A Survey on Measuring Cognitive Workload in Human-Computer Interaction

THOMAS KOSCH, HU Berlin, Germany
JAKOB KAROLUS, German Research Center for Artificial Intelligence, Germany
JOHANNES ZAGERMANN and HARALD REITERER, University of Konstanz, Germany
ALBRECHT SCHMIDT, LMU Munich, Germany
PAWEŁ W. WOŹNIAK, Chalmers University of Technology, Sweden

The ever-increasing number of computing devices around us results in more and more systems competing for our attention, making cognitive workload a crucial factor for the user experience of human-computer interfaces. Research in **Human-Computer Interaction (HCI)** has used various metrics to determine users' mental demands. However, there needs to be a systematic way to choose an appropriate and effective measure for cognitive workload in experimental setups, posing a challenge to their reproducibility. We present a literature survey of past and current metrics for cognitive workload used throughout HCI literature to address this challenge. By initially exploring what cognitive workload resembles in the HCI context, we derive a categorization supporting researchers and practitioners in selecting cognitive workload metrics for system design and evaluation. We conclude with three following research gaps: (1) defining and interpreting cognitive workload in HCI, (2) the hidden cost of the NASA-TLX, and (3) HCI research as a catalyst for workload-aware systems, highlighting that HCI research has to deepen and conceptualize the understanding of cognitive workload in the context of interactive computing systems.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; *HCI theory, concepts and models*; • **General and reference → Surveys and overviews;**

Additional Key Words and Phrases: Cognitive workload, workload assessment, categorization, questionnaires, workload-aware computing, physiological sensing, cognition-aware interfaces

ACM Reference format:
Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Comput. Surv.* 55, 13s, Article 283 (July 2023), 39 pages.
https://doi.org/10.1145/3582272

Thomas Kosch, Jakob Karolus, and Johannes Zagermann contributed equally to this research.

This research was funded by the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 and – Project-ID 414984028 – CRC 1404 FONDA.

Authors' addresses: T. Kosch, HU Berlin, Rudower Chaussee 25, 12489 Berlin, Germany; email: thomas.kosch@hu-berlin.de; J. Karolus, German Research Center for Artificial Intelligence, Trippstadter Str. 122, 67663 Kaiserslautern, Germany; email: jakob.karolus@dfki.de; J. Zagermann and H. Reiterer, University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany; emails: {johannes.zagermann, harald.reiterer}@uni-konstanz.de; A. Schmidt, LMU Munich, Frauenlobstraße 7a, 80337 Munich, Germany; email: albrecht.schmidt@ifi.lmu.de; P. W. Woźniak, Chalmers University of Technology, Chalmersplatsen 4, 412 96 Göteborg, Sweden; email: pawel.wozniak@chalmers.se.

0360-0300/2023/07-ART283 $15.00
https://doi.org/10.1145/3582272

1 INTRODUCTION

Due to the increasing number and complexity of novel computing systems, be it personal devices or public computers, the impact of cognitive workload during the interaction with human-computer interfaces has become a relevant factor affecting user experience. The research field of Human-Computer Interaction (HCI) develops designs, conducts user studies, and evaluates interactive computing systems, focusing on improving usability and user experience. Cognitive workload analysis during interface design and its posterior evaluation is an ongoing research area in HCI. While HCI researchers use the term "cognitive workload" extensively, there is, to date, no consensus on a definition of the concept, nor is there a gold standard for measuring it. At the same time, the assessment of cognitive workload remains a primary objective in HCI research. Cognitive workload is viewed as a factor that must be reduced or kept at an engaging level to achieve a satisfying user experience. By considering cognitive workload aspects, the HCI community strives to build improved user interfaces. Yet, the detailed history of how cognitive workload was adapted to HCI from the field of psychology [88, 89, 149] via HCI's human factors legacy remains uncharted. The HCI community used cognitive workload measures as early as 1994 [33], but the concept's generic applicability to computer systems remained a largely unexplored topic.

