than those using the lower cost interfaces. Maxwell and Azzopardi [96] observed a similar finding, with their results indicating that as the relative cost of querying increased, users issued fewer queries and examined more documents per query. As the information search task becomes more complex, users perceive higher levels of effort [91] and workload [120], and effort measures have been found to correlate with all information task characteristics associated with task difficulty (i.e., higher number of steps to achieve the task goal; intellectual task product; amorphous task goal). In tasks with the highest difficulty level, users are found to perform fewer actions, clicks, and bookmarks per query Capra et al. [18].

Similarly, as the users primary search goal is often to fulfil their information need, Yilmaz et al. [141] found that users will not spend time determining if a document is relevant, if they do not find the information they are looking for quickly or if they believe the relevant information is difficult to consume or process – in these cases, the user will often give up and move on to another document. Thus, even if a document actually is relevant to a query or the users information need, the utility of the document to the user will decrease if they have to exert more effort to find and understand the relevant parts of the document. Gwizdka [52] identified that average cognitive load varies by task stages and found that cognitive load was at its highest during query formulation and tagging of relevant documents as compared to examining search results and viewing individual documents. These studies suggest that in high cost search scenarios, users will most likely exert their effort in task stages of lower demand such as examining results or individual documents, but they will actively avoid task stages of higher demand such as issuing queries or making relevance assessments of documents (i.e., bookmarking documents).

Considering the extent to which CEL can influence user search decisions and behaviours, it is not surprising that much of the research in this area has focused on developing search systems and interfaces which endeavour to reduce the demands imposed on the user during the search process. For example, users have been found to exert less effort when using a system which leverages post-retrieval document clustering techniques compared to a system with more unstructured techniques [21]. When performing a search task on a structured interface, users reported significantly less workload than those using a standard (traditional) interface [37]. These findings were also consistent in relation to user effort, where users expend less effort when search results were presented in a visual representation (400 results per page) vs. a traditional vertical list (10 results per page)[47]. However, while the influence of CEL on user behaviour and decision-making during the search process is apparent, it is also acknowledged that it may not affect all users in the same way. For example, users with higher cognitive abilities are found to exert more effort as interface and task conditions increase in complexity [57]. Similarly, users with high working memory exert more effort during the search process than their low working memory counterparts [25], and users with high perceptual speed ability experience lower task demand than those with low perceptual speed ability [14].

While these examples represent only a small number of the studies examining CEL within ISR, it is clear from our discussion that cost, effort, and load play a key role in the search process - clearly highlighting the importance of incorporating these constructs into system evaluation frameworks. However, the influence and impact of cost, effort, and load during the search process cannot be fully understood and recognised until the field reaches a consensus in relation to their definition, which in turn, will inform more accurate measurement of these constructs.

2.2 Defining Cost, Effort, and Load

This section aims to provide an overview of CEL constructs and theory, specifically, we present accepted CEL definitions from Psychology and related fields in an effort to ground our discussions about CEL in ISR.

Research methods from Psychology suggest that in order to establish an appropriate method, it is important that the measurable construct is sufficiently defined relative to its nominal and operational meaning [113]. Nominal definitions, also known as *conceptualisation*, describe the meaning of the construct, whereas operational definitions, also know as *operationalisation*, explain precisely how the construct and its elements will be measured [73]. As conceptualisation precedes operationalisation, significant problems can arise when ambiguous and vague definitions are used. For example, when a specific term, i.e. effort, is defined in multiple ways then the operational properties which emerge from these definitions will also vary. As these operational properties dictate which elements of the construct are to be measured, then it is likely that a variety of different variables will emerge from these properties and later be used by researchers to measure what they believe to be the "same" construct. In reality, the lack of precision in the conceptual definitions provided at the outset will likely lead to conceptual overlap, ambiguous measures, and a loss of causality.

For over 50 years, CEL constructs have been discussed and explored extensively across disciplines such as Psychology [110, 123, 137], Ergonomics [20, 142], and Human Factors [31, 61]. Yet, universal definitions of these constructs are still yet to transpire. The rest of this section will briefly present the most commonly accepted definitions of CEL proposed by disciplines out-with ISR.

2.2.1 Cost. The term "cost" can be considered an abstract construct, assigned a variety of different interpretations depending on the field of research. In Cognitive Psychology and Neuroscience, cost is mostly referred to in relation to cognitive or mental costs [39, 135]. Humans are considered to value the effort that they exert - thus effectively treating the expenditure of effort as costly [136]. Human behaviour often reflects a constant trade-off between effort and reward [109] - however, the limited capacity of our cognitive resources constrains these trade-offs. Subsequently, the level of cognitive resources allocated to a specific task at a given moment, must be selected tactically on the basis of a cost-benefit analysis i.e., "should I spend or conserve my resources to achieve this goal?" [109]. Cost is often operationalised in these scenarios as "energy" or "fatigue" [12] - where individuals are hypothesised to only exert energy on tasks when these energetic costs are comparably low and reward benefits are comparably high. Cost-benefit analyses underlie much of the discussion surrounding the construct of cost. The Behavioural Economic approach assigns monetary value to quantify the costs involved in low vs. high effort tasks [135]. Temporal costs have also been considered in the trade-off process - where time is perceived as expensive, effort expenditure should be directed towards faster processes (at the expense of accuracy) to achieve the task goal [109]. While different fields of research discuss cost in different ways i.e. time, money etc., there appears to be a general consensus among perspectives that the exertion of effort involves the allocation of resources - and that these resources can be operationalised as "costs".

2.2.2 Effort. Effort, also referred to as mental, cognitive, or physical effort - has been branded one of the most intuitive and familiar aspects of human cognition [118]. Yet despite this ability to introspect on effort, examining the construct scientifically has not been an easy endeavour to achieve [118]. While universal definitions of effort are yet to emerge, there is convergence among the different definitions provided. Early definitions from Psychology describe effort as a volitional and intentional process which reflects what an individual is actively participating in, rather than what is passively happening to them [38]. In Cognitive Psychology, effort refers to the level and intensification of either mental or physical labour in the service of meeting the demands of a task or goal [64]. This implies that effort constitutes the

summation of mental labour over time in order to achieve a goal. Similarly, Kirschner [78] defines mental effort as the *"amount of cognitive capacity or resources that is actually allocated to accommodate the task demands"*. Definitions from Evolutionary Psychology, also align effort with the notion of "work"- proposing that individuals choose whether to exert effort based on the incentives available - akin to a worker who makes the decision to work depending on incentives such as salary.

2.2.3 Cognitive Load and Mental Workload. While cognitive load and workload evolved independently from within different disciplines, both constructs are theoretically underpinned by the same core assumptions [108]. *Cognitive Load Theory* (CLT) [122] and *Multiple Resource Theory* (MRT) [137] which emerge from the fields of Educational Psychology, and Ergonomics and Human Factors, respectively, can be considered the leading theories to describe both cognitive load and workload constructs. Both theories are similar in that they are closely related by their assumption of limited mental capacity and competing task demands [122, 137]. Before describing the theories in more detail, it is important to first address what is meant by the term "demand". Demand refers to the properties of the task that will regulate how much physical or mental exertion will be needed [64]. Sweller [121] proposes that contextual demands arise from the intrinsic qualities of the context (e.g. task difficulty, information presentation) which require resources.

Since its emergence in the 1980s, *Cognitive Load Theory* (CLT) has been primarily applied within the field of educational instruction and learning [126]. In CLT, cognitive load is defined as the total amount of mental activity imposed on the working memory at any given moment [123]. This definition reflects the origins of CLT and its emergence from the working memory model which emphasises the limited capacity of working memory and the abundant capacity of long term memory in the human brain [99]. The amount of cognitive load experienced by an individual is influenced by the number of elements simultaneously interacting within working memory. As working memory capacity is limited, there is a finite amount of information that working memory can handle at any one time. Therefore, if too much information occurs at once, the working memory becomes overloaded and the individual will be less likely to process information [122]. Perhaps the most defining feature of CLT is the discrimination between three different types of cognitive load; intrinsic (inherent characteristics of the task, i.e. difficulty), extraneous (load imposed by the context in which in the task is being performed), and germane (load imposed by the construction of schemas) [30]. These three types of cognitive load are proposed to be additive [111].

Mental workload is perhaps one of the most popular constructs examined in Ergonomics and Human Factors research, however researchers in the field are still yet to reach a universal consensus regarding its definition [142]. Although elements of CLT have been used to conceptualise the term "workload", definitions found in Psychology tend to align workload to Multiple Resource Theory, particularly in relation to the processes of task switching and allocation of attention [20]. MRT asserts that the human brain has a fixed quantity of mental resources of various types [125]. These resources can be characterised as a shared pool of energy that can be drawn on for a variety of simultaneous mental operations, including across different tasks, modalities, and processing [137]. The theory interprets performance decrements as the depletion of these resource pools which can occur when the performance of two or more tasks require a single resource [125]. Mental workload is inferred as the allocation of available resources to meet the demands of a task and the cognitive experience of the individual directly activated by those task demands. [20, 125]. Van Acker et al. [125] describe mental workload as conceptually very similar to cognitive load - with their underpinning theory as the only discerning feature. Nevertheless, we can clearly observe that MRT and CLT are closely linked, both conceptualised by the notion of task demands and resource consumption. As MRT proposes multiple resources available for allocation,

then perhaps it can be considered as a generalisation of CLT, which proposes the availability of only one cognitive resource.

2.2.4 How Are Cost, Effort, and Load Related? Cognitive load is considered a multi-dimensional construct encompassing aspects of both load and effort [85]. Cognitive load is imposed by the demands of the task parameters on our mental resources at a given point in time, and experienced by the individual. Effort is a volitional response to the load and refers to the total amount of cognitive resources allocated to attend to the task demands over time in order to achieve some kind of end goal [30]. Thus the relationship between effort and load appear relatively straightforward - effort is exerted by the individual, whereas cognitive load is experienced by the individual [30]. In terms of their relationship with cost, research in the domains of Cognitive Psychology, Neuroscience, and Economics, have widely considered effort, whether it be physical or mental, as costly [64] - including affective (i.e. fatigue [12]); temporal (i.e. time spent [109]); and economic costs (i.e. monetary [135]).

2.3 Summary

The discussion so far has highlighted that fields outwith ISR have faced struggles in providing explicit definitions of cost, effort, and load. Yet despite this general lack of consensus, there is still a clear delineation between these constructs in relation to their unique qualities. From this, we can also observe the extent to which these constructs relate. In this systematic review, we use these external representations of cost, effort, and load to ground our analysis of their treatment within the field of ISR. Not only do we utilise these interpretations as a benchmark to which we can compare definitions, but they also serve as a basis to examine how the existing measures used within ISR have the ability to accurately represent and reflect these complex constructs.

3 METHOD

The purpose of this systematic review is to firstly gather, present, and synthesise existing ISR research which discuss the measurement of cost, effort, and load, and secondly, to highlight the current challenges facing CEL definition and measurement within ISR. To this end, we hope to achieve our two main goals of: (1) establishing clearer definitions of these constructs, and (2) to provide a critical evaluation of existing measures and their future use within ISR research. In order to achieve this, a systematic review was chosen as the method of choice for this article as it has the potential to identify all of the relevant scholarly research on a particular topic, in an unbiased and reproducible manner. Conducting the systematic review involved five key steps: (*i*) formulate research questions; (*ii*) identify sources from which the articles would be selected; (*iii*) develop a search strategy and search terms/keywords; (*iv*) establish inclusion/exclusion criteria; (*v*) develop categories for coding and analysis. Each of these steps are described in more detail in the following sections.

3.1 Step 1: Research Questions

The following research questions aim to address our key goals outlined in Section 3. These questions were developed with the purpose of obtaining a broad and detailed understanding of how CEL is currently being defined, measured, and the relationships between these, across the field of ISR. We aim to use these findings to examine areas of convergence among researchers regarding definitions, which in turn, can help ground the development of clearer CEL definitions going forward. These questions will also allow us to critically examine the extent to which the measures used can

accurately reflect the conceptual properties of cost, effort, and load indicated by the definitions provided. To this end, this systematic review seeks 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 Cost, Effort, and Load 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 Cost, Effort, and Load, 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?

Table 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

Table 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

3.2 Step 2: Sources

The second step involved the identification of sources from which the articles would be selected. This involved identifying key sources where articles measuring CEL in the context of ISR were most likely to be published. In their systematic review of Interactive Information Retrieval (IIR) evaluations studies, Kelly and Sugimoto [76] provide a comprehensive list of 31 sources (17 journals & 14 conference publications), reviewed and approved by four IIR experts. This list was considered a reliable base from which to select relevant sources for the present review. After discussion with two ISR experts, 17 sources from the list outlined by Kelly [73] were selected, and 9 additional sources considered appropriate for the purpose of this review were also included. In total, 26 sources (See Table 1) were selected (9 journals; 16 conferences; 1 workshop). Manual searches were also used as a supplemental approach to the database searches in order to identify additional studies for review that may have been overlooked in the initial database search. The Cochrane Handbook recommends manual searching as a useful addition to electronic database search as some articles may not include the relevant search terms in their title or abstract [128]. In parallel to the primary database search, there is no prescribed or standard practice for conducting the manual search [128]. The Cochrane Handbook suggests two common strategies of manual search; (1) manually sifting through entire contents of journal issues or conference proceedings to identify all relevant papers; (2) scanning reference lists of identified relevant articles for additional relevant articles [105]. The latter strategy is also referred to as 'backward chaining' – i.e., moving backward through a chain of reference lists [48]. Alternatively, 'forward chaining' involves beginning with an article citation and then examining an index of authors who have cited this article i.e., follow the chain in a forward's direction [48]. Both methods of manual search were used to identify potentially relevant articles. A total of 26 articles from 16 different sources were identified through manual search (please refer to Table 2 for an overview of specific journals/conferences).

