Untangling Cost, Effort, and Load in Information Seeking and Retrieval

Molly McGregor
molly.mcgregor@strath.ac.uk
University of Strathclyde
Glasgow, UK

Leif Azzopardi
leifos@acm.org
University of Strathclyde
Glasgow, UK

Martin Halvey
martin.halvey@strath.ac.uk
University of Strathclyde
Glasgow, UK

ABSTRACT

When performing *Information Seeking and Retrieval* (ISR) activities, people submit queries, examine results, assess documents and engage with the information to make decisions and complete tasks. All these activities come at a "cost", but within the field of ISR there is no universally accepted definition of the concepts of *Cost, Effort and Load* (CEL). Instead, researchers have used the same terms interchangeably to describe similar but also different concepts. This lack of shared understanding has led to a disconnect between how these concepts are defined and discussed versus how they are interpreted and measured. Thus, the aim of this paper is two-fold: (*i*) to review the meaning of CEL related concepts used within ISR, and (*ii*) to create a shared taxonomy of the concepts relating to CEL in ISR. To seed our analysis, we conducted a literature review, where 397 papers were reviewed, and twenty-six papers that explicitly proposed measures or definitions of CEL were selected for analysis. By drawing upon theory from Psychology and other fields, we present the common definitions of CEL in order to ground our discussion of these concepts in ISR. We also highlight the issues associated with CEL measurement in ISR to help researchers reflect on the validity and precision of existing methods. We hope this perspectives paper serves as a basis for a taxonomy of how CEL concepts are used within ISR- where we have provided a series of working definitions that clearly delineate the different concepts being used, investigated and measured in ISR research.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**.

KEYWORDS

Cost; Effort; Cognitive Load; Workload

ACM Reference Format:
Molly McGregor, Leif Azzopardi, and Martin Halvey. 2021. Untangling Cost, Effort, and Load in Information Seeking and Retrieval. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21), March 14–19, 2021, Canberra, ACT, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3406522.3446026

CHIIR '21, March 14–19, 2021, Canberra, ACT, Australia
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8055-3/21/03…$15.00
https://doi.org/10.1145/3406522.3446026

1 INTRODUCTION

The nature of *Information Seeking, and Retrieval* (ISR) involves the user interacting with the system in a variety of ways, such as submitting queries, examining results, and engaging with information to make a decision [3]. It is well recognised that the search task and the system can affect the "load" on the user's cognitive resources during the search process [19], and that performing these search interactions requires "effort" which comes at a "cost" to the user [3]. Alongside measures of recall and precision [45], measures of the user's effort and cost have been incorporated within *Information Retrieval* (IR) evaluation frameworks since the 1960's. For example, in 1968, Cooper [9] introduced the *Expected Search Length* (ESL) metric to measure the expected user effort in terms of the average number of documents a user must browse in order to retrieve a given number of relevant documents. Under ESL, it assumed that the "effort" the user expends or the "load" they are placed under is directly proportional to the number of documents examined.

Cost, Effort, and Load (CEL), and their related concepts, are important and salient factors considered in most ISR studies. To date, CEL has been measured in a variety of different contexts and conditions, e.g. search engine result pages [6, 22]; search systems[2, 4, 27]; and interactive search tasks [21, 49, 56]. While these studies highlight the value of CEL measurement in ISR, it seems surprising that our literature review found less than thirty papers which explicitly define or propose methods to measure CEL. From these papers, ambiguities such as using CEL terms interchangeably within a single study and the existence of clear conceptual overlaps between CEL definitions are apparent.

So, despite over 50 years of CEL related discussions, and uses of, within the field of ISR, there appears to exist no universally accepted definitions of CEL and their related concepts. Thus, the aim of this paper is to untangle how CEL concepts have been used within ISR studies, and also to assess how these CEL concepts have subsequently been measured. To this end, we provide a review of contemporary ISR literature to ground our discussion and guide our future directions. In this perspectives paper, the contributions to the ISR community are as follows: Firstly, we aim to raise ISR researcher's awareness of the importance of providing clear and precise definitions of cost, effort, and load, and their related concepts. Secondly, we aim to draw ISR researcher's attention to the current methodological issues associated with CEL measurement and provide recommendations for improving the sensitivity, validity, and reliability of these methods. Finally, we provide working definitions of cost, effort, and load concepts with the aim of initiating the establishment of a shared framework of terminology and concepts.

2 BACKGROUND THEORY AND CONCEPTS

This section aims to provide an overview of the key concepts and theories surrounding CEL. Specifically, we present accepted CEL definitions from Psychology and related fields as a way to ground our discussions about CEL in ISR.

2.1 Defining Cost, Effort, and Load

Research methods from Psychology suggest that in order to establish an appropriate method, it is important that the measurable concept is sufficiently defined relative to its nominal and operational meaning [42]. Nominal definitions describe the meaning of the concept, whereas operational definitions explain precisely how the concept and its elements will be measured [25]. As nominal definitions preface the operational, significant problems can arise when ambiguous and vague nominal conceptualisations 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 concept are to be measured, then it is likely that a variety of different metrics and methods will emerge from these properties and later be used by researchers to measure what they believe to be the "same" concept. In reality, the lack of precision in the conceptual definitions provided at the outset will likely lead to conceptual overlap, unclear measures, and a loss of causality.

For over 50 years, CEL concepts have been discussed and explored extensively across disciplines such as Psychology [40, 51, 59], Ergonomics [7, 60], and Human Factors [11, 23]. Yet, universal definitions of these concepts are still yet to transpire. While it is beyond the scope of this current paper to review the last 50 years of CEL research, the next section will briefly present the most commonly accepted definitions of CEL proposed by disciplines out-with ISR.