Parallel notions of the concepts of cognitive workload emerged in HCI literature, resulting in an epistemic problem: On the one hand, cognitive workload is widely recognized as an important metric. On the other, no universal measures or definitions are recognized by the community. Despite this, research and practice use cognitive workload extensively to benchmark user interfaces where different human cognitive systems, such as memory, motor, visual, and auditory perception, are used. Traditionally, research in HCI employed questionnaires [90, 91] or think-aloud protocols [190] as cognitive workload measures. The NASA-TLX [90, 91] questionnaire has become a household name in HCI research, and it is widely taught in HCI education. With the advent of ubiquitous computing, sensing technologies enabled building systems that use physiological sensing data and user behavior as an objective indicator [38] of a user's mental workload. Sensing technologies provided researchers with a potent tool to objectively quantify the evoked workload of their user interface design in real-time and, going beyond, allow cognitive workload to be treated as a real-time input for user interface adaptations [70, 262]. This enables researchers to design for specific mental states of their users, providing a more fine-grained view on workload analysis or adapting interfaces to individual cognitive workload states.

Although well-researched literature exists about the evaluation of human work about cognitive and ergonomic factors [253], the missing consensus within the HCI field still leads to misuse of metrics and, ultimately, misunderstanding the concept of cognitive workload. Misinterpretation of research findings, lack of reproducibility, and unsuitable measurement modalities are possible threats if cognitive workload is not fully understood. Furthermore, researchers and practitioners are challenged by an abundance of cognitive workload assessments, potentially hindering a proper evaluation of computing systems adapting to user workload. Consequently, there is a need to categorize cognitive workload metrics in HCI research to support researchers and practitioners in their choice of evaluation metrics and allow a consistent interpretation of their findings within the scope of studies conducted in the HCI field.

This paper presents a systematic review of the concept and measurements of cognitive workload in HCI research. Through charting the various metrics used for cognitive workload, we identify opportunities and challenges for the future use of cognitive workload metrics in HCI. Further, we present a categorization providing insights into choosing appropriate cognitive workload metrics. Our work focuses explicitly on HCI's implicit understanding of cognitive workload rather than drawing from a theoretical background as suggested by past work [98]. Consequently, we seek to chart *how HCI understands and measures cognitive workload*. We contribute results that show that the HCI field has primarily measured cognitive workload using post-hoc questionnaires, most predominantly the NASA-TLX. Later work used physiological sensing to use workload as an input or conduct in-depth analyses of how workload varies during a task. Finally, our review shows opportunities for educating and guiding researchers to understand the role of cognitive workload in HCI, allowing them to conduct focused and reproducible research through an improved understanding of cognitive workload metrics.

REVIEW SYNOPSIS

The review provides researchers and practitioners interested in evaluating interactive systems with an overview of the theoretical background, metrics, and uses for cognitive workload in HCI research. In contrast to previous research reviewing work on cognitive workload [15, 60, 64, 65, 126, 186, 265], our survey focuses on cognitive workload assessments in the field of HCI to support researchers in their choice of cognitive workload measurement for their studies. We summarize metrics used throughout HCI research to assess cognitive workload for interface optimizations and adaptations. This paper makes four key contributions: (1) We contextualize our inquiry to introduce cognitive workload theory. (2) Next, we report on our review methodology and describe frequently used cognitive workload metrics in the research of HCI. (3) We then categorize cognitive workload metrics to support the selection of appropriate metrics in future HCI studies. This includes a step-by-step procedure to select appropriate cognitive workload measurements (see Figure 4) and an interactive paper library that can be extended by future research contributions from the HCI community,[1] hence providing researchers and practitioners with an up-to-date tool for an efficient selection of suitable cognitive workload metrics beyond this paper. (4) Finally, we present three research gaps to advance the field and discuss opportunities and possible pitfalls for future HCI research that uses cognitive workload as an evaluation metric or an input modality.

2 BACKGROUND

We begin our inquiry by exploring how HCI researchers and practitioners understand and use cognitive workload measures in their studies. Previous research has presented different theories or models originating from the field of educational [231, 234] and cognitive psychology [88, 89, 250–252]. A majority of these concepts originate from the field of human factors, where HCI was adapting the concepts for their workload assessments. Although the two fields, human factors and HCI, have deviated during the past years [84], several cognitive workload assessments from human factors were adopted for HCI research. In this context, the HCI community uses the general terms "cognitive workload", "cognitive load", or "mental workload" in most of their studies to describe the cognitive demand of a task. Although these theories lay the foundations for assessing cognitive workload in human factors, cognitive psychology, and pedagogy, they were not designed for the HCI field per se. Subsequently, mental workload assessments were used to evaluate the cognitive demand of interacting with user interfaces although they were not originally designed for the design and evaluation of user interfaces. For example, the NASA-TLX questionnaire [91], a method to assess pilot workload, was used as a mental workload measure in 1994 [33] in an HCI context for the first time.