3.3 Step 3: Search Strategy

The next stage involved keyword searches within the selected literature databases to identify papers. There was slight variation in search terms depending on the search database used, these are described below. The search term ***(effort***

OR cost* OR "mental workload" OR "cognitive load" OR workload) AND (search* OR "information retrieval" OR "information seeking") was used for all of the search databases except from the following three: 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). The IP&M and IJCCI database did not allow search terms with truncation symbols, therefore 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")***. All articles within the EuroHCIR database had to be manually searched as the Boolean search function was not available.

The search criteria used also depended on the options provided in each database and therefore differed between sources. Where possible, the search was refined to 'title' and 'abstract' search. However, there were instances where this was not possible, and to avoid missing relevant articles the search was limited to either the 'title', 'abstract', or in some cases a 'full-text' search (please see Table 1). The database search retrieved a total of 1,083 articles.

3.4 Stage 4: Inclusion/Exclusion Criteria

The aim of this step was to establish and evaluate inclusion and exclusion criteria (see Table 3) which would be utilised to systematically select and reject articles for review. This took place prior to the literature search. The development process was initiated with a collection of criteria that aligned with the target article type: the measurement of CEL in ISR. While the initial list of criteria was considered preliminary and therefore open to refinement as the review progressed, no additional adaptations to the criteria were made. Perhaps the most important criteria relate to the articles research topic and scope. Firstly, empirical studies (i.e. uses experimental methods) should measure at least once CEL construct in an ISR context. Non-empirical articles (i.e. a review article), should propose or describe at least one CEL construct in the context of ISR. Sources were limited to journals, conference proceedings and workshops. To ensure the retrieval of higher quality articles, it is necessary that the articles were scholarly publications and had been subject to peer-review. Both full length research articles and short papers were also included and should be written in English. While short papers are of a smaller scale, they were included as they are still expected to provide a detailed description of their methods which is of key interest to the aim of this review. To ensure the full scope of relevant articles were identified, a specific time frame was not established. Rather, the search was intended to yield articles which used these constructs from the start of the literature.

Table 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

The title, abstract, and results section of the 1,083 articles retrieved through the database search were screened using the inclusion/exclusion criteria. Please refer to Figure 1 for a complete overview of the full literature review process.

While it is common practice in a systematic review to screen only the title and abstract, the focus of this paper on methods deemed it appropriate to include the results section in the screening process to ensure no relevant articles were

overlooked. Following the initial screening, 67 articles were found to satisfy the inclusion criteria. Each article was then read in full to further assess eligibility. Two articles were excluded at this point as they did not measure CEL within an ISR context. This left 65 articles to be brought forward for analysis. Using these articles as a source, an additional 26 articles were retrieved through manual search; 'backward chaining' (N=16), and 'forward chaining' (N=10). Table 2, shows a list of the journals, conferences/workshops where the identified articles were published. In total, 91 articles were included in the final analysis.

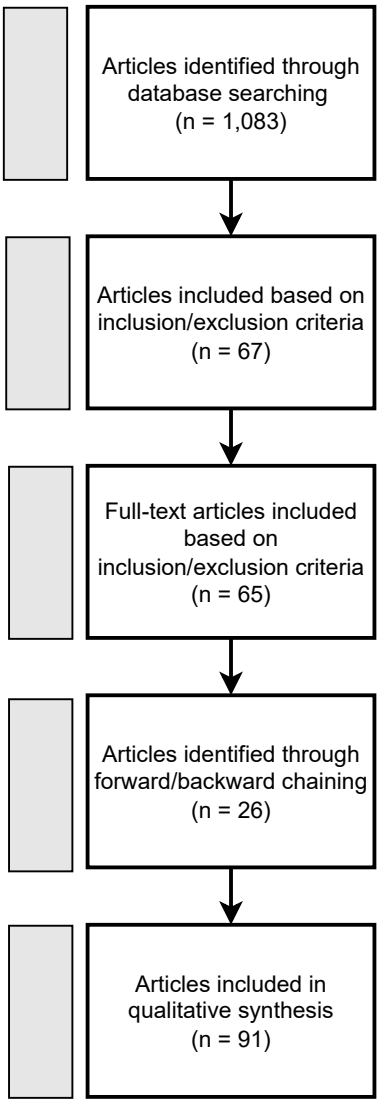

Fig. 1. Systematic Literature Review Process

3.5 Stage 5: Categories for Analysis

The categories for coding and analysis were developed according to the key research questions. The following elements which align with the scope of the research questions were examined and coded as appropriate:

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

- Studies that include/do not include a definition of cost, effort, and load.
- Similarities/differences between the conceptual categories associated with cost, effort, and load.
- Different terminology applied to cost, effort, and load.
- 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 cost, effort, and load and the methods used to measure these constructs.
- Studies which use/propose the same methods to measure different constructs.

4 RESULTS

4.1 Overview of Literature

Characteristics of Articles: The first articles examining CEL within ISR emerged within the period 1972-1982. Cost was one of the first CEL constructs to be examined, however the construct was not revisited until the period 2005-2015. Conversely, the interest in effort measurement within ISR has been quite consistent over the last four decades (except 1983-1993). Cognitive load and workload are constructs of more recent interest within ISR, with studies most prevalent during the last two decades. Please refer to Figure 2 for an overview.

CEL Construct Examined: Across the reviewed articles, effort was the most frequently examined construct ($N = 54$), with 49 of these articles empirically measuring the construct. Workload was the second most widely used construct ($N = 22$), with 21 articles measuring the construct empirically. Cognitive load was examined in 9 studies, 8 of which were empirical. Finally, 16 articles examined cost, 11 of which were empirical studies.

Study Type: The majority of reviewed articles ($N = 82$) employed an empirical user study. Lab-based and controlled user studies were by far the most widely used ($N = 69$), followed by online crowd-sourced user studies ($N = 7$); and naturalistic user studies ($N = 6$) conducted in a more realistic, less controlled environment.

The remaining articles ($N = 9$) examined CEL constructs non-empirically in a variety of different ways. Two studies proposed novel methods for measuring cognitive load and cognitive workload for use in a future study. Another two studies proposed formal models of user-sided costs in the context of conducting a search and interacting with a system. Others ($N = 2$) used simulated search sessions to support the experimental evaluation of retrieval systems in relation to

user cost and effort. Finally, three studies employed an analysis of existing search logs to predict user cost and effort during the search process.

Task: Across the majority ($N = 63$) of empirical studies, the task and/or topic were explicitly assigned to the user. These studies varied in relation to the specificity of the task description. The majority of these studies ($N = 22$) identified the task as an interactive web search task. Simulated work tasks ($N = 8$); collaborative search tasks ($N = 5$); and relevance judgement/annotation tasks ($N = 4$) were also frequently used to differentiate tasks. Other studies identified tasks according to the process i.e., browsing ($N = 3$); searching ($N = 8$); exploratory ($N = 11$), and the remaining studies characterised the task in relation to the end-product (i.e., known-item search ($N = 2$); essay/coursework search ($N = 3$); fact-finding ($N = 11$); product search ($N = 4$); comparative search ($N = 5$).

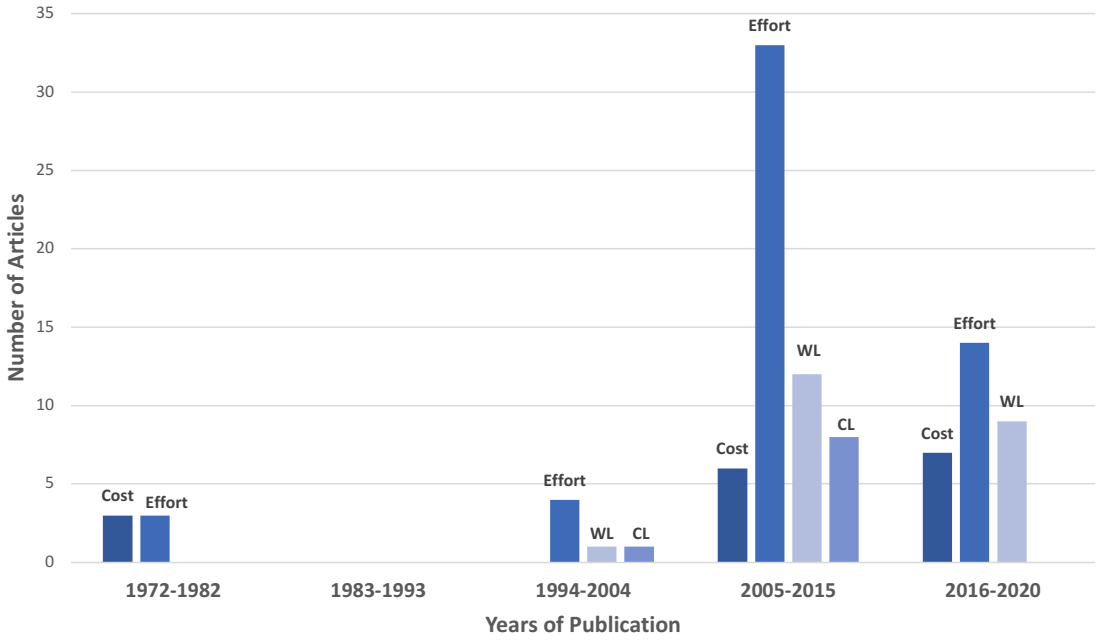

Fig. 2. Number of articles by year of publication.
Note: CL = cognitive load; WL = workload

Task Domain: Task domains related to media and leisure were most popular ($N = 29$) and included; news ($N = 8$); entertainment ($N = 4$); e-commerce ($N = 7$); general health ($N = 3$); travel ($N = 5$); weather ($N = 1$); and finding a job ($N = 1$). Eleven studies examined professional domains such as; journalism ($N = 7$); healthcare ($N = 4$); and research ($N = 2$). Nine studies examined different academic domains including; Science and Technology ($N = 3$); Politics ($N = 2$); History ($N = 1$); Education ($N = 2$); Genomics ($N = 1$). Finally, ten studies examined CEL in the context of digital libraries. The remaining studies ($N = 36$) reviewed did not specify a specific search domain, and rather just described a web-based search task with no further contextual information.

Corpus: The collection used was described in 73 of the 91 studies. Of these, the "web" was the most frequently used ($N = 39$), this includes studies which used the "open web" and those that used "closed web" – described by Kelly and

Sugimoto [76] as static web pages that have been manipulated in some way (i.e., features removed/added) before the study. TREC collections were used in 18 studies, and library collections were used in 5 studies. Finally, other collections such as CLEF ($N = 1$); and medical information databases ($N = 2$) were also used. The remaining studies ($N = 5$) represented unique or individual collections.

Subjects: The majority of subject groups ($N = 56$) were recruited from academia. Undergraduate and graduate students represent a considerable number of those recruited from this sector ($N = 45$), and came from a wide variety of disciplines (i.e., Humanities ($N = 5$), Social Science ($N = 4$), Library and Information Science ($N = 4$), Psychology ($N = 4$); Information Technology ($N = 4$), and Biology ($N = 4$)). Aside from students, university staff and members of faculty were also included as participants in some studies ($N = 6$). Finally, nine groups of participants from the academic sector were recruited but their position was unspecified. Twelve studies included subjects recruited from a variety of industry sectors. Only two studies shared the same type of participant group (i.e., GPs/Physicians). The other participant groups were diverse and were observed in only one study each - these included: web designers, IT company employees, engineers, sales reps, product managers, online search analysts, forest service personnel, and librarians. Finally, the remaining studies recruited subjects through crowd-sourcing ($N = 7$) or a specified population such as the general public ($N = 3$) and public library visitors ($N = 4$).

4.2 Definitions of CEL, and Related Constructs

Definitions of Cost, Effort, and Load: Of the 91 articles reviewed, 39 definitions of CEL were extracted from 38 articles; cost ($N = 8$); effort ($N = 20$); cognitive load ($N = 6$); and workload ($N = 5$). Following guidance from [82], the individual elements of each CEL definition were then separated, categorised and quantified. Based on this analysis, the definitions were then aligned with one of the five conceptual categories identified in our previous work [97]. Refer to Table 4 for an overview of each construct and their associated conceptual categories. Each of the five categories are based on definitions which can be characterised by the following descriptions:

Cost: The term "cost" was used in all sixteen articles which examined the construct. Two articles measuring cost also used the term effort interchangeably [91, 139]. For the eight articles which defined cost, two main conceptual categories were identified:

- *Time Orientated:* Cost was most frequently characterised in relation to the time spent during the user-system interaction ($N = 7$) [8–10, 22, 29, 91, 96]. For example, *"cost is often measured by the time of a series of actions, such as formulating queries, examining snippets, clicks on results etc"* [91]; and *"cost is often considered as the amount of time spent"* [9].
- *Interaction Orientated/Count Based:* Cost was also defined in relation to the interactions/number of actions occurring between the user and the system ($N = 4$)[9, 22, 96, 144]. For example, *"consider the cost of information search results from query generation, documentation examination, search engine result pages and task description examinations"* [144]. Note that three of the studies [9, 22, 96] mentioned provide definitions which align with both conceptual categories.

Conceptual Category	Construct & Number of Articles	Citation
Time-Orientated	Cost (N = 7)	[8–10, 22, 29, 91, 96]
	Effort (N = 2)	[83, 132]
Interaction Orientated/Count-Based	Cost (N = 4)	[9, 22, 96, 144]
	Effort (N = 7)	[27, 55, 56, 62, 66, 119]
Cumulative/Total Work	Effort (N = 8)	[21, 34, 36, 40, 59, 77, 104, 132]
	Cognitive Load (N = 2)	[23, 116]
	Workload (N = 4)	[5, 34, 70, 89]
Meta-Cognition/ Conscious Awareness	Effort (N = 3)	[77, 91, 115]
	Workload (N = 1)	[120]
Capacity Based/ Resource Bound	Effort (N = 1)	[53]
	Cognitive Load (N = 3)	[52, 133, 138]

Table 4. CEL Construct and Conceptual Categories

Effort: While the term "effort" was the most frequently used (N = 40), five studies used the term "cognitive effort", and four used the term "mental effort". Two articles measuring effort [27, 50] also used the term cognitive load interchangeably. Of the 49 articles examining effort, 20 provided a definition of the construct.