2.2 Cognitive Load and Mental Workload

While cognitive load and workload evolved independently from within different disciplines, both concepts are theoretically underpinned by the same core assumptions [38]. *Cognitive Load Theory* (CLT) [50] and *Multiple Resource Theory* (MRT) [59] which emerge from the field of Educational Psychology, and Ergonomics and Human Factors, respectively, can be considered the leading theories to describe both cognitive load and workload concepts. Both theories are similar in that they are closely related by their assumption of limited mental capacity and competing task demands [50, 59]. 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 [24]. Sweller [49] 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 [53]. In CLT, cognitive load is defined as the total amount of mental activity imposed on the working memory at any given moment [51]. 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 [34]. 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 [50]. 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) [10]. These three types of cognitive load are proposed to be additive [41].

Mental workload is perhaps one of the most popular concepts examined in Ergonomics and Human Factors research, however researchers in the field are still yet to reach a universal consensus regarding its definition [60]. 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 [7]. MRT asserts that the human brain has a fixed quantity of mental resources of various types [52]. 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 [59]. 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 [52]. 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. [7, 52]. Van Acker et al. [52] describes 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.

2.3 Effort

As with the conceptualisation of load, the concept of effort has faced similar ambiguity in relation to its origins and characterisations [57]. In the field of Psychology, effort is most often implicated as a mediating behaviour in numerous theories. For example, effort has been considered as a mediating variable in strategy selection - where effort is proposed to influence whether an individual will choose a high performance strategy (high effort) vs. a low performance strategy (low effort) [58]. Similarly, behavioural psychologist Clark Hull, offered the "law of least work" which relies on the notion that if presented with two options of similar reward, an individual is predisposed to avoiding options which require expending more work or effort [36]. From this perspective, effort is the mediator between the potential performance of an individual on a task versus how well they actually perform [47]. For example, an individual may have the ability to solve mathematical equations, but fail to solve a simple mathematical equation due to their unwillingness to exert effort.

Despite the general interest, the lack of direct examination has left the concept of effort without a universal, operational definition within the psychological domain [60]. To add to this confusion, it is commonplace for authors to use terms such as "mental effort" and "cognitive effort" interchangeably with the term "effort" [24, 47, 57]. While these terms appear to relate to the same cognitive processes, the term "physical effort" differs in that it refers to the regulation of our motor responses during a task. 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 [15]. 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 [24]. This implies that effort constitutes the summation of mental labour over time in order to achieve a goal.

2.4 Cost

Compared to the conceptualisation of effort and load, the definition of "cost" can be construed as more abstract in nature. Psychology has long considered humans as "cognitive misers", who strive to conserve cognitive effort and likewise avoid general effort exertion [58]. This implies that individuals value the effort they expend, treating effort as a cost [58]. Human behaviour consists of constant trade-offs between effort and reward [39]. How an individual performs in a certain task will rely partly on their decision to apply cognitive effort in the pursuit of attaining that reward. However, the limited information processing of the human brain is a constraint which underlies these trade-offs in that the central executive or the "control centre" of the brain directs these cognitive processing resources in line with our behavioural goals [39]. From this perspective, the level of cognitive processing allocated to a specific task at a given moment, is selected tactically on the basis of a cost-benefit analysis [39]. The notion of cost-benefit analyses underlie much of the discussion surrounding the concept of cost. Boksem and Tops [5] discussed energetic costs, predominantly as fatigue, which emerge from the cost-benefit analyses of whether to expend or conserve energy. Individuals are considered to only exert energy on a task when these energetic costs are comparably low and reward benefits are comparably high. Over time, this invested energy is proposed to build to the point where it eventually outweighs the benefits and subsequently drives abandonment behaviour (e.g. when people stop).

Other theories have adopted a behavioural economic approach which uses monetary value to quantify the costs involved in low vs. high effort tasks [57]. Temporal costs have also been used to characterise the trade-off processes an individual engages in during a task, where time is perceived to be expensive, effort expenditure should be directed toward faster and less accurate strategies to achieve the task goal [39]. There appears to be a shared consensus among researchers in terms of the underlying cognitive processes at the core of "cost", predominantly in the notion that the exertion of effort comes at some kind of "cost". Subsequently, while there are different theories assigning different attributes to these costs i.e. time, money etc., they do not contradict nor refute each other. Rather it appears that cost is multidimensional and that the value assigned to "cost" may be dependent on a variety of factors such as the individual, the context, and the goal.

2.5 How Are These Concepts Related?

Cognitive load is considered a multi-dimensional concept encompassing aspects of both load and effort [29]. 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 [10]. Thus the relationship between effort and load appear relatively straightforward - effort is exerted by the individual, whereas cognitive load is experienced by the individual [10]. 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 [24]. This shared consensus was highlighted in the Section 2.4, in that the expenditure of effort comes at some degree of cost (including affective (i.e. fatigue [5]); temporal (i.e. time spent [39]); and economic costs (i.e. monetary [57]).

The discussion so far has highlighted both the unique qualities of cost, effort, and load, and how these concepts relate. The rest of this paper will focus specifically on CEL definition and measurement within the field ISR.

3 REVIEW PROCESS

To examine the extent to which CEL and their related concepts have been explicitly defined and measured in ISR, and to provide a basis for analysis and discussion, a literature review was conducted.

3.1 Research Questions

In order to fulfil the key aims of this perspectives paper and provide a valuable contribution to the ISR community, the papers identified in the following literature review were analysed and discussed in accordance with the following four research questions:

RQ1 What do ISR researchers mean by cost, effort, and load?

RQ2 What are the similarities and differences between cost, effort, and load?

RQ3 How have cost, effort, and load been measured in ISR?

RQ4 What are the relationships between the measures used, and cost, effort, and load concepts?