The term workload characterizes the required mental effort to complete a task ("work") with several constraints ("load"), such as fixed periods or large batches of tasks [172]. Extensive cognitive workload influences an individual's task performance and ability to cope with concurrent tasks. Whereas cognitive workload is often associated with a negative experience, it can be essential to elicit positive effects during daily activities. Nevertheless, a permanent high cognitive workload

[1] https://www.thomaskosch.com/cl-paper-library - last access 2023-01-24.

harms attention and focus when there is an imbalance between the individuals' cognitive abilities and the task difficulty, potentially distracting the person and eliciting mistakes. Cognitive workload occurs when events demand the mental resources of individuals in a specific context. These events can occur within an external context, such as interruptions caused by notifications, or an internal context, where the demand is inherent to the task. Multiple events that simultaneously impact a single processing channel lead impact cognitive processing performance [250–252]. How well a person copes with these events depends highly on their cognitive prowess and the type of events. For example, users interrupted by notifications revert faster to their previous mental state when multiple disruptions during long-term tasks demand higher cognitive abilities [147]. During times of high cognitive workload, the body allocates resources needed for dealing with the load. This includes changes in cortical activity, blood flow, an increase in respiration rate and electrodermal activity, or differences in eye behavior [56].

2.1 Understanding Cognitive Workload

Before we explore different workload theories and how they are employed in HCI research, we will clarify the understanding and term usage between them. Specifically, we will clarify the terms "cognitive load" and "mental workload". The term "cognitive load" was coined by Sweller [231] to describe the instructional design of pedagogical methods. Later, the three terms *intrinsic load*, *germane load*, and *extraneous load* were introduced to separate cognitive load into different modalities [234]. Although these concepts do not originate from the HCI field, we decided to include them in this survey paper since an increasing number of HCI papers started to mention the cognitive load theory to justify their workload assessment (cf. [6, 98]).

More recent work investigated the additive interaction between the two different concepts "cognitive load" (i.e., *intrinsic load*, *germane load*, and *extraneous load*) and "mental workload" [75]. Their results revealed an impact between task difficulty and time pressure, modulating the alertness, subjective, and psychophysiological workload measures of users. Thus, the definition and interpretation of cognitive workload including its underlying theories and models is a subject of controversy in psychology [57]. In the following, we describe concepts that theorize cognitive workload from an HCI perspective. These initial concepts led to a categorization of workload modalities later adopted for studies in the field of HCI. We begin by exploring theories, measurements, and assessments adopted in past HCI studies.

Since it originated in the 1980s, Sweller's cognitive load theory has become an acknowledged field within the field of instructions and learning design [64]. Sweller developed the cognitive load theory [231–233], a theoretical framework that explains how learning is limited through instructional design. Later, the theory divides the relationship between instructional design and problem-solving into three components: *intrinsic load*, *germane load*, and *extraneous load* [234]. Initially, Sweller's cognitive load theory originated from educational psychology and was unrelated to how interfaces impact user workload in HCI research. However, it provides an initial understandable separation between the different allocated mental resource fragments. Sweller assumes that each component cumulatively allocates and demands mental resources. *Intrinsic load* describes the inherent complexity of a task. Thus, the efficacy of a person to solve a given task heavily depends on their skill and ability to understand patterns or sequences. *Intrinsic load* is not trivial to manipulate since the intrinsic task complexity is associated with a person's necessary individual cognitive resources and proficiency. Keeping *intrinsic load* at a *sweet spot*[2] is considered to foster task engagement while reducing frustration [263]. *Germane load* represents

[2] A *sweet spot* is the optimal proximal zone, where a user is not affected by frustration or boredom that originates from cognitive over or underload.

the effort to process patterns within a task. Realizing new schemes or patterns that help solve a task may increase task engagement and foster learning. Thus, the maximization of *germane load* is emphasized as a crucial factor when designing engaging user interfaces. However, this is highly disputed among psychologists, where *intrinsic load* and *germane load* are considered to be identical [109, 218]. *Extraneous load* is manipulated by the task representation that can be perceived via human perception. For example, in the HCI context, a well-designed interface visualization allows for an easier interpretation of data. Relative to the other workload types, *extraneous load* is theoretically trivial to manipulate since the task representation can be exchanged through design. *Extraneous load* can be minimized to avoid unnecessary allocations of cognitive resources to resolve the intrinsic task complexity. Hence, *extraneous load* is directly related to the interface's representation and user interpretation. Work in HCI exists using *extraneous load* as the relationship between the visual representation of a task and measured workload [6, 98].