These definitions aligned with all five conceptual categories.

- *Cumulative/Total Work* The majority of studies (N = 8) [21, 34, 36, 40, 59, 77, 104, 132] characterised effort in terms of the cumulative or total amount of physical/mental work that the individual applies towards an outcome. For example, *"the total work done to achieve a particular goal"* [11], and *"how much work must an assessor exert"* [59].
- *Interaction Orientated/Count Based:* Similar to the definitions of cost, the second category (N = 7) [27, 55, 56, 62, 66, 119] to emerge characterises effort as the interaction or number of actions which occur between the system and the user. For example, *"we assume that users perform actions to make progress on a search task; every action costs effort"* [62] and *"counts of actions that require a cognitive assessment (e.g., evaluate a SERP) or production (e.g., enter a query) from a searcher"* [55].
- *Meta-cognition/Conscious Awareness:* The third category (N = 3) [77, 91, 115] to emerge refers to 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"* [115]. Note that the definition provided by Kim and Rieh [77] aligned with two conceptual categories.
- *Time Orientated:* Again, similar to the characterisation of cost, two studies [83, 132] 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"* [83].

- *Capacity Based/Resource Bound:* Finally, one study [53] refers to 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"* [53].

A number of the effort definitions provided use other CEL terms when describing the construct. For example,*"search effort is the cost of acquiring information from the search engine"* [66].

Cognitive Load: The term "cognitive load" was the most frequently used ($N = 7$), followed by "mental load" ($N = 1$). Of the nine articles which examined cognitive load, six definitions were provided. Two main categories emerged from the definitions.

- *Capacity Based/Resource Bound:* Also used to describe effort, the majority of studies ($N = 3$) [52, 133, 138] characterised cognitive load in relation to the notion of limited mental capacity and resources. For example, *"cognitive load is closely related to the notion of limited mental resources"* [52] and *"people's cognitive capacities are so limited that they can process only limited information chunks concurrently"* [133].
- *Cumulative/Total Work:* The definition used in two articles [23, 116], aligns with the "work" related elements also identified in the characterisation of effort . For example, *"cognitive load is usually evaluated according to the quantity of information to be memorised and the amount of processes involved to perform the task"* [23].

Workload: The term "workload" was the most frequently used ($N = 12$), followed by "cognitive workload" ($N = 5$), "mental workload" ($N = 4$). Of the 22 articles which examined workload, five definitions were provided. These definitions correspond to the following two categories:

- *Cumulative/Total Work:* Similar to how effort was described, the first category (N = 4) [5, 34, 70, 89] to emerge for definitions of workload relates to the cumulative or total amount of physical/mental work. For example, *"can be intuitively defined as the amount of cognitive work necessary for a person to complete a task over time"*
- *Meta-cognition/Conscious Awareness:* Finally, workload was characterised in one article [120] by the category which also emerged from effort definitions. For example, *"mental workload represents the subjective experience of a decision maker"* [120].

Terms such as effort and cost were used in definitions of workload. For example, *"mental workload (or mental effort) is the term used to describe the mental cost of accomplishing task demands that is the individual reaction to the objective requirements of the task"* [34]. Similarly, *"the mental effort involved in performing any given task"* [5], and *"mental workload refers to the amount of perceived effort induced by a particular task"* [70].

Demand: Eleven articles measuring effort [53, 55, 56, 115]; cognitive load [52, 116, 133]; and workload [5, 70, 89, 120] provided a definition of the term "demand". Similar to the analysis of the CEL constructs, all identified definitions were organised into categories. This analysis revealed two categories of how "demand" is discussed in the literature:

- Task demand ($N = 7$)
- System demand ($N = 5$).

Task demand relates to the demands which arise from the search task itself. Task characteristics such as difficulty [89]; time pressure [89]; and complexity [133] were all described in relation to task demand. Simultaneous task requirements were also considered to impose higher demand on the user [5].

System demand is generally described as the the demand imposed on the user by the information retrieval system. Elements of the system such as the interface and information displays [52]; modality (i.e., visual, auditory) [89]; and interruptions [89] were all considered to affect the demands imposed on the user.

Resources: Fifteen articles measuring cost [91]; effort [34, 53, 55, 56, 77, 115]; cognitive load [14, 23, 52, 116, 133]; and workload [9, 34, 89, 120] provided a definition of the term "resources". Across all of the CEL constructs, "resource" definitions followed a similar theme, referring to the users:

- Cognitive resources ($N = 11$)
- Limited capacity of resources ($N = 8$)

Cognitive resources are the *internal resources* that the user has available to them in order to process stimuli. The majority of definitions refer to the resource - working memory [9, 14, 52, 55, 120]. However, other cognitive resources such as perception [9, 89], attention [9, 77, 133], and motor control [9] are also mentioned.

The second theme to emerge from the resource definitions relates to the limited capacity of these resources. This refers to the notion that there is only a finite amount of resource that can be allocated to a task/system demand at any one time. Some studies [9, 52, 55, 56] describe this as a "constraint" or "limitation" of human mental capacity.

Theory: Overall, seven articles referenced an established theory to ground their research. Two dominant theories were discussed:

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

4.3 CEL Measurement in ISR

Objective Methods: From the 91 articles reviewed, 59 articles used or proposed objective methods to measure cost, effort, and load. For the purpose of this review, we consider objective methods as an impartial measurement, with a quantifiable outcome. For empirical studies, search interaction logs were the most frequently used method, with 48 articles using at least one search interaction measure. In nine of these studies, search interaction logs were used alongside another objective measure, such as dual task [50, 77, 115], and eye tracking [22, 27, 47, 55, 133, 144]. Compared to search interaction logs, other objective methods were used less frequently, with only twelve articles using eye tracking measures; eight using the dual-task method; and one article using a combination of multiple physiological measures such as; electroencephalogram (EEG), Temperature, Electrocardiogram (ECG), electro-dermal activity (EDA).

For studies which were not empirical, nine studies proposed objective measures of cost, effort, and load. Again, search interaction logs were most common, with seven articles [7, 10, 29, 41, 66, 83, 141] proposing the use of search interaction logs as a method to measure cost and effort. The remaining two studies proposed the use of two physiological methods: functional near-infrared spectroscopy (fNIRS)[92] and electroencephalogram (EEG) [138] as a measure of users cognitive load. Table 5 provides an overview of the objective methods which were used and proposed, including a breakdown of the constructs measured.

Dependent Variables Used for Objective Methods. Each of the four objective methods used across the studies had dependent variables associated with them. There were 23 dependent variables derived from the search interaction logs. These variables can be assigned to three measurement categories: *total interaction counts; rates of interaction;* and *time-based.* Only cost and effort were measured via total interaction counts and rates of interaction, whereas time based measures were used to measure all three CEL constructs. Five dependent variables were derived from eye-tracking. All variables were used as a measurement of effort, three were used to measure load (cognitive load and workload), and

only one variable was used as a measure of cost. The dual-task method was associated with three dependent variables. All variables were used as measures of effort, and two variables measured load (cognitive load and workload). The dual task method was not used in any study to measure cost. Finally, each of the physiological measures had only one associated dependent variable - all of which were used as measures of load (cognitive load and workload). Refer to Table 8 for a more detailed overview.

Objective Method Proposed/Used	Construct & Number of Articles	Citation
Search Interaction Logs	Cost ($N = 14$)	[7–10, 22, 29, 80, 91, 96, 107, 114, 129, 139, 144]
	Effort ($N = 38$)	[1, 11, 18, 21, 40, 41, 47, 65, 77, 83, 98, 104, 106, 112, 115, 131] [3, 17, 24, 27, 50, 53, 55–57, 57–59, 62, 66, 79, 86, 88, 90, 106, 119, 132, 141]
	Cognitive Load ($N = 2$)	[116, 133]
Eye Tracking	Cost ($N = 1$)	[144]
	Effort ($N = 9$)	[22, 27, 28, 47, 53–56, 67]
	Cognitive Load ($N = 1$)	[133]
	Workload ($N = 3$)	[34, 70, 101]
Dual Task	Effort ($N = 3$)	[50, 77, 115]
	Cognitive Load ($N = 4$)	[23, 33, 52, 116]
	Workload ($N = 1$)	[143]
Other Physiological (EEG, Temperature, ECG, EDA, fNIR)	Cognitive Load ($N = 1$)	[138]
	Workload ($N = 2$)	[70, 92]

Table 5. Objective methods used/proposed to measure CEL in reviewed articles

Subjective Methods. For the purpose of this review, we consider subjective methods as those which rely on human judgement of some kind. This also includes self-report judgements which may relate to objective search behaviours, for example, "what were the total number of sources consulted during your search?". Across all the reviewed articles, 39 studies used subjective methods to measure CEL. Four types of subjective method were used across all studies, see Table 6 for an overview of the subjective tool used and the construct measured. The NASA-Task Load Index was the most frequently used subjective method, with 22 articles using the tool as a measure of cost, effort, and load (workload and cognitive load). Eleven of these articles used the full six-component version (physical demand; mental demand; temporal demand; performance; frustration; and effort). To shorten the test, three articles [3, 5, 13] used the raw version, where each of the six components is weighted according to the context using a separate instrument. Five studies [5, 24, 116, 117, 143] omitted individual components if they were considered less relevant to the task. Two articles [35, 46] did not specify which version they used. The majority of studies (64%) which used the NASA-TLX as a measure

of workload used the tool to compare user workload across search systems [2, 3, 8, 13, 35, 37, 46, 71, 74, 80, 116]. The remaining 36% of studies used the tool to compare user workload across search tasks [3, 5, 14, 24, 89, 120, 143].

Self-designed questionnaires were used in 14 articles measuring cost, effort, and workload. These questionnaires were all different, varying in relation to format, scale, and unit of analysis. See Table 9 for the article, the specific questions used, and the unit of analysis. Finally, the Workload Profile (WP) [89] and the Mental Workload Test (MWT) [34] were used in two studies as a tool to measure workload.

Subjective Method Proposed/Used	Total Number of Articles	CEL Construct Measured & Number of Articles	Citation
NASA Task Load Index (NASA-TLX)	22	Cost ($N = 1$)	[80]
		Effort ($N = 4$)	[6, 24, 59, 130]
		Cognitive Load ($N = 2$)	[116, 117]
		Workload ($N = 17$)	[2, 3, 13, 14, 24, 37, 46, 74, 80] [4, 5, 8, 35, 71, 89, 120, 143]
Self-Designed Questionnaire	14	Cost ($N = 2$)	[91, 114]
		Effort ($N = 11$)	[19, 45, 68, 69, 72] [77, 79, 87, 115, 119, 124]
		Workload ($N = 2$)	[19, 134]
Workload Profile (WP)	1	Workload ($N = 1$)	[89]
Mental Workload Test (MWT)	1	Workload ($N = 1$)	[34]

Table 6. Subjective methods used to measure CEL in reviewed articles

Objective and Subjective Methods. Ten articles used a combination of both objective and subjective measures of CEL within a single study. The majority of these ($N = 6$) [59, 79, 90, 91, 114, 143] used two different methods, with the most common combination as self designed questionnaires and search interaction logs ($N = 4$) [79, 90, 91, 114]. Three articles [77, 115, 116] used more than two different methods. See Table 7 for a more detailed overview.

Independent Variables From the 91 reviewed articles, 121 independent variables were extracted and categorised into five groups: task characteristics; search engine results page (SERP); system; individual differences; and document/web-page.

- *Task Characteristics.* Task characteristics (35%) were the most widely examined independent variable across all studies (cost: $N = 3$; effort: $N = 29$; cognitive load: $N = 5$; workload: $N = 7$). More specifically, the different task characteristics manipulated in these studies comprise six sub-categories: *structure; product; determinability; goal;*

complexity; and *difficulty.* Task complexity was the most popular task characteristic introduced across the studies ($N = 21$) and the majority of studies [24, 34, 56, 57, 101, 119, 120, 133, 144] designed tasks at only two levels of complexity (simple vs. complex) - where simple was a "fact-finding" task and complex an "information gathering task". Studies which designed tasks of up to five levels of complexity [3, 65, 91, 114] tended to base these tasks around Anderson and Krathwohl's Taxonomy of cognition, i.e., *remember, understand, analyse, evaluate, and create.* While task difficulty was identified as an independent task characteristic [51, 52], the task manipulations were very similar to that of task complexity (i.e., fact-finding vs. information gathering) as were the terms used (i.e., simple vs. complex). Three studies examined the effects of task product (factual vs. intellectual) on user effort [27, 68, 69].

Task goal was also examined at two levels, mainly the effects of a specific vs. amorphous goal on user effort [27, 68, 69]. Task determinability was used in three studies [3, 17, 18] which used similar manipulations but with varying numbers of conditions. For example, Capra et al. [18] initially examined how four levels of task determinability (unspecified; specified items; specified dimension; specified both items and dimension) affect user effort, and in a later study examined the effects of six levels of determinability (unspecified; items; objective dimension; subjective dimension; items + objective dimension; items + subjective dimension) [17]. Finally, task structure was usually described as *simple, hierarchical,* or *parallel* and investigated in relation to its effect on user effort [19, 57, 104] and cognitive load [19, 52].