3.2 Literature Review Process

The first stage of the search process was to identify the key sources from which the studies would be selected. In their systematic review of Interactive Information Retrieval (IIR) evaluation studies, Kelly and Sugimoto [26] provide a comprehensive list of 31 sources (17 journals & 14 conference publications) reviewed and approved by four IIR experts. While it was not feasible for the scope of this present paper to examine all 31 sources, eight of these sources were selected after discussion with two ISR experts. Although not included in [26], the Conference on Human Information Interaction and Retrieval (CHIIR) was also included as a key source. Table 1 shows the list of sources consulted. Sources were limited to journals and conference proceedings and included full length research papers, short papers and brief communications. Although of a smaller

Table 1: Source and Publications Examined (T = title only search; T-A = title & abstract search)

Source Title	# of Papers Examined	# of Papers Included
Information Processing & Management (IP&M)	24 (T)	1 [44]
Journal of the Association for Information Science & Technology	12 (T)	2 [13, 20]
ACM/IEEE Joint Conference on Digital Libraries (JCDL)	12 (T)	1 [16]
ACM International Conference on Information and Knowledge Management (CIKM)	43 (T)	0
Proceedings for the Association of Information Science & technology (ASIS&T)	11 (T)	6 [4, 8, 18, 27, 37, 61]
ACM Special Interest Group on Information Retrieval Conference (SIGIR)	122 (T-A)	3 [22, 54, 62]
Conference on Human Information Interaction & Retrieval (CHIIR)	48 (T-A)	4 [6, 31, 35, 43]
Information Interaction in Context (IIiX)	14 (T-A)	2 [33, 48]
Conference on Human Factors in Computing Systems (CHI)	111 (T)	0

scale, short papers and brief communications were considered appropriate for inclusion as they are still expected to provide clear details about their methods which is the key interest of this paper.

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 used for the Association for Computing Machinery Database ACM Digital Library (DL) and the American Society for Information Science (ASIS&T) Digital Library is as follows: ***(effort OR cost* OR "mental workload" OR "cognitive load" OR workload)***. The following search term was used for the Journal of Information Processing and Management (IP&M) database:***(effort OR cost OR "mental workload" OR "cognitive load" OR workload)***.

Initially, the search query was filtered to search the keywords within the title-abstract, however this yielded a large volume of papers across all three databases. For example, for all sources searched within the ACM DL database a total of 2,429 papers were retrieved in the title-abstract search. Upon a manual scan, many were focused out-with the ISR literature. Therefore, due to the large number of papers retrieved and the time- consuming nature of manually reviewing and identifying relevant studies, it was considered appropriate that the literature search would focus primarily on papers where keywords appeared in the "title". Additionally, as the aim of this paper focuses on studies which explicitly propose CEL methods, it seemed likely that a CEL term would appear in the title of such studies. For searches that yielded 10 or less papers in title-only search, a title-abstract search was then also included. This was the case for 3 conferences; the Conference on Human Information Interaction & Retrieval (CHIIR); the Conference on Information Interaction in Context (IIiX) and ACM Special Interest Group on Information Retrieval Conference (SIGIR), all other searches remained title only. For all searches a time span of 20 years (September 2000-September 2020) was defined in the search criteria. This was chosen as a means to focus on more contemporary methods of CEL measurement in ISR. A total of 397 papers published between 2000-2020 were retrieved from the database searches.

This collection was then refined by selecting papers that fulfilled the following inclusion criteria: (1) where the primary goal of the study was to explicitly propose a method of measuring CEL or its related concepts; (2) where CEL is examined within an ISR context, i.e. the user is engaged in interactive searching; and (3) where the method involves the active participation of people. To identify the relevant studies, all 397 abstracts were manually reviewed. Nineteen papers were identified as satisfying the inclusion criteria. Finally, seven more papers (sources: Human Computer Information Retrieval Symposium (HCIR) [21]; European Conference on Information Retrieval (ECIR) [14]; Advances in Human-Computer Interaction Journal [46]; Journal of Innovation in Health Informatics [2]; Decision Support Systems Journal [55]; Proceedings of International Joint Conference on Web Intelligence & Intelligent Technology (WI-IAT) [30]; and Computers in Human Behaviour [56]), were identified through reference crawling of the nineteen papers already retrieved. Thus, the final overall corpus of literature consisted of twenty-six papers published between 2000-2020.

4 RESULTS

4.1 RQ1: CEL Definitions in ISR

To answer RQ1, all studies found were analysed to identify whether they provided an explicit definition of the measured CEL concept. From the 26 papers selected for review, eight did not provide an explicit explanation of what they mean by the CEL concepts used in their study. All identified definitions were then extracted and organised according to their CEL concept. Following guidance from [28], the individual elements of CEL definitions were then separated, categorised and quantified. The categories are discussed in the following sections.

Cost: The term "cost" was used in all four studies. For the three studies which explicitly defined cost, two main categories were identified:

- *Interaction Oriented/Count Based:* Cost was characterised as the interactions/ number of actions occurring between the user and the system ($N = 2$).
- *Time Orientated:* Cost was also characterised in terms of time spent during the user-system interaction ($N = 2$).

Effort: Twelve papers in total measured effort or its related concepts. "Effort" was the most frequently used term ($N = 10$), with "mental effort" and "cognitive effort" only referred to once. Eleven of the twelve studies provided an explicit definition. While no definition was the same, three main categories emerged:

- *Cumulative/Total Work:* The first category ($N = 4$) to emerge characterises effort in terms of the cumulative or total amount of physical/mental work that the individual applies towards an outcome.

- *Interaction Oriented/Count Based:* The second category (N = 5) involves the characterisation of effort as the interaction or number of actions which occur between the system and the user.
- *Meta-cognition/Conscious Awareness:* Finally, the third category (N = 2) to emerge refers to effort as a volitional and intentional process in which the individual is consciously aware. Mental effort, cognitive effort, and effort all shared similar conceptual elements with no obvious distinctions.