In contrast, "mental workload" is related to the cognitive demand of a task itself [172]. In contrast to the aforementioned cognitive load theory, mental workload refers directly to cognitive resources allocated by a task. The attention and task performance are more severely impacted the more mental workload is required. In this context, Wickens' suggested a multiple resource theory, proposing that human operators do not have one information processing channel [250, 251], but several processing channels that can be addressed simultaneously. If one information processing channel is overly demanded, for example, through the sequential processing of a task, the more the performance is expected to decrease. Thus, users need more information processing resources, especially when performing multiple tasks on the same information channel, leading to errors and lowered task performance. However, simultaneously addressing different information channels, such as auditory or visual, does not necessarily cause mental overload.

The concepts of Sweller's cognitive load theory and Wickens' multiple resource theory are different, yet, they are used by the HCI community in a mixed way (cf. [6, 98]). The aim of this work is not to unify these theories. Rather, we aim to show how cognitive workload is measured by the HCI community and provide an instructional guide to choosing suitable workload measures. This paper systematically collects and summarizes workload assessment to guide the HCI community in selecting appropriate workload measurements. To avoid confusion regarding the terminology, we use the terms "cognitive workload", "cognitive load", and "mental workload" synonymously to describe workload imposed through the instructional system design of user interface visualizations (e.g., *extraneous load*) or cognitive demand of users who process information.

2.2 Cognitive Workload in HCI Research: A Tool for Designing and Evaluating User Interfaces

Well-designed user interfaces are meant to support their users in their operation, preventing permanent high and low workload while fostering appropriate cognitive engagement. Therefore, cognitive workload has emerged as a frequently used metric for designing and evaluating interactive systems in HCI research. The assessment of cognitive workload helps interface designers in two ways: using a constructive approach (e.g., during early design phases), where cognitive workload is measured to improve the design of novel interactive systems, or using a summative approach (e.g., using existing systems) to compare the cognitive workload evoked by, for example, different visualizations or interaction techniques. Additionally, cognitive workload can be used as an input to design engaging adaptive interfaces that, for example, adapt visualizations in real-time. This concept is often referred to as *Cognition-Aware Computing* [37] and *Workload-Aware Computing* [128].

While most users can cope with changes in cognitive workload for a short time, the long-term management of high cognitive workload can be challenging. Over- or under-challenged user states are the result, leading to frustration or boredom [203]. In addition, health issues, such as burnout, mental exhaustion, or physical tiredness, may arise under prolonged exposure to cognitive workload [170]. To further elaborate on the role of cognitive workload in HCI research, we provide an example of a user interface inducing high workload and then present how HCI research uses cognitive workload assessments to improve this interface. For example, interfaces that require users to search for items or memorize lengthy menu structures can increase cognitive workload by straining their memory [24]. On the other hand, interacting with a lengthy menu list with many nestings for a short time might only impact the cognitive resources. However, fewer cognitive resources are available to solve the actual task if users are repeatedly forced to navigate the same complex menu structure at varying times. Here, the visual complexity of the representation of the menu structure might strain the user's cognitive resources. The visual complexity can be reduced by (1) providing, for example, shortcuts for repetitive actions directly on the interface, (2) including the task at hand as context information (e.g., only showing relevant items), or (3) sensing the cognitive workload through physiological sensing to predict following actions of a user. A typical HCI user study would compare different user interface alternatives (e.g., different visual representations of a menu) and measure the user's cognitive workload using questionnaires (e.g., NASA-TLX [91]) or physiological sensing. These assessments are tools for HCI researchers to design interfaces matching the cognitive resources of their users through the visual complexity of the representation of the user interface, hence manipulating extraneous load. Over the past 30 years, several metrics for cognitive workload have been explored by HCI research. Physiological sensing has gained attention as a measure of cognitive workload while, at the same time, new questionnaires have emerged to measure cognitive workload in various contexts. Hence, it becomes increasingly more challenging to maintain an overview of the most appropriate cognitive workload measurement modalities.