- *The Search Engine Results Page (SERP).* The second most commonly used (31%) independent variable across all of the articles (cost: $N = 8$; effort: $N = 8$; cognitive load: $N = 3$; workload: $N = 13$) was the search engine results page, or the query interface. This is the page that users encounter when they navigate a search engine to enter a query and find specific information. Across the reviewed studies, the SERP was adapted in a variety of different ways. The majority of these studies ($N = 14$) examined the effects of result list adaptations. For example, additional result list elements/tools such as; knowledge modules [2]; entity cards [13]; information structuring tools [80]; tag clouds [57]; and additional categories [143] have all been examined in relation to their effect on the users cost [80], effort [57], and workload [2, 13, 143]. The presentation of results have also been examined in various ways, including the effects of; blocked vs. interleaved results list on user effort and workload [3], SERP size (3,6,10 results per page) on user workload [74]; list vs. overview on user effort [57]; text-based vs. graphic based on user effort and cognitive load [19]; and arc vs. list vs. grid display on user workload [134]. Four studies examined the effects of manipulations at the query level of the SERP. For example, Gerwe and Viles [45] studied the effects of an advanced query interface with additional search features vs. a basic query interface on user effort. Edwards et al. [37] examined the effect of a standard query box vs. a structured query box (series of boxes for query terms) on user workload. Other studies have focused on the effects of query suggestions on user cost [8] and cognitive load [71]. Finally, the effect of the SERP result quality (low vs. high relevance) on user effort [91] and cost [107] has been studied. Alongside how user effort may be impacted by both the opinion (consistent vs. inconsistent) and credibility [112] of SERP results.
- *System.* The system was used as the independent variable in thirteen of the studies (cost: $N = 1$; effort: $N = 10$; cognitive load: $N = 1$; workload: $N = 1$). The majority of studies ($N = 12$) compared two independent systems (i.e., system A vs. system B) in relation to their effect on user cost, effort, and load. For example, Kim and Rieh [77] and Rieh et al. [115] compared the user effort invested while using both a library system and a web system. The remaining study examined user effort in relation to two systems which shared commonalities but had different back-end algorithms [72].

	Self- Designed Questionnaire	NASA-TLX	Mental Workload Test	Search Interaction Logs	Dual-Task	Eye-Tracking
Cost [91, 114]	X			X		
Effort [77, 115]	X			X	X	
Effort [59]		X		X		
Effort [79, 91]	X			X		
Workload [34]			X			X
Workload [143]		X			X	
Cognitive-Load [116]		X		X	X	

Table 7. Studies which use both subjective & objective measure Cost, Effort, or Load

- *Participant Individual Differences:* Eleven articles (cost: $N = 1$; effort: $N = 8$; workload: $N = 8$) investigated individual differences such as cognitive ability, skill, and experience, as an independent variable. Working memory was the most frequently examined cognitive ability, with studies interested in its effects on user effort [24, 53, 55, 57] and workload [3, 24]. Other cognitive abilities which were examined across studies include; associative memory [14]; perceptual speed [3, 14]; verbal closure [57]; inhibition [3]; visualisation ability [14]; and spatial ability [120]. Besides cognitive ability, the effect of typing speed on user costs [107], and domain knowledge on user effort [28] were also of interest. Finally, the level of user experience was included as an independent variable in the measure of user effort [40, 104].
- *Document/Web-page/Landing-page:* Finally, the document or web-page was considered an independent variable in eight studies (effort: $N = 6$; cognitive load: $N = 3$). In the majority of studies, the level of document relevance was examined in relation to user effort [54, 59, 87, 91, 130]. The remaining studies which measured cognitive load, adapted the visual complexity of documents (e.g., number of elements included on the page) [23, 133].

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 8. Dependent variables used/proposed to measure CEL in reviewed articles

Construct	Questions	Unit of Analysis
Effort		
	[45] 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)
	[72] One question: "How much effort did it take to complete the task?"	Scale from 1 (very little effort) to 7 (a lot of effort)
	[68] 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)
	[69] 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)
	[87] One question: "rate your effort to answer this question well"	Scale from 1 (low) to 5 (high)
	[19] 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)
	[77, 115] One question: Rate your "effort invested in searching"	Scale from 0 (no effort) to 10 (a great deal of effort)
	[91] Perceived time estimation	Perceived time (s) examining each document
	[79] 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		
	[114] 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)
	[91] One question: "Estimate the duration spent on searching" (after each task)	Perceived dwell time (s)

Table 9. CEL constructs measured and the questions used in self-designed questionnaires

4.4 Relationship between Constructs and Measures:

This section describes how each construct and its respective conceptual categories have been measured within ISR. Table 10 provides a detailed overview of the conceptual categories associated with each construct and the measures/dependent variables used to measure them. Figure 3 provides a broader overview of these relationships, and more clearly illustrates areas of overlap between the different constructs, conceptual categories, and measures.

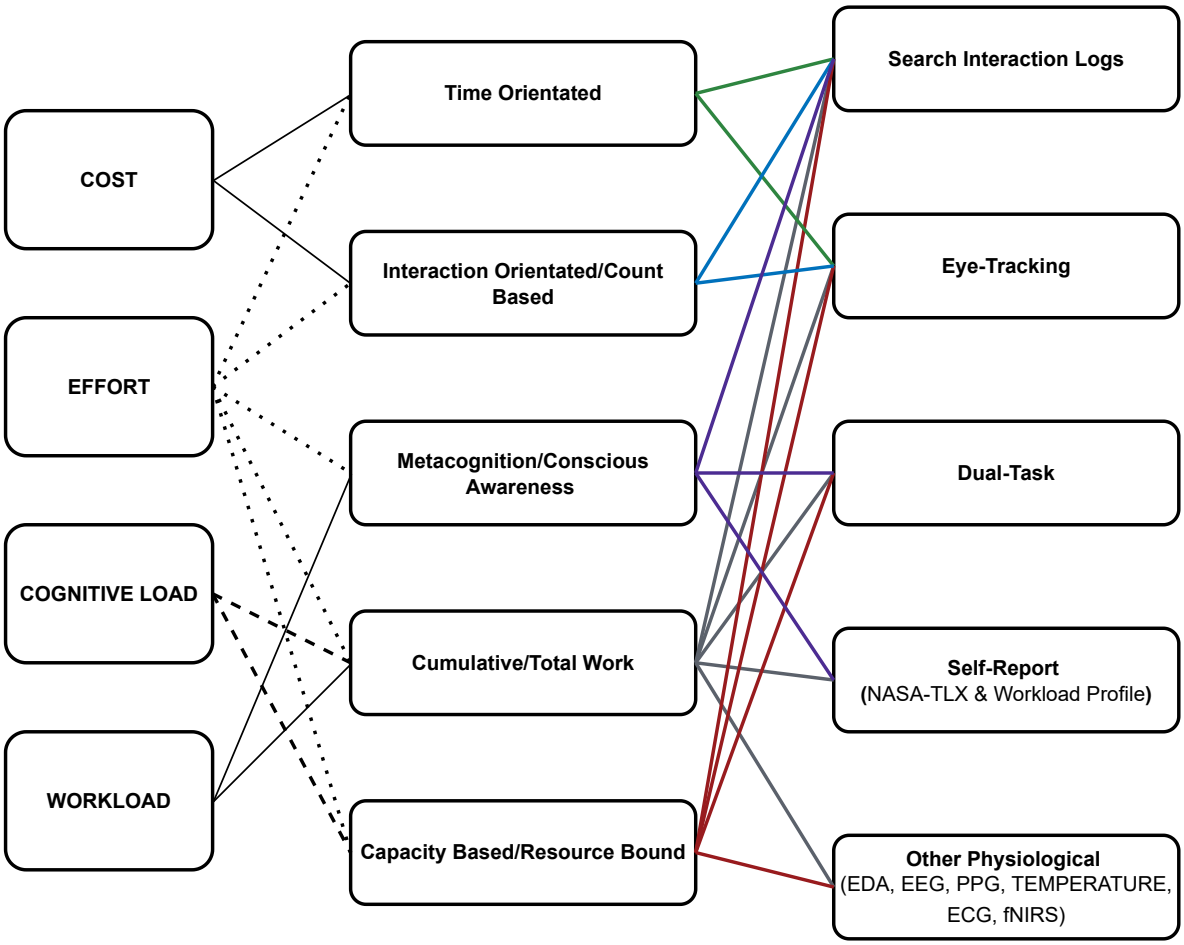

Fig. 3. Relationship between Constructs and Measures

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 10. Relationships between Constructs and Measures

5 DISCUSSION

Based on the analysis presented in the previous section, this section will provide a critical overview of the relationship between how CEL constructs are defined and how they are measured within ISR.

5.1 Cost

Our review highlights that the construct of cost has faced somewhat sporadic examination within the field of ISR over the last several decades. Interest in the construct appears to decrease from the early 1990s, before making a revival almost two decades later in 2009. If we look more closely at the treatment of cost between these two time periods, we can see a distinct shift from the measurement of "fiscal costs" to "temporal costs". The examination of "fiscal costs" seems well justified considering the landscape of the field from the late 1970s to early 1990s - where researchers examined cost as a means to justify the use of online bibliographic IR compared to manual IR within organisations. The emergence of online search engines in the early 1990s, led to a highly competitive era for search, with network latency and download speeds becoming increasingly influential on user experience [83]. With this brought revived interest in examining cost, and following several published papers by Azzopardi and colleagues [7–10, 96], cost has generally been treated in relation to "temporal costs" as a means to evaluate system efficiency and effectiveness.

The measurement of cost generally reflects how it has been conceptualised within ISR. Studies which characterise cost as *Time-Orientated* mainly used time-based measures derived from search interaction logs. For the remaining studies which did not provide a definition of cost, a similar trend was observed, where search interaction logs, and particularly time-based measures were most commonly used. If we examined each of these studies and their respective measures of cost in isolation, then it is likely that we could claim some degree of internal validity as time-based measures may accurately reflect the time-oriented nature of cost. However, while the unit of measurement "time" is consistent across the studies, it is operationalised in a multitude of different ways - such as the time taken to; view search results, complete a search task, formulate a query, and read and assess documents. Subsequently, it becomes difficult to make comparisons between these studies- as each imply that cost has a "fixed" value. However, is the time taken to complete a search task more costly than the time taken to issue a query? Furthermore, the reviewed studies only consider "time" in their operationalisation of cost - but is "time" really the only cost paid by the user during the ISR process? Perhaps a limitation of these studies is the scope and generalisability of the research. Within this review, the majority of studies examining cost used a laboratory-based study with university students as participants. In these scenarios, costs were mostly representative of user-sided costs. However, if we consider how cost has been operationalised outwith ISR in relation to monetary or human resource, can we assume that similar costs are also paid or spent by the user within certain ISR processes? If we consider a professional search context, then it is likely that the amount of time, money, and human resources spent during the search process are costly to both the user *and* the organisation.

5.2 Effort

The examination of effort within ISR has followed a similar trajectory as cost, receiving some attention during the 1970s-1980s, before tapering off until it's revival as a construct of interest in the late 1990s. In these early studies, effort was often conflated with cost and primarily discussed and measured in relation to an organisation, and the amount of labour (time/salary) involved in using different systems. Perhaps the rise of search engines in the 1990s prompted the resurgence of effort examination within ISR, when it was acknowledged that effort played a key role in influencing user search decisions and behaviours. Between 2009-2018, interest in examining effort peaked within ISR. However, while

constructs such as cost became associated with specific research groups around this time and therefore treated fairly consistently across studies, the examination of effort became almost more disconnected. Rather it appears that there are few research groups which consistently examine effort. Moreover, for studies which examine effort less explicitly, citations to key works in the area are often non-existent.

All five conceptual categories were represented by the definitions of effort - perhaps indicating that either the authors have different interpretations of effort, that the construct of effort is multi-faceted, or that there is simply no standard and widely accepted definition of effort. Not only did the definitions of effort span across a range of conceptual categories, similar observations were found in relation to the way that effort has been operationalised within ISR. Search interaction log measures such as *total interaction count* and *time-based* variables were used to operationalise effort in studies from all conceptual categories. However, if we take the measures, "task completion time" or "number of queries issued", these are unlikely to operationalise effort as described by all five conceptual categories. When we examine studies which did not include a definition, the range of measures used to operationalise effort become even greater. This leads to the question of why effort has been defined and measured in so many ways within ISR? If we take a closer look at the definitions provided, it seems that researchers tend to label individual components of effort rather than defining the "overall" construct. Take these definitions for example, *"effort reflects a voluntary allocation of effort that can be reported by the individual"*, and *"the total work done to achieve a particular goal"*. If we compare these definitions to existing theory, which characterises effort as both subjective [38] and reflective of the "work done" [64], then these definitions are relatively representative. However, these definitions only define one dimension of effort - rather than the construct as a whole. This problem combined with a general lack of clarity on the conceptualisation of effort (both within ISR and other domains), has perhaps led to the proliferation of the number of variables which are then operationalised as "effort" within ISR. This "kitchen sink" type approach to labelling or measuring a variety of search interaction variables as "effort" is perhaps not based on a lack of agreement between researchers, but rather a lack of common language to define effort in its entirety.

Self-designed questionnaires were also a commonly used measure of effort. While self-report tools can perhaps offer a more explicit indication of users perceived effort than search interaction logs variables, several issues came to light on examination of these questionnaires. As demonstrated in Table 9, it is clear that there is generally very little standardisation in relation to the questions, scales, unit of analysis, and format used in these types of questionnaires. Questionnaires used to measure effort employed a range of different scales: (i) a seventeen point scale (1-17), (ii) a seven point scale (1-7), (iii) a five point scale (1-5), (iv) eleven point scale (0-10). This lack of standardisation of thresholds in these scales, may suggest different levels of effort, e.g., 5 = high level of effort vs. 17 = high level of effort - again making it difficult to make comparisons between studies. While these scales may assume face validity, the notion of whether an individual can accurately rate their own cognitive capacity is questionable. Subsequently, this can lead to issues in comparing individual ratings.