Load: The term "cognitive load" was most frequently used (N = 5), followed by "mental workload" (N = 3) and "workload" (N = 3). From the eleven studies explicitly measuring load, seven definitions were extracted. No explicit definitions were provided for the term "workload". Two main categories emerged from the definitions:

- *Capacity Based / Bounded Resource:* All "cognitive load" definitions (N = 5) shared similarities in terms of their reference to the notion of limited mental capacity and resources. Only one of the "mental workload" definitions aligned with this category.
- *Cumulative / Total Work:* The remaining two definitions used to characterise "mental workload" closely align with the "work" related elements also identified in effort.

4.2 RQ2: CEL Similarities and Differences

Similarities between CEL Concepts: Both effort and cost were characterised by the *Interaction Oriented/Count Based* category. This may align with the idea that exerting effort during a user-system interaction comes with some degree of cost. However, it also highlights how the conceptual overlap between terms may lead to confusion about what we are subsequently measuring i.e. is the system-user interaction and the number of actions performed indicative of effort or cost? Similarly, conceptual elements found in the *Cumulative/Total Work* category have been used to define both effort and load concepts. While effort is recognised in other domains as a facet of cognitive load [29], the overlap in conceptual elements means there is again confusion about whether the *Total Work* measured in a study is representative of effort or load.

Differences between CEL Concepts: The only category which differentiates effort from cost and load concepts is the *Meta-cognition/ Conscious Awareness* element, which suggests that effort is considered a subjective phenomenon whereas cost and load are not. This supports claims from other domains which consider effort as volitional and intentional [15, 24]. The *Capacity Based/Resource Bound* category distinguishes load from effort and cost. As the notion of limited resource capacity underpin CLT and MRT, it is not surprising that these elements distinguish load from cost and effort. Finally, the *Time Orientated* aspect of cost makes a differentiation between cost and effort. This supports previous research which consider time as an expended resource during effort exertion [39].

4.3 RQ3: CEL Measurement in ISR

Similar to the lack of universally accepted definitions of CEL, there appears to be no single or standardised method to measure these concepts [51]. Following an analysis of the literature, a collection of subjective and objective methods used to measure CEL and their related concepts were identified. A general overview of the five most commonly used methods are described below.

Subjective Measures: Self-report or subjective measures are frequently used in ISR to measure cognitive load and perceived mental effort [51]. Hart and Staveland [23] boldly claim that *"subjective ratings may come closest to tapping the essence of mental workload"*. These measures rely on the assumption that an individual can make a reliable and valid assessment of the amount of load experienced within a specific context [51]. The review of ISR studies highlighted that self-designed questionnaires were used in three studies [16, 44, 48] measuring effort, and two other studies to measure mental workload [56] and cost [43]. While self-designed questionnaires were popular for the measure of effort, the NASA Task Load Index (NASA-TLX) was also identified as a dominant scale (N=8) in the measure of effort [22, 54] and load [2, 6, 14, 30, 46, 62]. The measure consists of 6 component scales (physical demand; mental demand; temporal demand; performance; frustration; and effort) which are weighted according to the context using a separate instrument [23]. The ratings of the 6 scales are then averaged to compute the final overall score from 0-100, known as the overall task load index [23]. To shorten the test, some studies employed the 'Raw TLX' [2, 6], where the weighting process is not included. Other studies [2, 46, 61] dropped individual sub-scales if they were considered as less relevant to the task.

Objective Measures: Search interaction logging (N=18), dual task methods (N = 7), and eye tracking (N = 7), were identified as the three most popular methods to objectively measure CEL in the ISR studies reviewed. Search interaction logging was frequently used in the measure of cost and effort, and generally, is one of the most commonly used methods for data collection within the field of ISR evaluation [25]. Dual task methods are underpinned by the notion of limited cognitive resources and in particular, the theory of Multiple Resources [59]. The method involves the user engaging in two tasks simultaneously, a primary task and an auxiliary task. The core premise of the dual-task method is that by varying the amount of cognitive load an individual experiences with the secondary task and then observing their subsequent performance, the researcher can gain a degree of insight as to which aspects of the primary task are most demanding, or perhaps most engaging [25]. Eye tracking was the most commonly used physiological measure of CEL concepts in the ISR studies reviewed. Previous studies have identified a significant correlation between blink activity (latency and rate), pupil size, and fixation duration, with varying levels of working memory demand [32]. Pupillary response in particular has been used to measure mental load in Information Science and Human-Computer Interaction (HCI) research [20].

4.4 RQ4: Relations b/w Measures & Concepts

The next section provides an overview of the relationships between CEL concepts and how they have been measured in the ISR literature (see Figure 1 for a full overview of relationships).

Cost: Search interaction logs were used to measure both categories of cost. Metrics such as number of queries issued, query length, pages saved, and snippets viewed, were used as indicators of *Interaction Orientated/Count-Based* costs. Whereas metrics such as dwell time and total time to enter queries were used in *Time-Oriented*