In this review, we investigate how this definition was applied in the design of interactive systems. To this end, we summarize how extraneous load was evaluated in previous HCI user studies and encapsulate the measures used to foster a shared understanding of cognitive workload assessments. This review focuses on the field of HCI in a computing sense, i.e., the interdisciplinary research domain and community historically built around the ACM Conference on **Human Factors in Computing Systems (CHI)** [84], usually classified by ACM as Human-Centered Computing. Thus, our inquiry addresses conceptualizations of cognitive workload within studies of discretionary use of computers [84]. We recognize that our analysis focuses solely on this academic tradition and that other communities, e.g., **Human Factors Engineering (HFE)** or Ergonomics, have built a more thorough understanding of workload (i.e., in the domain of human-machine interaction). We aim to show current practices, execution, and interpretation of cognitive workload assessments within the HCI community.

3 REVIEW METHODOLOGY

Our review follows an iterative process of identifying relevant venues, filtering, and analyzing interest publications. This section describes our approach to the systematic review. We outline the inclusion criteria of our literature search and the phases of the refinement process (exclusion criteria). We followed the PRISMA procedures [121, 174] for systematic reviews and meta-analyses. The complete PRISMA is depicted in Figure 1. We built our literature corpus by systematically identifying relevant papers through a keyword search in all proceedings and journals related to the field of HCI.

3.1 Systematic Search

The first phase of our review was a systematic keyword search in the digital library databases relevant to the HCI field.

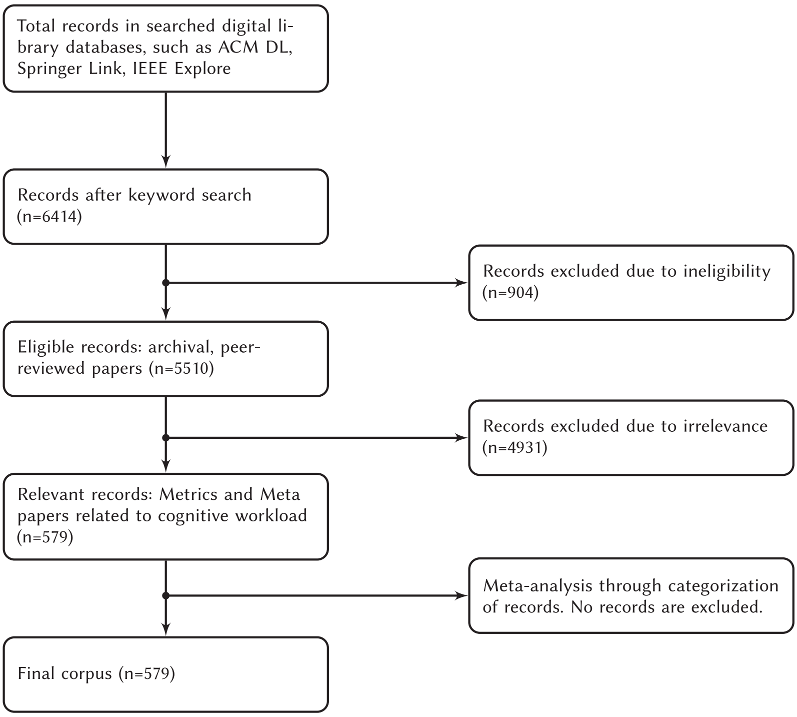

Total records in searched digital library databases, such as ACM DL, Springer Link, IEEE Explore
Records after keyword search (n=6414)
Records excluded due to ineligibility (n=904)
Eligible records: archival, peer-reviewed papers (n=5510)
Records excluded due to irrelevance (n=4931)
Relevant records: Metrics and Meta papers related to cognitive workload (n=579)
Meta-analysis through categorization of records. No records are excluded.
Final corpus (n=579)

Fig. 1. PRISMA flow chart highlighting the stages of our systematic search as detailed in Sections 3.1 and 3.2.