5.3 Cognitive Load

Studies examining cognitive load appeared within ISR research almost two decades after cost and effort - and to this date, cognitive load has received far less attention. Few studies provided a definition of cognitive load. This finding may be linked back to the lack of formal theory used within studies- as the construct of cognitive load is situated within the fundamentals of CLT. With only few definitions to extract conceptual elements from, the inferences we can make about the conceptual categories and their measures are subsequently limited. The studies which defined cognitive load as *Capacity Based/Resource Bound* used or proposed measures derived from dual task, eye tracking,

and electroencephalogram (EEG). As these methods are considered as direct and objective and designed to capture the dynamic and instantaneous properties of cognitive load [15], they appear to be an appropriate operationalisation of the constructs *Capacity Based/Resource Bound* properties. Furthermore, these methods were also primarily used within studies which did not provide a definition of cognitive load, perhaps indicating some degree of consensus among researchers both in relation to their understanding of cognitive load and a general awareness of the appropriate methods of measurement. However, in order for these measures to accurately reflect the conceptual properties of cognitive load it is important that they are analysed at the correct level of granularity. That is, in order to examine the detailed trends and patterns of cognitive load, the data needs to be examined at the fine-grain level. However, across the reviewed studies this was not often the case. For many of studies which used the dual-task method, reaction time and missed event measures were averaged across the entire search session. Similar observations were made for eye-tracking measures such as fixation duration and number of eye fixations. When results are analysed at the session level, only a static, post-hoc analysis of cognitive load can be made, potentially masking the dynamic interplay between demand and the consumption of user resources. It is important to point out however, that in most cases, the studies which provided a definition of cognitive load, analysed the data at a finer granularity e.g., task segment level. Further supporting the assertion, that conceptualisation should precede operationalisation for effective measurement.

The task and interface were frequently used as independent variables in studies examining cognitive load, which may allude to the recognition of the different load types (i.e., task - intrinsic; interface - extraneous). However, similar to the lack of definition and reference to existing theory, the majority of studies examining cognitive load made no explicit reference to or distinction between the different types of load (i.e., intrinsic, extraneous, germane) in their experimental design. There are several issues with this. Firstly, studies which measure cognitive load as a singular construct are unable to determine which type of load is consuming the users cognitive resources during the search task. This may be particularly problematic in studies which used both task and interface manipulations within a single study. While perhaps these studies could still claim the measurement of "overall load", assuming the load types are additive, the problems associated with lack of distinction of load types may be more pronounced in studies which inadvertently attempt to manipulate only one load type - as more conclusive, and likely conflated inferences may be assigned to the findings. The lack of reference to CLT and the different types of load also raises the question as to what empirical basis the experimental manipulations of the task and interface were based on.

5.4 Workload

Similar to cognitive load, workload was not examined within ISR until the late 1990s, with most research attention for the construct occurring in the last decade. As with cognitive load, there were few definitions of workload provided. Studies which defined workload in relation to the *Cumulative/Total Work* used a range of different methods including eye-tracking, self-report, and physiological methods. However, for studies which did not include a definition of workload, there was an overwhelming reliance on the NASA-TLX as the method of measurement. This observation is not unique to the domain of ISR. Since its conception in the 1980s, the NASA-TLX remains the most widely used tool for measuring workload. It has even been argued that the NASA-TLX is so heavily associated with workload that it has almost become definitive of the construct [31]. This may provide an explanation as to why the majority of studies in this review which used the NASA-TLX did not provide any definition of workload. Hart and Staveland [61], the developers of the tool promote it as the most valid and sensitive indicator of the construct. Thus, while the integrity of the tool itself is perhaps less questionable, there have been several issues with the way that it has been employed within ISR. Firstly, the tool is designed to compare workload among tasks, however the majority of studies used it to compare workload

among systems- subsequently this observation questions the validity of the claims made in these studies. Secondly, the demands imposed by the task are likely to fluctuate on a moment-to-moment basis [95]. However, the majority of studies administered the NASA-TLX following task completion. In this case, the assessment of workload relies on the integration of multiple memories from the task, which in some cases can be done reasonably accurately, in others cases not. This integration may be further compounded and disproportionately influenced by recall of episodes of peaks or deviations in workload [95]. As a result, the overall workload scores obtained post-hoc, may not accurately reflect the users experienced workload throughout the task. Finally, is the NASA-TLX applicable to ISR specific tasks? The NASA-TLX was designed for use in Aviation and is still predominantly used as a tool of workload measurement in high-pressure operational environments such as Air Traffic Control, Military, and Healthcare where tasks include flying, driving, surgically operating etc.[60]. As ISR tasks are generally of a lower demand than these examples, it is uncertain whether the tool is sensitive enough to detect lower levels of workload. Furthermore, the question of whether items such as "physical demand" is relevant to ISR tasks also remains uncertain.

6 IMPLICATIONS

This review highlights a body of high-quality literature which mostly all demonstrate and evidence the use of a measure - exemplifying potential capabilities and limitations which can be used to inform future research. In section 5 we discussed the relationship between how these constructs are being conceptualised and measured - highlighting areas of similarity and difference among ISR researchers and potential issues with current CEL examination. This next section will discuss the implications of these issues, and provide suggestions for future CEL research.

6.1 Conceptualisation and Operationalisation of CEL

It has been proposed that the field of ISR, is largely driven by innovation and technology as opposed to the development or use of theory [73]. This focus on the applied and practical elements of science, has led to the prominence of results over explanation - leaving many studies lacking theoretical motivation [73]. The findings of this review may reflect this. Firstly, the majority of studies did not provide explicit definitions of CEL and instead relied upon intuitive notions of these constructs, rather than the use of accepted or established definitions from existing theory. Similarly, rather than framing CEL measurement within well-established definitions and theory, it appears that researchers may rely on the *face validity* of an instrument - where the measure is not formally validated, but makes intuitive sense and is therefore accepted as an appropriate measure by the research community [73]. As a result, the last 50 years of ISR research has experienced little maturation or progress in relation to our understanding and measurement of CEL. Rather, it appears there still exists no universal definitions or standardised solutions for measuring CEL, and no single method that could be recommended as "gold standard" following this review. Thus, in order to discover these "gold standard" methods of CEL measurement within ISR, it is necessary that the first step of effective measurement, *conceptualisation*, is fulfilled. Without the provision of definitions and theories of CEL within ISR, this area of research will struggle to advance. Therefore, in order to achieve maturation of CEL understanding and measurement, it is important that the field of ISR coordinates efforts to ground CEL research in established theory.

This lack of conceptualisation has ultimately led to issues with the operationalisation of CEL constructs - if the researchers interpretation of the construct is unknown, then how do we know if the instrument is accurately measuring its conceptual properties? Without this knowledge, the *construct validity* (i.e., the extent to which the instrument measures the intended construct) of the measurement is uncertain [63]. This issue is further exacerbated by the overlap in measures assigned to different constructs. In order for studies to produce meaningful data, it is important that these

methods are validated to ensure they are measuring what they claim to measure. Very few of the studies examined in this review discussed or attempted to validate measures of CEL. While validating measures often involves conducting a number of studies developed exclusively around the measure [73], there are less arduous steps researchers can take to improve the validity of measures. Providing clear and concise definitions of the construct, combined with reference to well-established theory which provides clear justification for the use of a particular measure, can lead the field towards effective operationalisation of the construct - which in turn will improve the validity, comparability, and standardisation of CEL research and measurement [43].

In our recent perspectives paper [97], we created a tentative framework for defining CEL and their related constructs, based on how they have been used, investigated, and measured within ISR research. We developed these with the hope that the ISR research community can benefit from these working definitions and we actively encourage others to use and build on these definitions within their own research to help develop a more unified approach in understanding and researching CEL constructs in ISR. These working definitions and the respective diagram have since been updated according to new insights gained from this present review, specifically in relation to the physical resources that are available to the user. Demand on the users physical resources during the search process became more salient following this review. The previous framework placed emphasis on the internal resources of the user in relation to cognition, for example the use of their working memory and perception. However, search interactions such as clicks, scrolling, and typing were all frequently implicated as measures of user effort in this present review. We deemed such actions to consume not only the users cognitive resources, but also their physical resources such as strength, motor action, and metabolic energy. Additionally, the consumption of physical resources during the search process and the implications on user effort, is an important factor to consider from an accessibility standpoint in future research.

Below we describe our framework and show how these constructs are related in Figures 4 and 5.

Resources: According to CLT and MRT, people have multiple resources available to them. In the context of ISR, we can generalise these resources and delineate them as: (*i*) **internal resources** that pertain to the user. These can be either cognitive (e.g. working memory, attention, etc.) or physical (e.g. metabolic energy, strength, etc.), and; (*ii*) **external resources**, which the user has available to them (e.g. time, money, labour, etc.)

Resource Capacity: All resources are limited in capacity (e.g. the number of items that can be held in working memory or the amount of time available to complete a task). The capacities of resources are not fixed, and may vary over time. For example, through practice or training a user may increase their working memory capacity, but if they are stressed or fatigued, then this capacity may be reduced. Alternatively, if a deadline is suddenly moved forward then the amount of time available is decreased, however if the deadline is extended, then the amount of time available is increased.

Demand: Demands emerge from the properties of the task, system, and more generally the context. 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 and will fluctuate throughout the course of the task.

Load: In alignment with theory from Psychology, we categorise cognitive load and workload under the umbrella term “load”. Given a particular resource, and the demand imposed by the task, system, and context, we can generalize the construct of load from CLT in the context of ISR to refer to the amount of resource (internal or external) being consumed at a given point in time.

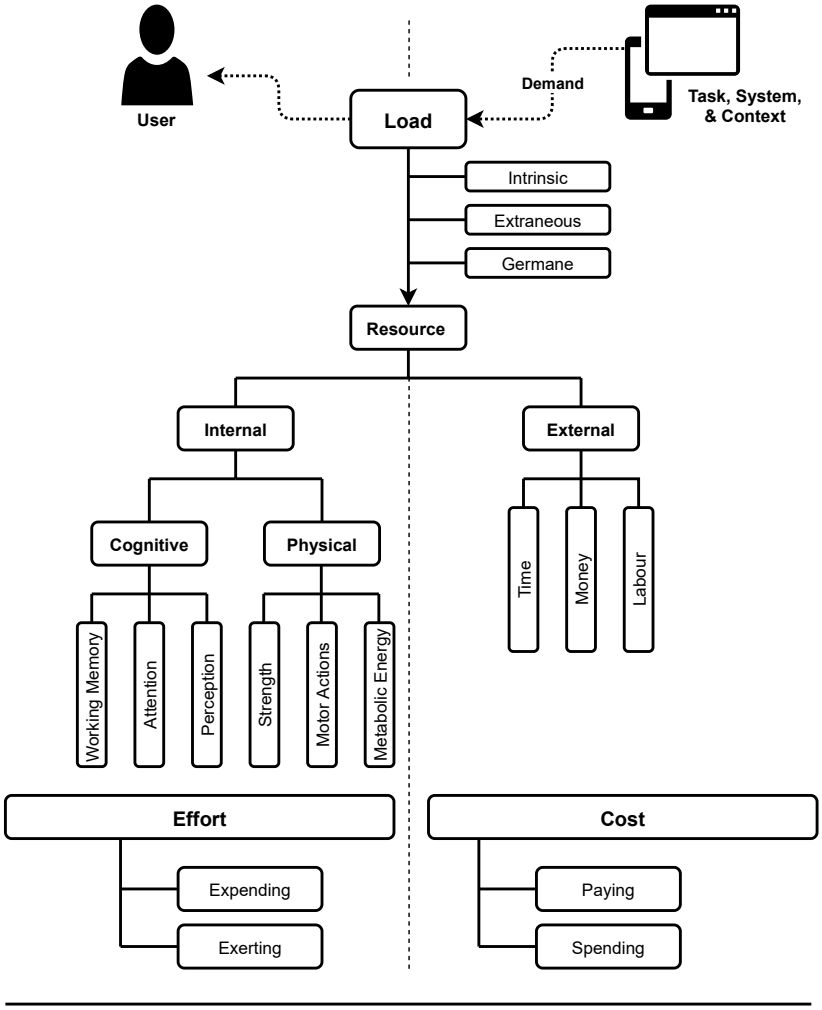

Fig. 4. The relationships between CEL constructs

Overload: Taken together, the construct of overload occurs when the demands of the task, system and context exceed the capacity of the resource(s). For example, if the amount of working memory or attention required exceeds the individual's capacity they are likely to experience overload.

Effort: In the context of ISR, we see effort as a user-sided construct that reflects the total amount of *internal* resources that are *exerted* or *expended*, over a given period of time, in order to meet the demands of the task, system and context. In Figure 5, the bottom plot shows how effort is related to load, where effort is the total load experienced over time (i.e. the area under the curve).

Cost: We delineate cost from effort specifically in terms of the resources they relate to. Cost is considered with respect to *external* resources (e.g. money, time, human resources etc.) that are *spent* or *paid* by the user in order to meet the demands of the task, system and context.

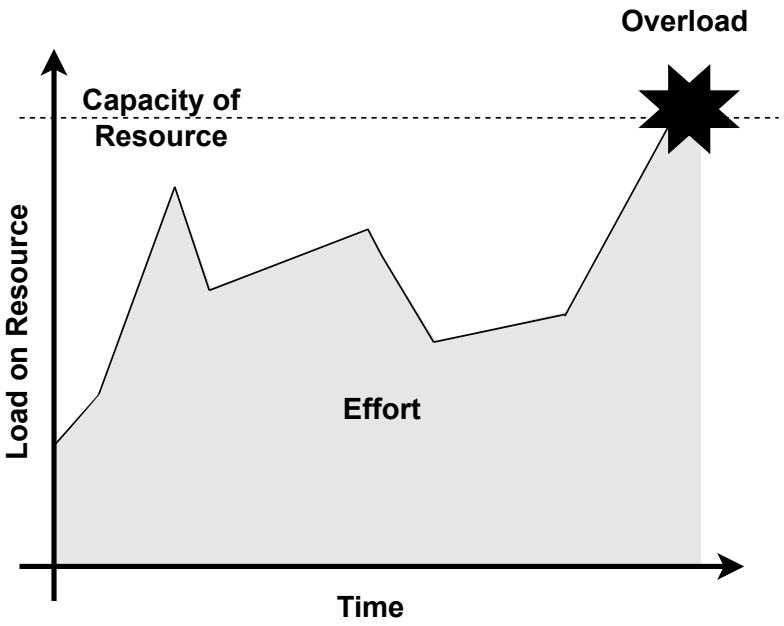

Fig. 5. A graphical depiction of the relationship of the load experienced by a user over time for a given internal resource. When the load demanded by the task, system and context exceeds the capacity of the user's resource, then they hit overload. The effort experienced by the user is the total load over time (i.e. the area under the curve).