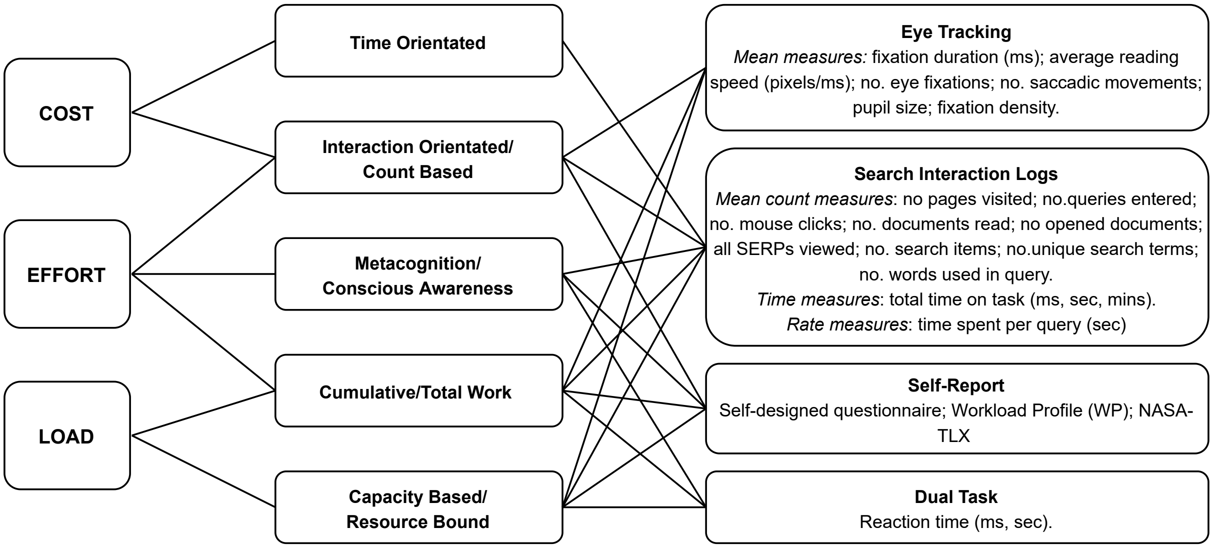
COST
EFFORT
LOAD
Time Orientated
Interaction Orientated/ Count Based
Metacognition/ Conscious Awareness
Cumulative/Total Work
Capacity Based/ Resource Bound
Eye Tracking
Mean measures: fixation duration (ms); average reading speed (pixels/ms); no. eye fixations; no. saccadic movements; pupil size; fixation density.
Search Interaction Logs
Mean count measures: no pages visited; no.queries entered; no. mouse clicks; no. documents read; no opened documents; all SERPs viewed; no. search items; no.unique search terms; no. words used in query.
Time measures: total time on task (ms, sec, mins).
Rate measures: time spent per query (sec)
Self-Report
Self-designed questionnaire; Workload Profile (WP); NASA-TLX
Dual Task
Reaction time (ms, sec).

Figure 1: Relationship between conceptual categories of CEL and their measurement (incl. metrics)

characterisations of cost. Eye-tracking was used to measure number of fixations on visited pages and task descriptions in the context of cost as an *Interaction-Orientated/Count-Based* concept.

Effort: Search interaction logs were also used to measure three conceptual categories. The number of mouse actions (clicks & scrolling) and time on task were used as effort indicators in all three conceptual categories. In *Cumulative/Total Work* and *Interaction/Count-Based* categories, number of queries, number of documents opened, and number of search terms issued were used as indicators of effort. Eye tracking was used in *Interaction Orientated/Count-Based* depictions of effort and reflected user-system interactions such as; reading speed, number of eye fixations on documents, and duration of fixation on documents. Only one of the studies that used self-report methods conceptualised effort in terms of *Meta-cognition/Conscious Awareness.* The NASA-TLX was used as a measure of effort relative to *Cumulative/Total Work.*

Load: Dual task and eye tracking methods were used to measure both characterisations of load. Pupillary response and fixation duration were used as indicators of load when measured via eye tracking. Time-on-task was the only indicator of load taken from search interaction logs. Note that workload measures have not been included in Figure 1, as no explicit definitions were provided.

4.5 Summary of Results

As previously discussed, operational properties emerge from the nominal definitions provided at the offset. If we were to examine each individual CEL study and their respective measures in isolation, then we could claim some degree of internal validity as the measures quite closely align with the conceptual elements provided. For example, if we take the conceptual properties of cost i.e. *Time* and *Interaction/Count Based*, then measures such as search interaction logs and eye tracking lend themselves well to measuring these indicators. Similarly, dual-task and eye-tracking measures used to measure load make intuitive sense as they are considered as sensitive measures in assessing the dynamic nature of load and the notion of limited capacity [25]. However, problems arise when we try to make comparisons between studies claiming to measure the same CEL concept, due to a variety of disparities in their measures such as the granularity of measurement (e.g. clicking a mouse vs. reading a document), level of analysis (e.g. task stage vs. session level), and the reliability and validity of methods used. The next section will discuss these main issues in more detail.

5 MAIN ISSUES

5.1 Ambiguity between Concepts and Measures

Firstly, there is clear overlap in the metrics used to measure different CEL concepts. Search interaction data was used in the measurement of all three CEL concepts with often the same metrics used to measure different concepts. For example, "number of queries issued", "number of documents opened", and "number of pages viewed" have been implicated as both a measure of cost [33, 43] and effort [8, 21, 27]. Likewise, "time-on-task" was used as a metric for all CEL concepts: cost [43, 62]; effort [8, 48, 54]; and load [13, 46], implying that time-on-task is an adequate indicator of all three concepts. Similar commonalities were observed in eye tracking methods where "number of fixations" were used as metric for both effort [18] and cost [62], and "duration of fixations" as a metric for both effort [18, 21] and load [55]. These observations not only highlight a clear overlap between the different CEL concepts and their measurement in ISR, but they also reflect the challenging aspect of using interaction methods more generally — how can we correctly relate these signals to complex CEL concepts?