3.1.1 Inclusion Criteria for Venues. We decided to include all proceedings and journals in the research area of HCI in our search. To get an objective overview, we used Microsoft Academic[3] search to obtain conference and journal rankings in HCI. For conferences, we decided to include the top 20 venues and six conferences in the top 40, all known to the authors as featuring several papers highly relevant for this review. For journals, we included all top 10 journals. We additionally confirmed the relevance of the venues via CORE[4] rankings. We accumulated all matching papers (cf. Section 3.1.2) within all the digital libraries, which contained at least one of the conferences or journals listed.[5] We adjusted the search syntax to work with the search APIs provided by the publishers. This process resulted in an initial corpus of 6414 papers. The final cut-off date for all venues was the 30th of April 2020.

3.1.2 Inclusion Criteria for Keywords. We have introduced an interpretation of the term "cognitive workload" and its implication for interaction in Section 2. We discovered that the term is used with different semantics during this survey. To this end, we iteratively refined our keywords and evaluated these on a base corpus of top-rated venues as presented in Section 3.1.1. Starting from the initial keyword "cognitive workload", we identified synonyms used in literature and extended our set of keywords. The final regular expression used in the review is presented in Equation (1). This expression matches all commonly used terms and synonyms related to cognitive workload. Using the expression in the relevant databases resulted in our initial set of papers:

$$((cognitive \mid mental)\ (load \mid workload)) \mid (working\ memory) \tag{1}$$

[3] Service discontinued, previously available via https://www.academic.microsoft.com.
[4] https://www.core.edu.au - last access 2023-01-24.
[5] Publishers changed for some venues.

3.2 Review Process

Our systematic search yielded an initial corpus of 6414 papers in total; 4408 from conference proceedings and 2006 from journals. This section outlines how we filtered the papers to reach our final corpus. We employed two exclusion criteria and a final categorization of our paper selection as a meta-analysis step.

3.2.1 Exclusion Criteria. As detailed below, we reviewed the initial corpus concerning paper eligibility and relevance.

Paper Eligibility. We only considered archival, peer-reviewed papers, i.e., journal and conference papers, full and short papers, and book chapters. This approach eliminates workshop summaries, posters, works in progress, and other non-archival publications. We excluded 904 papers in this step.

Relevance. We used a double decision scheme to filter the initial corpus of eligible papers (5510 records). Four researchers participated in the filtering process. Per publication, two researchers independently assessed whether the publication was relevant to the survey. Using researcher triangulation, we ensured that at least two researchers had read every paper in the final corpus. No author had assessed the relevance of their work. Papers were considered relevant if they exhibited one of the following characteristics:

- **Metrics paper:** These papers discuss a system or application, including a form of cognitive workload that was measured, either using it as an evaluation metric or as an input channel.
- **Meta paper:** These papers discuss a concept of cognitive workload on a meta-level, such as definitions, application concepts, and possible frameworks for HCI research.

Each publication rated relevant by both researchers was included in the final corpus. In addition, we excluded another 4931 papers during this step. This final corpus (579 records) was then subjected to our analysis, extracting employed modalities in the categorization step.

3.2.2 Categorization. We aimed to understand the content structure within the corpus through this categorization, especially the employed modalities to measure cognitive load. Thus, we additionally categorized metrics papers into works that used subjective metrics and those that used objective metrics. Finally, we directly filtered by employed modalities, such as questionnaires or gaze, in a more refined categorization. During this step, we did not exclude any publications.

To systematically code all papers and build a taxonomy of all cognitive workload measures used in HCI, we used a qualitative analysis approach where we listed all cognitive load measures used in all papers in the corpus. In a coding meeting, four researchers then agreed on a naming scheme for the methods to achieve a uniform matrix of measures in papers. We then used affinity diagramming to iterative group papers and methods to reach a final taxonomy of methods. While applying this coding method does not guarantee that the proposed taxonomy is the only correct one, as no qualitative analysis can offer such a result, our approach ensures that the structure is comprehensive.

3.3 PRISMA Flow Graph

In the following, we present the complete PRISMA flow graph for our systematic review as detailed in this section (cf. Sections 3.1 and 3.2). It depicts the individual stages of paper inclusion and exclusion (right-hand side).

3.4 Literature Corpus

Our final corpus includes a total of 579 records: 482 records from 26 conferences[6] and 97 records from 11 journals. The complete list is available on our website.[7] Among the irrelevant papers, most were excluded due to the ambiguity of the term "working memory". Especially in computer science venues, this is a technical term used in processor architecture. Moreover, papers often used cognitive workload to motivate research or explain results. Here, cognitive workload is merely a hypothetical factor contributing to results. These works did not actively engage with metrics or concepts related to cognitive workload and were subsequently excluded.