In the context of ISR, the above definitions come together as follows: During an interactive search task, demands will arise from the characteristics of the search task itself (i.e. task difficulty) and also from the system (i.e. search engine result page layout, etc.). The user has internal (cognitive and physical) resources they can draw on to attend to these demands, such as holding information in their working memory, or they may draw on external resources, such as asking a colleague for help. If the demands become too high, these resources will reach their upper limit, and the user will experience overload. In this case, the user may experience a decline in performance or stop the search task altogether. In order to allocate resources across the duration of the task and to reach their task goal, the user must consciously exert some kind of physical (e.g. typing a query) and cognitive activity (e.g. examining a results page). The amount of effort exerted will depend on the amount of load experienced. As the user reaches the end of their search task, cost can be considered as the external resources consumed or spent, for example the time spent on the task.

6.2 Problems with Existing CEL Measurement

This review has illustrated that there has been little innovation in the development of CEL measures within ISR. Search interaction logs as methods of cost and effort measurement are as prevalent now as they were 40 years ago. While measures derived from search logs offer an unobtrusive, relatively easy to collect proxy for cost and effort their use has been criticised. For example, implicating time as an indicator of cognitive processing has been argued as "too simplistic", under the contention that it does not sufficiently consider the effects of other confounding variables such as the task, topic, and other individual characteristics [75]. The use of measures derived from search interaction logs assume that users will behave in a predisposed or stereotypical way, however as every user interaction with a system

and its information is unique in terms of the physical, cognitive, and affective experience - is it possible for these fixed values to accurately capture the individual experiences of the user? It has been argued that the value of search interaction signals can only be obtained if there is a strong consideration for the context, purpose, and nature of what is being examined [73]. However, even in these circumstances, there is still uncertainty as to which construct is actually being measured. Measures of cost and effort such as number of mouse clicks have been implicated in other studies outside this review as strong indicators of topical interest [26, 140]. Similarly dwell time and time-on-task, frequent measures of cost and effort, have also been used as measures of user interest [26]; user satisfaction [42]; and relevance [49]. Considering their prevalence and long-term use, it is likely that search interaction logs will remain a prominent method of cost and effort measurement within ISR, therefore it is important that these measures are used as accurately as possible. As discussed, measures derived from search logs used in isolation are unlikely able to represent and explain complex phenomenon such as cost and effort. However, methodological triangulation is a technique which can help increase the credibility and validity of research findings [113]. This approach generally promotes the use of multiple data collection methods within a research study in order to overcome any fundamental biases which may occur from using a single method [113].

More generally, effort and workload were the most widely measured constructs across the reviewed studies. A key observation however, was that both effort and workload were rarely treated as the primary focus of the study. Rather it appeared that many studies included effort and workload measures, particularly in the form of self-report, as a "quick and dirty" approach for gathering user perceptions of the task or interface demands. While subjective methods can provide an indication of the overall task demand or the total work performed by the user, the ad-hoc and subjective nature of these measures can tell us very little about which specific elements or aspects of the task, system, or interface led to higher effort exertion or workload. Cognitive load on the other hand, when measured correctly, has the potential to inform researchers of specific moments in time when the user experiences increased load. As high levels of cognitive load can be detrimental to both the individual and task performance, understanding where these higher levels of load occur can be very beneficial to advancing interface design, and developing tools which could support the user during the performance of more complex tasks. Considering the important insights to be gained from measuring cognitive load, the construct itself was examined the least of all CEL constructs. More interestingly, cognitive load has not been measured within ISR since 2014- around the same time that examination of workload took off and has since prevailed. We can only speculate reasons why this may be - perhaps the complex nature of cognitive load combined with the lack of a "quick and dirty" measure has rendered the construct less appealing to measure? Nevertheless, we argue that there may be more insights to be gained from measuring constructs such as cognitive load within ISR - particularly in understanding what specific elements of the search process contribute to a users cognitive load and how this may impact system usability and user task performance.

6.3 Suggestions for Future Research

Similar to the prevalence of search interaction logs as measures of cost and effort, the NASA-TLX has become the silver bullet of workload measurement within ISR. However, the over-use of this tool may be indicative of the wider issue, where face validity alone is consideried adequate justification of an instruments utility. Considering the problems associated with the use of the NASA-TLX within ISR and self-report measures more generally, perhaps it is necessary for the field to explore alternate methods of workload measurement. Fields outside ISR, such as Human Factors have in recent years advocated the use of physiological measures (see Charles and Nixon [20] for a review) as a robust measure

of mental workload. Signals derived from electrocardiographic, respiratory, dermal, and blood pressure measures are all found to discriminate between the mental workload imposed by task type, task demand, and task difficulty [20].

While we have argued the advantages of cognitive load examination, these can only be gained if the different types of cognitive load are acknowledged when measuring cognitive load. While the findings of this review highlight a lack of measurement instruments applicable to the different types of load, utilising tools from other domains, such as the the self-report questionnaire by Klepsch et al. [81] designed to measure all three different load types could offer a useful alternative. However, acknowledging the well-known shortfalls of using self-report tools to measure dynamic constructs such as cognitive load, future studies should aim to triangulate these with objective methods, to enhance the validity of findings [15]. As with workload measurement, other domains outwith ISR such as Cognitive Neuroscience have demonstrated the use of highly sensitive and precise physiological measures of cognitive load, i.e. functional Magnetic Resonance Imaging (fMRI), where resource consumption in the brain can be measured directly and has the potential to distinguish between the three types of load [123]. While these types of measures offer a promising future for cognitive load measurement, their widespread deployment and use during in-situ or naturalistic experiments is currently quite limited. Nonetheless, it is worth exploring a variety of methods to measure CEL constructs to determine the trade-off between their scalability and accuracy. As a larger body of research examining the different types of cognitive load emerges, a richer understanding of which techniques are most applicable to each load type can be gained and in turn, some degree of unity can be formed in the standardisation of cognitive load measures within ISR.

Finally, this review highlighted that many of the studies may lack *external validity*, i.e., how well the results of a study can be expected to apply to other settings. The majority of studies included higher education students as participants, often using tasks which were not specific to their area or domain of study. As students are not the only users of search systems, these results are likely not very generalisable to real-world populations. There is also the possibility that CEL is more salient in other contexts and populations - which may not be well reflected among a student cohort or tasks which are fairly inconsequential. If we consider professional search, then increased cognitive load during a search task may pose significant problems. For example, information overload is a common problem within Legal IR [127], largely due to the demanding characteristics of documents and the need to find accurate and complete information in a limited amount of time [84]. Missing relevant documents due to these increased demands can have costly consequences such as undermining knowledge acquisition and even access to justice [93, 127].

Similarly, when conducting systematic reviews within the Healthcare domain, the systematic reviewer is to assess or "screen" which articles are pertinent to the research question - and which are not [102]. As with the Legal domain, the consequences of missing a relevant document are high -where potentially life-saving medical research could be missed. Considering the importance of CEL within professional search domains such as legal search or healthcare, this review highlighted that very little research has directly examined these constructs within these search contexts. Thus, to gain a fuller understanding of the complexities and nuances of CEL, the ISR community should strive to recruit a more diverse user base to participate in these studies, particularly in search contexts where problems with CEL may be more salient and consequential.

6.4 Acknowledgement of Contributions and Limitations of Review

On a final note, it is important to acknowledge the contributions made by researchers of CEL within ISR and the limitations of this review. Firstly, all of the articles included this review have provided a significant contribution to our understanding of CEL within ISR. The purpose of this review is not intended as a critique of these research efforts but rather to discover and analyse areas of divergence and consensus between researchers in relation to how CEL is defined

and measured. Secondly, this review does not claim to include every article written about CEL within ISR - rather this set is only representative of CEL studies which fulfilled our inclusion criteria, and even then it is still possible that some articles were missed due to limitations of the database search or through human error. Additionally, coding the articles, particularly in relation to the definitions and measures was not always a straightforward task. Definitions were not always stated explicitly and subsequently these were more difficult to identify and extract than studies which provided explicit definitions. Similarly, some studies were not explicit about which units of measurement were used as indicators of CEL, particularly in the results section of articles. In these cases, the more qualitative nature of coding may imply our reported results are less replicable than others derived from more explicit definitions and measures. Finally, it is important to raise the issue that we did not consider other constructs related to cost, effort, and load such as difficulty and complexity in our working definitions.

7 CONCLUSIONS

The purpose of this review was to document and examine the constructs of cost, effort, and load, as they are currently defined and measured in studies across the field of ISR. A key goal of this review was to discover and analyse areas of convergence among ISR researchers in relation to how each CEL construct is defined and subsequently measured. Despite the lack of definitions provided, the review identified several conceptual similarities among researchers. This convergence may provide a potential starting point for development towards a unified approach to CEL definition within ISR. Even without a unified approach to CEL definition within ISR, it is hoped that future research will focus efforts on providing clear and precise definitions and ground empirical work in well established theory. By framing CEL measurement within the boundaries of theory and concise conceptualisation, the assumptions and limitations of existing measures should become more transparent.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their helpful suggestions and feedback. This research work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No 860721.

REFERENCES

[1] Brian Amento, Will Hill, Deborah Hix, Robert Schulman, and Loren Terveen. 2003. Experiments in Social Data Mining. *ACM Transactions on Computer-Human Interaction* 10, 1 (2003), 54–85. https://doi.org/10.1145/606658.606661

[2] Ioannis Arapakis, Luis A. Leiva, and B. Barla Cambazoglu. 2015. Know your onions: Understanding the user experience with the knowledge module in web search. *International Conference on Information and Knowledge Management, Proceedings* 19-23-Oct- (2015), 1695–1698. https://doi.org/10.1145/2806416.2806591

[3] Jaime Arguello and Bogeum Choi. 2019. The effects of working memory, perceptual speed, and inhibition in aggregated search. *ACM Transactions on Information Systems* 37, 3 (2019). https://doi.org/10.1145/3322128

[4] Jaime Arguello and Anita Crescenzi. 2019. Using Principal Component Analysis to Better Understand Behavioral Measures and Their Effects. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '19)*. Association for Computing Machinery, New York, NY, USA, 177–184. https://doi.org/10.1145/3341981.3344222

[5] Ferran Ariza, Dipak Kalra, and Henry W.W. Potts. 2015. 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* 22, 4 (2015), 379–390. https://doi.org/10.14236/jhi.v22i4.85

[6] Sandeep Avula, Jaime Arguello, Robert Capra, Jordan Dodson, Yuhui Huang, and Filip Radlinski. 2019. Embedding search into a conversational platform to support collaborative search. *CHIIR 2019 - Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (2019), 15–23. https://doi.org/10.1145/3295750.3298928

[7] Leif Azzopardi. 2011. The economics in interactive information retrieval. *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2011), 15–24. https://doi.org/10.1145/2009916.2009923

[8] Leif Azzopardi, Diane Kelly, and Kathy Brennan. 2013. How Query Cost Affects Search Behavior Categories and Subject Descriptors. *Sigir* (2013), 23–32. https://doi.org/10.1145/2484028.2484049

[9] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: An information foraging based measure. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018* (2018), 605–614. https://doi.org/10.1145/3209978.3210027

[10] Leif Azzopardi and Guido Zuccon. 2016. An Analysis of the Cost and Benefit of Search Interactions. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/2970398.2970412

[11] Earl Bailey and Diane Kelly. 2011. Is amount of effort a better predictor of search success than use of specific search tactics? *Proceedings of the ASIST Annual Meeting* 48 (2011). https://doi.org/10.1002/meet.2011.14504801077

[12] Maarten A.S. Boksem and Mattie Tops. 2008. Mental fatigue: Costs and benefits. *Brain Research Reviews* 59, 1 (2008), 125–139. https://doi.org/10.1016/j.brainresrev.2008.07.001

[13] Horatiu Bota, Ke Zhou, and Joemon M Jose. 2016. 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 (CHIIR '16)*. Association for Computing Machinery, New York, NY, USA, 131–140. https://doi.org/10.1145/2854946.2854967

[14] Kathy Brennan, Diane Kelly, and Jaime Arguello. 2014. 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 (IIiX '14)*. Association for Computing Machinery, New York, NY, USA, 165–174. https://doi.org/10.1145/2637002.2637022

[15] Roland Brünken, Jan L. Plass, and Detlev Leutner. 2003. Direct measurement of cognitive load in multimedia learning. *Educational Psychologist* 38, 1 (2003), 53–61. https://doi.org/10.1207/S15326985EP3801{_}7

[16] Jake Brutlag. 2009. Speed matters for google web search. *Google. June* (2009), 2009.