This question leads to the key issue of whether these methods are actually measuring what they claim to measure. Search interaction logs cover a variety of different actions ranging from a simple mouse click to viewing a whole document. As observed from the studies [4, 8, 33, 44, 62], these actions are often *"count based"*. However, it is not clear how we can compare, say the total number of clicks to the total number of queries examined? Are clicks more costly, more effortful, or more load inducing than queries? Going beyond counts of interactions, how do we compare different interaction times, for example, is the time spent reading a document more costly, etc., than the time spent browsing result items? And how can we then subsequently compare the CEL between interactions like clicks vs scrolling, reading, querying, etc.? While we can acknowledge that all cognitive processes occur over time, the amount of time taken to complete a task or reach a goal can be influenced by a variety of individual and contextual factors (i.e. prior knowledge, complexity of the task, relevance) [51]. Even under strict experimental conditions where these factors are controlled, Sweller [51] argues that there remains little theoretical relation between time-on-task and user experience of effort. Thus, the issue with search interactions in general, is their assumption that actions have fixed values of "cost" or "effort". However, every user interaction with a system and its information is unique in terms of the physical, cognitive, and affective experience. So, while the actions which emerge from the interaction such as querying, examining a document, and so on, can be directly observed, they only offer an indirect glimpse into the cognitive activity of the user [25].

5.2 Level of Analysis

Only two [20, 46] out of the eleven studies which measure cognitive load or its related concepts, explicitly consider its dynamic nature and the different types of load as proposed by CLT. Firstly, referring back to MRT and CLT, the interaction between cognitive resources and performance reflects a constant and dynamic interplay [59]. Therefore, studies which analyse results at the task session level can only really provide a static, post-hoc indication of cognitive load- and the shifts in cognitive demands on the user during the search interaction are unlikely to be exposed through average values. Therefore, it is important for researchers to acknowledge that when cognitive load differences are absent at the session level, it is possible that they may exist at the task-stage level [19].

Secondly, studies that measure cognitive load as a singular concept [13, 55, 56] ultimately struggle to characterise which cognitive load type is consuming the users cognitive capacity during a search task. Identifying which processes have caused the perceived amount of cognitive load, i.e. understanding that the load experienced by the user was caused by the layout of the interface (extraneous load) as opposed to the complexity of the task (intrinsic load) could help researchers reach more valid conclusions and in turn, inform better system design decisions. Schmutz et al. [46] used the dual task method - which is a method designed to measure instantaneous cognitive load. When they compared the subjective load ratings obtained at the end of the session, they did not correlate with the averaged load scores from the dual task method. This misalignment between the two measures, both supposedly measuring load, reveal the possibility that both methods are measuring different dimensions of load or that any differences in load which may have arisen were masked by the averaging of the dual-task performance data. It could also be that the participants are falling prey to the peak-end rule, such that their experience is shaped by the peak and end events - rather than the average. More generally, these findings highlight the potential danger of using static measures such as post-hoc self-report in isolation, where it may be unclear which type of load the user has interpreted through the question items.

5.3 Questionnaires / Self Report Measures

Historically, ISR researchers have been less concerned about the validity and reliability of their subjective measures perhaps as much as other domains who use self-report measures and questionnaires to gather data from their participants [25]. Studies who use measures which have not gone through the rigorous scrutiny of validity and reliability testing may be subject to measurement error imposed by the items themselves [25]. Each of the three CEL concepts were measured at least once via self-designed questionnaires [16, 43, 44, 48, 56]. These questionnaires tended to follow different formats with distinct variations in the number of units used for the scales. For example, three self-designed questionnaires measuring effort [16, 44, 48] used a range of different scales: (i) binary "yes" or "no" responses [48], (ii) eleven point scale (0-10) [44], and (iii) a seventeen point scale (1-17) [16]. The lack of standardisation of thresholds in these questionnaire scales may suggest different levels of effort, e.g. 10 = high level of effort [44] vs. 17 = high level of effort [44]. One study [16] measured effort using answers to eight search behaviour questions. While the questions themselves reflect common search interaction metrics used in other studies of effort (i.e. number of Boolean operators used), the justification of the weighting system used to calculate the effort score was poorly explained. For example, question 8: *"how many terms do you typically enter before submitting your search"* was given higher weighting over the other seven search behaviours with little empirical explanation as to why.

Problems with the CEL term used and the unit used to measure it also came to light. For example, the search behaviour *"time"* was employed as a unit of effort in the questionnaire of [48] i.e. *"It took me less than 5 minutes to complete the task?"*, but used as a unit of cost in another [43] i.e.*"how long did it take to finish the task?"*. Subsequently, these studies make claims that the same metrics have measured different concepts. This raises the question of whether these questionnaires are really measuring the CEL concept they are claiming to measure?

While the dangers of using self-designed questionnaires have been highlighted, using well-established self-report measures can also be troublesome. Across all of the studies reviewed, the NASA-TLX was the most popular subjective tool of measurement. Over half of the studies employing the NASA-TLX [2, 14, 17, 30, 46] used the tool for the purpose of comparing workload across search systems. However, the tool is designed to distinguish among tasks, not systems [25]. Consequently, there arises uncertainty regarding the validity of claims made in these studies. It was also found that little connection was established between the definitions used-if any, and the NASA-TLX in studies which explicitly measure workload [6, 14, 62]. This raises the question of why this measure was chosen in the first place? If we refer back to the problems related to level of analysis, the NASA-TLX is another example of a post-hoc static measure being used to measure a concept that is considered dynamic and instantaneous. Perhaps the lack of definitions in these studies subsequently led to the conceptual properties of workload being overlooked - and therefore not appropriately measured? In the field of Human Factors, there has been much discussion regarding the unjustified popularity and subsequent over-reliance of the NASA-TLX as a workload measurement tool [11]. With critics arguing that the tool lacks sensitivity and predictive validity compared to other questionnaires, and has failed to continuously adjust towards optimal validity since its emergence [11]. Dekker and Hollnagel [12] sum up this argument, claiming that *"workload is a measure defined by consensus rather than reference to a model"*. Concerns about the theoretical and empirical foundations of the NASA-TLX makes it disconcerting that the tool is now synonymous with the definition of workload across fields [11]. The absence of explicit definitions of workload in ISR studies which have used the NASA-TLX may further reflect this issue.

6 CHALLENGES AND FUTURE DIRECTIONS

While CEL measures and concepts have been central within many ISR studies it is clear that we need to begin to address the issues highlighted in this paper and create a stronger theoretical and conceptual underpinning of CEL in the field of ISR. The two main challenges are discussed in this section alongside recommendations for future CEL research.

Challenge 1 - Definition of CEL Concepts: There is no universal definition or previously established theory on CEL concepts within the ISR literature. We further observed that few studies explicitly define CEL concepts. Instead, most studies relied upon intuitive notions of CEL, rather than the use of accepted or established definitions from existing theory. To differentiate between CEL concepts in the ISR context, we have created a tentative framework for defining these CEL concepts. Below we describe our framework and show how these concepts are related in Figure 2.

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. 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.

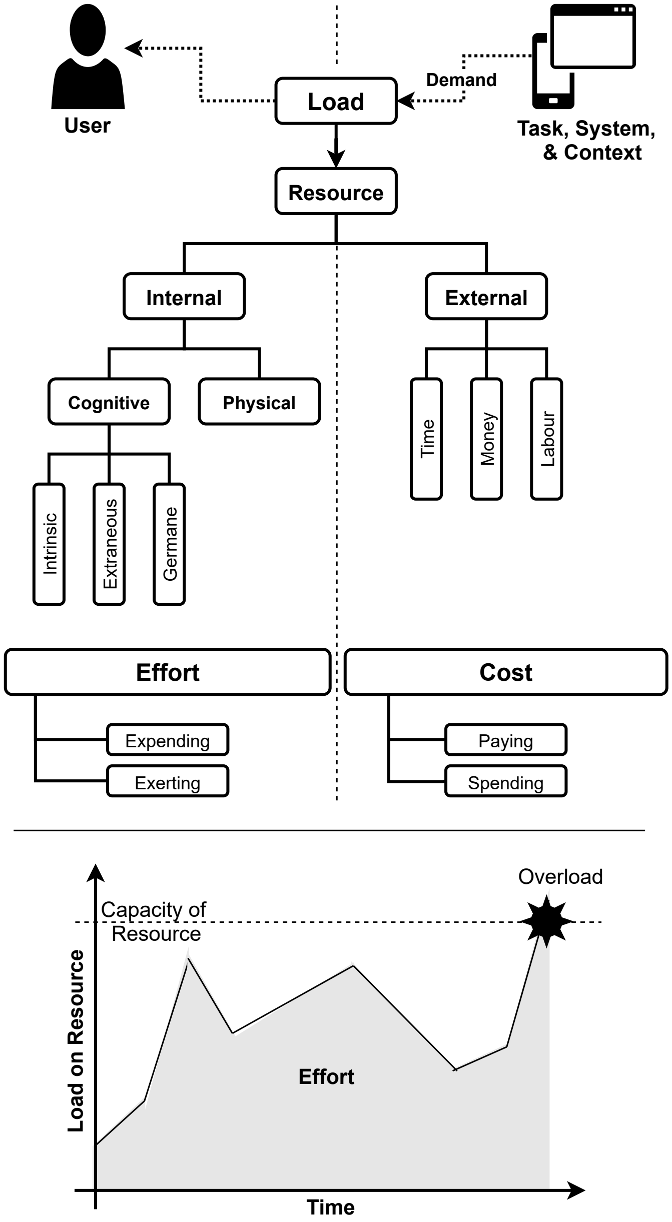

Figure 2: Top: The relationships between CEL Concepts. Bottom: A graphical depiction of the relationship of the load experienced by a user over time for a given internal resource. When the load demanded by the task, system and context exceeds the capacity of the user's resource, then they hit overload. The effort experienced by the user is the total load over time (i.e. the area under the curve).

Load: Given a particular resource, and the demand imposed by the task, system, and context, we can generalize the concept 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.

Overload: Taken together, the concept 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 concept 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 2, the bottom plot shows how effort is related to load, where effort is the total load expended 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.

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.

Challenge 2 - Measurement of CEL Concepts: The second challenge relates to the precision and accuracy of the methods currently used to measure CEL in ISR studies. This review has highlighted that although subjective methods are commonly used to measure CEL concepts, the majority of self-report measures are not validated. Furthermore, they do not adequately capture or delineate between the different properties of the CEL concepts identified. When considering questionnaires claiming to measure effort [16, 27, 44] for instance, we observed a large variety of questions, with different formats, scales and units – making them almost impossible to compare. In addition, these measures were also often administered post-hoc, which makes it unclear which aspect of effort is being assessed i.e. total effort, average effort, peak effort, end effort, etc.? Similarly, subjective methods used to measure load, such as the NASA-TLX, also tend to be assessed post-hoc — but load is inherently dynamic and instantaneous, so can a post-hoc assessment accurately capture this? To help validate and support self-report findings, we recommend using a multi-method approach to CEL measurement — combining both subjective and objective methods. However, a key direction for the field is to establish reliable and valid self-report instruments which can accurately represent CEL concepts within an ISR context.