[17] Robert Capra, Jaime Arguello, Heather O'Brien, Yuan Li, and Bogeum Choi. 2018. The effects of manipulating task determinability on search behaviors and outcomes. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018* (2018), 445–454. https://doi.org/10.1145/3209978.3210047

[18] Rob Capra, Jaime Arguello, and Yinglong Zhang. 2017. The Effects of Search Task Determinability on Search Behaviour. In *ECIR 2017*. 108–121. https://doi.org/10.1007/978-3-319-56608-5

[19] Robert Capra, Gary Marchionini, Jung Sun Oh, Fred Stutzman, and Yan Zhang. 2007. Effects of structure and interaction style on distinct search tasks. *Proceedings of the ACM International Conference on Digital Libraries* May 2014 (2007), 442–451. https://doi.org/10.1145/1255175.1255267

[20] Rebecca L. Charles and Jim Nixon. 2019. Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics* 74, September 2016 (2019), 221–232. https://doi.org/10.1016/j.apergo.2018.08.028

[21] Hsinchun Chen, Haiyan Fan, Michael Chau, and Daniel Zeng. 2003. Testing a Cancer Meta Spider. *International Journal of Human Computer Studies* 59, 5 (2003), 755–776. https://doi.org/10.1016/S1071-5819(03)00118-6

[22] Ye Chen, Yiqun Liu, Ke Zhou, Meng Wang, Min Zhang, and Shaoping Ma. 2015. Does vertical bring more satisfaction? Predicting search satisfaction in a heterogeneous environment. *International Conference on Information and Knowledge Management, Proceedings* 19-23-Oct- (2015), 1581–1590. https://doi.org/10.1145/2806416.2806473

[23] Aline Chevalier and Maud Kicka. 2006. Web designers and web users: Influence of the ergonomic quality of the web site on the information search. *International Journal of Human Computer Studies* 64, 10 (2006), 1031–1048. https://doi.org/10.1016/j.ijhcs.2006.06.002

[24] Bogeum Choi, Robert Capra, and Jaime Arguello. 2019. The Effects of Working Memory during Search Tasks of Varying Complexity. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 261–265. https://doi.org/10.1145/3295750.3298948

[25] Hwan Hee Choi, Jeroen J.G. van Merriënboer, and Fred Paas. 2014. Effects of the Physical Environment on Cognitive Load and Learning: Towards a New Model of Cognitive Load. *Educational Psychology Review* 26, 2 (2014), 225–244. https://doi.org/10.1007/s10648-014-9262-6

[26] M. Claypool, P. Le, M. Wased, and D. Brown. 2001. Implicit interest indicators. *International Conference on Intelligent User Interfaces, Proceedings IUI* (2001), 33–40. https://doi.org/10.1145/359784.359836

[27] Michael J. Cole, Jacek Gwizdka, Chang Liu, and Nicholas J. Belkin. 2011. Dynamic assessment of information acquisition effort during interactive search. *Proceedings of the ASIST Annual Meeting* 48 (2011). https://doi.org/10.1002/meet.2011.14504801149

[28] Michael J. Cole, Jacek Gwizdka, Chang Liu, Nicholas J. Belkin, and Xiangmin Zhang. 2013. Inferring user knowledge level from eye movement patterns. *Information Processing and Management* 49, 5 (2013), 1075–1091. https://doi.org/10.1016/j.ipm.2012.08.004

[29] Michael D. Cooper. 1972. A cost model for evaluating information retrieval systems. *Journal of the American Society for Information Science* 23, 5 (1972), 306–312. https://doi.org/10.1002/asi.4630230505

[30] Ton de Jong. 2010. Cognitive load theory, educational research, and instructional design: Some food for thought. In *Instructional Science*, Vol. 38. 105–134. https://doi.org/10.1007/s11251-009-9110-0

[31] J. C.F. de Winter. 2014. Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology and Work* 16, 3 (2014), 289–297. https://doi.org/10.1007/s10111-014-0275-1

[32] Alan R. Dennis and Nolan J. Taylor. 2006. Information foraging on the web: The effects of "acceptable" Internet delays on multi-page information search behavior. *Decision Support Systems* 42, 2 (2006), 810–824. https://doi.org/10.1016/j.dss.2005.05.032

[33] Simon Dennis, Peter Bruza, and Robert McArthur. 2002. Web searching: A process-oriented experimental study of three interactive search paradigms. *Journal of the American Society for Information Science and Technology* 53, 2 (2002), 120–133. https://doi.org/10.1002/asi.10015

[34] Leandro L. Di Stasi, Adoración Antolí, Miguel Gea, and José J. Cañas. 2011. A neuroergonomic approach to evaluating mental workload in hypermedia interactions. , 298–304 pages. https://doi.org/10.1016/j.ergon.2011.02.008

[35] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. 2020. Interactive Evaluation of Conversational Agents: Reflections on the Impact of Search Task Design. *ICTIR 2020 - Proceedings of the 2020 ACM SIGIR International Conference on Theory of Information Retrieval* (2020), 85–88. https://doi.org/10.1145/3409256.3409814

[36] Crudge Sarah E. and Johnson Frances C. 2007. Using the repertory grid and laddering technique to determine the user's evaluative model of search engines. *Journal of Documentation* 63, 2 (1 2007), 259–280. https://doi.org/10.1108/00220410710737213

[37] Ashlee Edwards, Diane Kelly, and Leif Azzopardi. 2015. The impact of query interface design on stress, workload and performance. In *European Conference of Information Retrieval (ECIR)*, Vol. 9022. 691–702. https://doi.org/10.1007/978-3-319-16354-3{_}76

[38] Howard Egeth and Daniel Kahneman. 1975. Attention and Effort. *The American Journal of Psychology* 88, 2 (1975), 339. https://doi.org/10.2307/1421603

[39] David F. Feldon, Gregory Callan, Stephanie Juth, and Soojeong Jeong. 2019. Cognitive Load as Motivational Cost. *Educational Psychology Review* 31, 2 (2019), 319–337. https://doi.org/10.1007/s10648-019-09464-6

[40] Carol Hansen Fenichel. 1981. Online searching: Measures that discriminate among users with different types of experiences. *Journal of the American Society for Information Science* 32, 1 (1981), 23–32. https://doi.org/10.1002/asi.4630320104

[41] T. Flynn, P. A. Holohan, M. S. Magson, and J. D. Munro. 1979. Cost effectiveness comparison of online and manual bibliographic information retrieval. *Journal of Information Science* 1, 2 (1979), 77–84. https://doi.org/10.1177/016555157900100204

[42] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating Implicit Measures to Improve Web Search. *ACM Trans. Inf. Syst.* 23, 2 (4 2005), 147–168. https://doi.org/10.1145/1059981.1059982

[43] Maria Gäde, Marijn Koolen, Mark Hall, Toine Bogers, and Vivien Petras. 2021. A Manifesto on Resource Re-Use in Interactive Information Retrieval. *CHIIR 2021 - Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* March (2021), 141–149. https://doi.org/10.1145/3406522.3446056

[44] Dennis Galletta, Raymond Henry, Scott McCoy, and Peter Polak. 2004. Web Site Delays: How Tolerant are Users? *Journal of the Association for Information Systems* 5, 1 (2004), 1–28. https://doi.org/10.17705/1jais.00044

[45] Paul Gerwe and Charles L Viles. 2000. User Effort in Query Construction and Interface Selection. In *Proceedings of the Fifth ACM Conference on Digital Libraries (DL '00)*. Association for Computing Machinery, New York, NY, USA, 246–247. https://doi.org/10.1145/336597.336679

[46] Roberto González-Ibáñez, José Luis Varela-Otárola, and Carlos Barrera-Pulgar. 2016. Evaluating Body-Centered Interactions in an Image Search Task. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)*. Association for Computing Machinery, New York, NY, USA, 269–272. https://doi.org/10.1145/2854946.2854998

[47] Roberto González-IbáñTez, Verónica Proaño-Ríos, Gary Fuenzalida, and Gonzalo Martinez-Ramirez. 2017. Effects of a visual representation of search engine results on performance, user experience and effort. *Proceedings of the Association for Information Science and Technology* 54, 1 (2017), 128–138. https://doi.org/10.1002/pra2.2017.14505401015

[48] Maria J. Grant and Andrew Booth. 2009. A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal* 26, 2 (2009), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

[49] Qi Guo and Eugene Agichtein. 2012. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web* (2012), 569–578. https://doi.org/10.1145/2187836.2187914

[50] Jacek Gwizdka. 2008. Revisiting search task difficulty: Behavioral and individual difference measures. *Proceedings of the ASIST Annual Meeting* 45 (2008). https://doi.org/10.1002/meet.2008.1450450249

[51] Jacek Gwizdka. 2010. Assessing Cognitive Load on Web Search Tasks. *The Ergonomics Open Journal* 2, 2 (2010), 114–123. https://doi.org/10.2174/1875934300902020114

[52] Jacek Gwizdka. 2010. Distribution of cognitive load in Web search. *Journal of the American Society for Information Science and Technology* 61, 11 (2010), 2167–2187. https://doi.org/10.1002/asi.21385

[53] Jacek Gwizdka. 2013. Effects of working memory capacity on users' search effort. *ACM International Conference Proceeding Series* (2013). https://doi.org/10.1145/2500342.2500358

[54] Jacek Gwizdka. 2014. Characterizing Relevance with Eye-Tracking Measures. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX '14)*. Association for Computing Machinery, New York, NY, USA, 58–67. https://doi.org/10.1145/2637002.2637011

[55] Jacek Gwizdka. 2017. 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 (CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 341–344. https://doi.org/10.1145/3020165.3022148

[56] J Gwizdka and MJ Cole. 2011. Least effort? Not if I can search more. *In Proceedings of the 5th Workshop on Human-Computer Interaction and Information Retrieval*. 2, L (2011), 2012. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.306.8297&rep=rep1&type=pdf

[57] Jacek Gwizdka and Irene Lopatovska. 2009. The role of subjective factors in the information search process. *Journal of the American Society for Information Science and Technology* 60, 12 (2009), 2452–2464. https://doi.org/10.1002/asi.21183

[58] Jacek Gwizdka and Ian Spence. 2006. What can searching behavior tell us about the difficulty of information tasks? A study of web navigation. *Proceedings of the ASIST Annual Meeting* 43 (2006). https://doi.org/10.1002/meet.14504301167

[59] Martin Halvey and Robert Villa. 2014. Evaluating the effort involved in relevance assessments for images. (2014), 887–890. https://doi.org/10.1145/2600428.2609466

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

[61] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[62] Jiyin He, Marc Bron, Arjen De Vries, Leif Azzopardi, and Maarten De Rijke. 2015. Untangling result list refinement and ranking quality: A framework for evaluation and prediction. *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* i (2015), 293–302. https://doi.org/10.1145/2766462.2767740

[63] Roberta Heale and Alison Twycross. 2015. Validity and reliability in quantitative studies. *Evidence-Based Nursing* 18, 3 (2015), 66–67. https://doi.org/10.1136/eb-2015-102129

[64] Michael Inzlicht, Amitai Shenhav, and Christopher Y. Olivola. 2018. The Effort Paradox: Effort Is Both Costly and Valued. *Trends in Cognitive Sciences* 22, 4 (2018), 337–349. https://doi.org/10.1016/j.tics.2018.01.007

[65] Bernard J. Jansen, Danielle Booth, and Brian Smith. 2009. Using the taxonomy of cognitive learning to model online searching. *Information Processing and Management* 45, 6 (2009), 643–663. https://doi.org/10.1016/j.ipm.2009.05.004

[66] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and Predicting Graded Search Satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. Association for Computing Machinery, New York, NY, USA, 57–66. https://doi.org/10.1145/2684822.2685319

[67] Jiepu Jiang, Daqing He, and James Allan. 2014. 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* (2014), 607–616. https://doi.org/10.1145/2600428.2609633

[68] Jiepu Jiang, Daqing He, and James Allan. 2017. Comparing in situ and multidimensional relevance judgments. *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017), 405–414. https://doi.org/10.1145/3077136.3080840

[69] Jiepu Jiang, Daqing He, Diane Kelly, and James Allan. 2017. Understanding ephemeral state of relevance. *CHIIR 2017 - Proceedings of the 2017 Conference Human Information Interaction and Retrieval* (2017), 137–146. https://doi.org/10.1145/3020165.3020176

[70] Angel Jimenez-Molina, Cristian Retamal, and Hernan Lira. 2018. Using psychophysiological sensors to assess mental workload during web browsing. https://doi.org/10.3390/s18020458

[71] Maryam Kamvar and Shumeet Baluja. 2008. Query Suggestions for Mobile Search: Understanding Usage Patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1013–1016. https://doi.org/10.1145/1357054.1357210

[72] Weimao Ke, Cassidy R. Sugimoto, and Javed Mostafa. 2009. Dynamicity vs. effectiveness: Studying online clustering for scatter/gather. *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009* (2009), 19–26. https:

//doi.org/10.1145/1571941.1571947

[73] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1-2 (2009), 1–224. https://doi.org/10.1561/1500000012

[74] Diane Kelly and Leif Azzopardi. 2015. How many results per page? (2015), 183–192. https://doi.org/10.1145/2766462.2767732

[75] Diane Kelly and Nicholas J. Belkin. 2004. Display time as implicit feedback: Understanding task effects. *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2004), 377–384.