On the other hand, objective measures currently used within the ISR field such as dual task and eye-tracking, already offer dynamic and real-time assessment of cognitive load during a search task. However, the key challenge for researchers is ensuring that the unit of analysis used to measure cognitive load can accurately reflect its conceptual properties (i.e. dynamic, instantaneous). For example, examining cognitive load at finer granularity, e.g. at the task-stage level, rather than calculating average values across a task session may help researchers facilitate this. To more accurately visualise the detailed trends and patterns of cognitive load in real-time, a key progression in ISR research would involve the use of more sophisticated sensors. For example, the field of Cognitive Neuroscience offers highly sensitive and precise 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 [51]. Alternatively, within a HCI context, Abdelrahman et al. [1] present thermal imaging as a less expensive and perhaps less obtrusive method - where users facial temperature is representative of cognitive load fluctuations. While these measures offer an exciting avenue for future CEL research within ISR, 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 concepts to determine the trade-off between their scalability and accuracy.

7 CONCLUSIONS

Although CEL concepts, measures and measurement has offered useful insights across the ISR literature, we have highlighted the key issues with how these concepts are currently being used and subsequently measured. It is hoped that the ISR research community can benefit from the working definitions provided in this perspectives paper, 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 concepts in ISR. With this unified approach, we anticipate a refinement of existing CEL measures alongside the development of new ones, inevitably making comparison and generalisation across studies more achievable, and lead the field to a greater understanding of these complex concepts.

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] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive Heat: Exploring the Usage of Thermal Imaging to Unobtrusively Estimate Cognitive Load. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (9 2017). https://doi.org/10.1145/3130898

[2] 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

[3] Leif Azzopardi, Diane Kelly, and Kathy Brennan. 2013. How Query Cost Affects Search Behavior. In *Sigir.* 23–32. https://doi.org/10.1145/2484028.2484049

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

[5] 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

[6] 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

[7] 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

[8] Michael J. Cole, Jacek Gwizdka, Chang Liu, and Nicholas J. Belkin. 2011. Dynamic assessment of information acquisition effort during interactive search. https://doi.org/10.1002/meet.2011.14504801149

[9] William S. Cooper. 1968. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation* 19, 1 (1968), 30–41. https://doi.org/10.1002/asi.5090190108

[10] 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

[11] 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

[12] Sidney Dekker and Erik Hollnagel. 2004. Human factors and folk models. *Cognition, Technology & Work* 6, 2 (2004), 79–86. https://doi.org/10.1007/s10111-003-0136-9

[13] 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

[14] 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

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

[16] Paul Gerwe and Charles L. Viles. 2000. User effort in query construction and interface selection. In *Proceedings of the ACM International Conference on Digital Libraries.* 246–247. https://doi.org/10.1145/336597.336679

[17] Roberto González-Ibáñez, Carlos Barrera-Pulgar, and José Luis Varela-Otárola. 2016. Evaluating touch-based interactions in an image search task. In *CHIIR 2016 - Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval.* 265–268. https://doi.org/10.1145/2854946.2854999

[18] 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. In *Proceedings of the Association for Information Science and Technology*, Vol. 54. 128–138. https://doi.org/10.1002/pra2.2017.14505401015

[19] 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

[20] 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

[21] J Gwizdka and MJ Cole. 2011. Least effort? Not if I can search more. In *In Proceedings of the 5th Workshop on Human-Computer Interaction and Information Retrieval.*, Vol. 2. 2012. https://doi.org/10.1.1.306.8297{&}rep=rep1{&}type=pdf

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

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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. In *Proceedings of the American Society for Information Science and Technology*, Vol. 42. https://doi.org/10.1002/meet.14504201155

[28] 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

[29] 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

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

[31] Cheng Luo, Xue Li, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, and Shaoping Ma. 2017. Investigating users' time perception during web search. *CHIIR 2017 - Proceedings of the 2017 Conference Human Information Interaction and Retrieval* (2017), 127–136. https://doi.org/10.1145/3020165.3020184

[32] Stewart Martin. 2014. Measuring cognitive load and cognition: metrics for technology-enhanced learning. *Educational Research and Evaluation* 20 (2014), 592–621. https://doi.org/10.1080/13803611.2014.997140

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

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

[35] Prithima Reddy Mosaly, Lukasz Mazur, and Lawrence B. Marks. 2016. Usability evaluation of electronic health record system (EHRs) using subjective and objective measures. In *CHIIR 2016 - Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval.* 313–316. https://doi.org/10.1145/2854946.2854985

[36] E. N. and Clark L. Hull. 1943. Principles of Behavior. An Introduction to Behavior Theory. *The Journal of Philosophy* 40, 20 (1943), 558. https://doi.org/10.2307/2019960

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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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 &amp; Retrieval (CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 285–288. https://doi.org/10.1145/3176349.3176890

[44] 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

[45] G. Salton. 1970. Evaluation problems in interactive information retrieval. *Information Storage and Retrieval* 6, 1 (1970), 29–44. https://doi.org/10.1016/0020-0271(70)90011-2

[46] 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

[47] 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

[48] Georg Singer, Ulrich Norbisrath, and Dirk Lewandowski. 2012. Ordinary Search Engine Users Assessing Difficulty, Effort, and Outcome for Simple and Complex Search Tasks. In *Proceedings of the 4th Information Interaction in Context Symposium (IIIX '12)*. Association for Computing Machinery, New York, NY, USA, 110–119. https://doi.org/10.1145/2362724.2362746

[49] 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

[50] 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

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

[52] 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

[53] 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

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

[55] 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

[56] Erik Wästlund, Torsten Norlander, and Trevor Archer. 2008. The effect of page layout on mental workload: A dual-task experiment. *Computers in Human Behavior* 24, 3 (2008), 1229–1245. https://doi.org/10.1016/j.chb.2007.05.001

[57] 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

[58] 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

[59] 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

[60] 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

[61] Yinglong Zhang and Jacek Gwizdka. 2014. Effects of tasks at similar and different complexity levels. In *In Proceedings of tthe 77th ASIS&T Annual Meeting, Seattle, WA, USA.*, Vol. 51. https://doi.org/10.1002/meet.2014.14505101093

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