[76] Diane Kelly and Cassidy R. Sugimoto. 2013. A systematic review of interactive information retrieval evaluation studies, 1967-2006. *Journal of the American Society for Information Science and Technology* 64, 4 (2013), 745–770. https://doi.org/10.1002/asi.22799

[77] Yong-Mi Kim and Soo Young Rieh. 2006. Dual-task performance as a measure of mental effort in searching a library system and the Web. *Proceedings of the American Society for Information Science and Technology* 42, 1 (2006), n/a–n/a. https://doi.org/10.1002/meet.14504201155

[78] Paul A. Kirschner. 2002. Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and Instruction* 12, 1 (2002), 1–10. https://doi.org/10.1016/S0959-4752(01)00014-7

[79] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)*. Association for Computing Machinery, New York, NY, USA, 121–130. https://doi.org/10.1145/2854946.2854961

[80] Aniket Kittur, Andrew M Peters, Abdigani Diriye, Trupti Telang, and Michael R Bove. 2013. Costs and Benefits of Structured Information Foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2989–2998. https://doi.org/10.1145/2470654.2481415

[81] Melina Klepsch, Florian Schmitz, and Tina Seufert. 2017. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology* 8, NOV (2017). https://doi.org/10.3389/fpsyg.2017.01997

[82] Marina Krnic Martinic, Dawid Pieper, Angelina Glatt, and Livia Puljak. 2019. Definition of a systematic review used in overviews of systematic reviews, meta-epidemiological studies and textbooks. *BMC Medical Research Methodology* 19, 1 (2019), 1–12. https://doi.org/10.1186/s12874-019-0855-0

[83] Cody C.T. Kwok, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. *Proceedings of the 10th International Conference on World Wide Web, WWW 2001* (2001), 150–161. https://doi.org/10.1145/371920.371973

[84] Steven Lastres, A. 2012. Legal Research Comes of Age. (2012). http://www.lexisnexis.com/documents/pdf/20130806061418_large.pdf?sf17880724=1

[85] Jimmie Leppink, Fred Paas, Cees P.M. Van der Vleuten, Tamara Van Gog, and Jeroen J.G. Van Merriënboer. 2013. Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods* 45, 4 (2013), 1058–1072. https://doi.org/10.3758/s13428-013-0334-1

[86] Yuelin Li and Nicholas J Belkin. 2010. An exploration of the relationships between work task and interactive information search behavior. *Journal of the American Society for Information Science and Technology* 61, 9 (2010), 1771–1789. https://econpapers.repec.org/RePEc:bla:jamist:v:61:y:2010:i:9:p:1771-1789

[87] Qiaoling Liu, Yandong Liu, and Eugene Agichtein. 2010. Exploring Web Browsing Context for Collaborative Question Answering. In *Proceedings of the Third Symposium on Information Interaction in Context (IIiX '10)*. Association for Computing Machinery, New York, NY, USA, 305–310. https://doi.org/10.1145/1840784.1840830

[88] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different users, different opinions: Predicting search Satisfaction with mouse movement information. *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2015), 493–502. https://doi.org/10.1145/2766462.2767721

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

[90] Cheng Luo, Xue Li, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, and Shaoping Ma. 2017. Investigating Users' Time Perception during Web Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 127–136. https://doi.org/10.1145/3020165.3020184

[91] Cheng Luo, Yiqun Liu, Tetsuya Sakai, Ke Zhou, Fan Zhang, Xue Li, and Shaoping Ma. 2017. Does Document Relevance Affect the Searcher's Perception of Time?. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. Association for Computing Machinery, New York, NY, USA, 141–150. https://doi.org/10.1145/3018661.3018694

[92] Horia A. Maior, Matthew Pike, Max L. Wilson, and Sarah Sharples. 2013. Directly evaluating the cognitive impact of search user interfaces: A two-pronged approach with fNIRS. *Euro HCIR Workshop Proceedings* 1033 (2013), 43–46.

[93] Stephann Makri, Ann Blandford, and Anna L. Cox. 2008. Investigating the information-seeking behaviour of academic lawyers: From Ellis's model to design. *Information Processing and Management* 44, 2 (2008), 613–634. https://doi.org/10.1016/j.ipm.2007.05.001

[94] Yazdan Mansourian and Nigel Ford. 2007. Search persistence and failure on the web: A "bounded rationality" and "satisficing" analysis. *Journal of Documentation* 63, 5 (2007), 680–701. https://doi.org/10.1108/00220410710827754

[95] Gerald Matthews, Joost De Winter, and P. A. Hancock. 2020. What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science* 21, 4 (2020), 369–396. https://doi.org/10.1080/1463922X.2018.1547459

[96] David Maxwell and Leif Azzopardi. 2014. Stuck in traffic: How temporal delays affect search behaviour. *Proceedings of the 5th Information Interaction in Context Symposium, IIiX 2014* (2014), 155–164. https://doi.org/10.1145/2637002.2637021

[97] Molly Mcgregor, Leif A. Azzopardi, and Martin Halvey. 2021. Untangling Cost, Effort, and Load in Information Seeking and Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 151–161.

[98] Anna Mikkonen. 2015. Books ' Interest Grading and Fiction Readers ' Search Actions During Query Reformulation Intervals Categories and Subject Descriptors. (2015), 27–36.

[99] G Miller. 1956. The magical number seven plus minus two. *Psych. Rev.* 63 (1956), 81–97.

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

[101] Prithima Reddy Mosaly, Lukasz Mazur, and Lawrence B. Marks. 2016. 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* (2016), 313–316. https://doi.org/10.1145/2854946.2854985

[102] Cynthia Mulrow, Deborah Cook, Maureen O Meade, and W Scott Richardson. 1997. Systematic Review Series Series Editors: Selecting and Appraising Studies for a Systematic Review Selecting Studies for Systematic Reviews. *Annals of Internal Medicine* 127 (1997), 531–537.

[103] Fiona Fui Hoon Nah. 2004. A study on tolerable waiting time: How long are Web users willing to wait? *Behaviour and Information Technology* 23, 3 (2004), 153–163. https://doi.org/10.1080/01449290410001669914

[104] Ragnar NordlieOslo and Nils Pharo. 2013. Search transition as a measure of effort in information retrieval interaction. *Proceedings of the ASIST Annual Meeting* 50, 1 (2013), 1–7. https://doi.org/10.1002/meet.14505001044

[105] Denise O'Connor, Sally Elizabeth Green, and Julian P T Higgins. 2008. Defining the review question and developing criteria for including studies. In *Cochrane Handbook for Systematic Reviews of Interventions* (first ed.), Julian PT Higgins and Sally Green (Eds.). John Wiley & Sons, United States of America, 8–94.

[106] Suvi Oksanen and Pertti Vakkari. 2012. 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 (JCDL '12)*. Association for Computing Machinery, New York, NY, USA, 199–202. https://doi.org/10.1145/2232817.2232855

[107] Kevin Ong, Mark Sanderson, Kalervo Järvelin, and Falk Scholer. 2018. QWERTY: The effects of typing on web search behavior. *CHIIR 2018 - Proceedings of the 2018 Conference on Human Information Interaction and Retrieval* 2018-March (2018), 281–284. https://doi.org/10.1145/3176349.3176872

[108] Longo L. Orru G. 2019. Human Mental Workload: Models and Applications. Communications in Computer and Information Science. *Communications in Computer and Information Science* 1012, February (2019), 267. https://doi.org/10.1007/978-3-030-14273-5

[109] A. Ross Otto and Nathaniel D. Daw. 2019. The opportunity cost of time modulates cognitive effort. *Neuropsychologia* 123, May 2018 (2019), 92–105. https://doi.org/10.1016/j.neuropsychologia.2018.05.006

[110] Fred Paas, Juhani E. Tuovinen, Huib Tabbers, and Pascal W.M. Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist* 38, 1 (2003), 63–71. https://doi.org/10.1207/S15326985EP3801{_}8

[111] F. G. Paas, J. J. Van Merriënboer, and J. J. Adam. 1994. Measurement of cognitive load in instructional research. *Perceptual and Motor Skills* 79, 1 Pt 2 (1994), 419–430. https://doi.org/10.2466/pms.1994.79.1.419

[112] Suppanut Pothirattanachaikul, Yusuke Yamamoto, Takehiro Yamamoto, and Masatoshi Yoshikawa. 2019. Analyzing the effects of document's opinion and credibility on search behaviors and belief dynamics. *International Conference on Information and Knowledge Management, Proceedings* (2019), 1653–1662. https://doi.org/10.1145/3357384.3357886

[113] Paul Price, Rajiv Jhangiani, and I-Chant Chiang. 2015. Research Methods in Psychology. In *Research Methods in Psychology* (2nd ed.). Pressbooks.com, 322. https://doi.org/10.1016/0022-3999(95)00555-2

[114] Manasa Rath, Souvick Ghosh, and Chirag Shah. 2018. Exploring Online and Offline Search Behavior Based on the Varying Task Complexity. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 285–288. https://doi.org/10.1145/3176349.3176890

[115] Soo Young Rieh, Yong Mi Kim, and Karen Markey. 2012. Amount of invested mental effort (AIME) in online searching. *Information Processing and Management* 48, 6 (2012), 1136–1150. https://doi.org/10.1016/j.ipm.2012.05.001

[116] Peter Schmutz, Silvia Heinz, Yolanda Métrailler, and Klaus Opwis. 2009. Cognitive Load in eCommerce Applications—Measurement and Effects on User Satisfaction. *Advances in Human-Computer Interaction* 2009 (2009), 1–9. https://doi.org/10.1155/2009/121494

[117] Chirag Shah and Roberto González-Ibáñez. 2011. Evaluating the synergic effect of collaboration in information seeking. *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* January (2011), 913–922. https://doi.org/10.1145/2009916.2010038

[118] Amitai Shenhav, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L. Griffiths, Jonathan D. Cohen, and Matthew M. Botvinick. 2017. Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience* 40, December 2016 (2017), 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

[119] Georg Singer, Ulrich Norbisrath, and Dirk Lewandowski. 2012. Ordinary search engine users assessing difficulty, effort, and outcome for simple and complex search tasks. *IIiX 2012 - Proceedings 4th Information Interaction in Context Symposium: Behaviors, Interactions, Interfaces, Systems* (2012), 110–119. https://doi.org/10.1145/2362724.2362746

[120] Cheri Speier and Micheal G Morris. 2003. The Influence of Query Interface Design on Decision-Making Performance. *MIS Quarterly* 27, 3 (2003), 397–423. https://www.jstor.org/stable/30036539

[121] John Sweller. 1988. Cognitive Load During Problem Solving: Effects on Learning - Sweller - 2010 - Cognitive Science - Wiley Online Library. *Cognitive science* 285 (1988), 257–285. http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog1202_4/abstract

[122] John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction* 4, 4 (1994), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

[123] John Sweller. 2018. Measuring cognitive load. *Perspectives on Medical Education* 7, 1 (2018). https://doi.org/10.1007/s40037-017-0395-4

[124] Pertti Vakkari and Saila Huuskonen. 2012. Search effort degrades search output but improves task outcome. *Journal of the American Society for Information Science and Technology* 63, 4 (2012), 657–670. https://econpapers.repec.org/RePEc:bla:jamist:v:63:y:2012:i:4:p:657-670

[125] Bram B. Van Acker, Davy D. Parmentier, Peter Vlerick, and Jelle Saldien. 2018. Understanding mental workload: from a clarifying concept analysis toward an implementable framework. *Cognition, Technology and Work* 20, 3 (2018), 351–365. https://doi.org/10.1007/s10111-018-0481-3

[126] Jeroen J.G. Van Merriënboer and John Sweller. 2005. *Cognitive load theory and complex learning: Recent developments and future directions.* Vol. 17. 147–177 pages. https://doi.org/10.1007/s10648-005-3951-0

[127] Marc van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law* 25, 1 (2017), 65–87. https://doi.org/10.1007/s10506-017-9195-8

[128] Matt Vassar, Paul Atakpo, and Melissa J. Kash. 2016. Manual search approaches used by systematic reviewers in dermatology. *Journal of the Medical Library Association* 104, 4 (2016), 302–304. https://doi.org/10.3163/1536-5050.104.4.009

[129] Manisha Verma and Emine Yilmaz. 2017. Search costs vs. User satisfaction on mobile. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10193 LNCS (2017), 698–704. https://doi.org/10.1007/978-3-319-56608-5{_}68

[130] Robert Villa and Martin Halvey. 2013. Is relevance hard work? Evaluating the effort of making relevant assessments. *SIGIR 2013 - Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2013), 765–768. https://doi.org/10.1145/2484028.2484150

[131] Robert Villa and Joemon M. Jose. 2012. A study of awareness in multimedia search. *Information Processing and Management* 48, 1 (2012), 32–46. https://doi.org/10.1016/j.ipm.2011.03.005

[132] Tung Vuong, Miamaria Saastamoinen, Giulio Jacucci, and Tuukka Ruotsalo. 2019. Understanding user behavior in naturalistic information search tasks. *Journal of the Association for Information Science and Technology* 70, 11 (2019), 1248–1261. https://doi.org/10.1002/asi.24201

[133] Qiuzhen Wang, Sa Yang, Manlu Liu, Zike Cao, and Qingguo Ma. 2014. An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems* 62 (2014), 1–10. https://doi.org/10.1016/j.dss.2014.02.007

[134] Austin R. Ward and Rob Capra. 2020. Immersive Search: Using Virtual Reality to Examine How a Third Dimension Impacts the Searching Process. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), 1621–1624. https://doi.org/10.1145/3397271.3401303

[135] Andrew Westbrook and Todd S. Braver. 2015. Cognitive effort: A neuroeconomic approach. *Cognitive, Affective and Behavioral Neuroscience* 15, 2 (2015), 395–415. https://doi.org/10.3758/s13415-015-0334-y

[136] Andrew Westbrook, Daria Kester, and Todd S. Braver. 2013. What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference. *PLoS ONE* 8, 7 (2013), 1–8. https://doi.org/10.1371/journal.pone.0068210

[137] Christopher D Wickens. 2002. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science* 3, 2 (2002), 159–177. https://doi.org/10.1080/14639220210123806

[138] Max L. Wilson. 2011. Evaluating the cognitive impact of search user interface design decisions. *Euro HCIR Workshop Proceedings* 763 (2011), 27–30.

[139] I. Chin Wu and Pertti Vakkari. 2014. Supporting navigation in Wikipedia by information visualization: Extended evaluation measures. *Journal of Documentation* 70, 3 (2014), 392–424. https://doi.org/10.1108/JD-10-2012-0138

[140] Beverly Yang and Glen Jeh. 2006. Retroactive answering of search queries. *Proceedings of the 15th International Conference on World Wide Web* 1 (2006), 457–466. https://doi.org/10.1145/1135777.1135845

[141] Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. 2014. Relevance and effort: An analysis of document utility. *CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management* (2014), 91–100. https://doi.org/10.1145/2661829.2661953

[142] Mark S. Young, Karel A. Brookhuis, Christopher D. Wickens, and Peter A. Hancock. 2015. State of science: mental workload in ergonomics. *Ergonomics* 58, 1 (2015), 1–17. https://doi.org/10.1080/00140139.2014.956151

[143] Yinglong Zhang and Jacek Gwizdka. 2014. Effects of tasks at similar and different complexity levels. *Proceedings of the ASIST Annual Meeting* 51, 1 (2014). https://doi.org/10.1002/meet.2014.14505101093

[144] Yinglong Zhang and Jacek Gwizdka. 2016. Rethinking the cost of information search behavior. *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2016), 969–972. https://doi.org/10.1145/2911451.2914742

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