4 COGNITIVE WORKLOAD METRICS IN HUMAN-COMPUTER INTERACTION

Informed by the records of our literature corpus, we describe current practices for cognitive workload measurements in HCI. This section aims to provide a comprehensive overview of individual metrics, their respective characteristics, and their relation to each other to inform researchers about the functionality of the measurement. Please refer to Section 5 on how these metrics are employed in HCI studies, including a step-by-step guide on choosing an appropriate metric. Additionally, our website provides an interactive search-and-filter interface supporting researchers and practitioners to find relevant literature – and with this suitable cognitive workload measurements – for their work. We present the whole spectrum from subjective (e.g., questionnaires) and objective (e.g., physiological) measurements. If not stated differently, details on the analyses of the measures can be found in the referenced literature. As an overview, we provide Figure 2 and Figure 3 depicting numbers for selected modalities[8] and the relationships among them.

4.1 Questionnaires

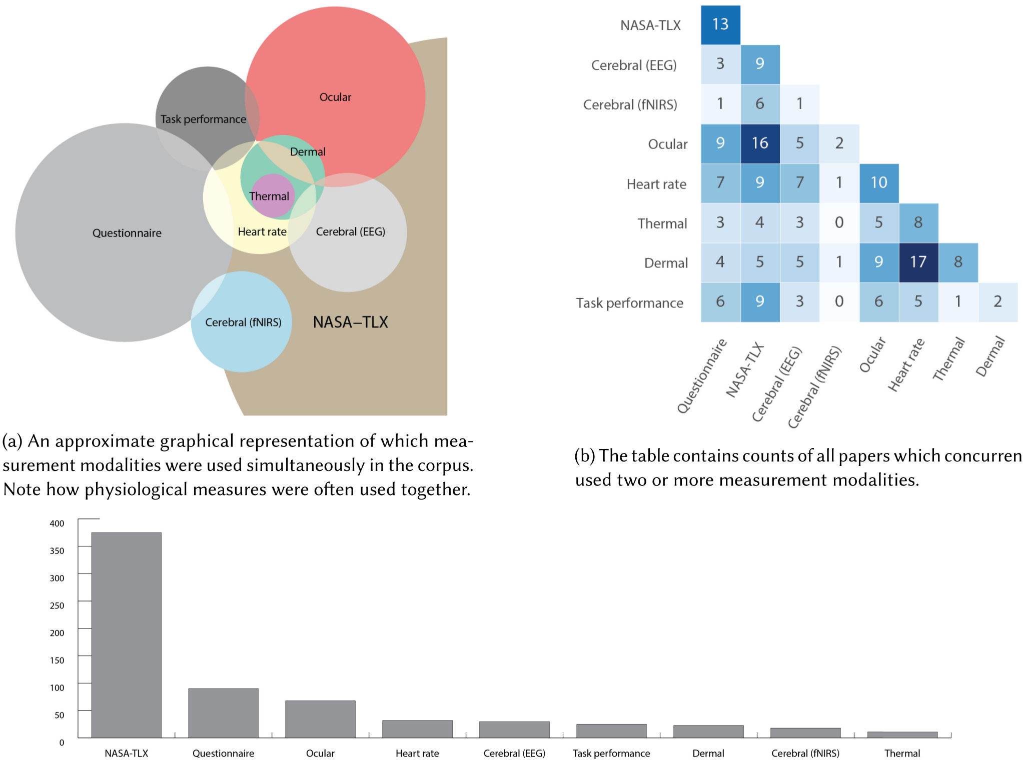
Questionnaires incorporate questions to infer the experienced cognitive workload of a single trial or whole experiment. This is often indicated by a score that is assigned by participants. The magnitude of the score (i.e., lowest and highest score) depends on the design of the questionnaire itself. Questionnaires have two key advantages, which make them popular among HCI researchers: First, they can be easily deployed by asking a participant to fill in questionnaires after an experimental condition. This removes the inconvenience of rather complex workload measures that require a prior rigorous setup. Secondly, they provide a predefined and straightforward procedure to analyze the workload score. The cognitive demand of the presented stimuli is quickly quantified through automatic or manual analysis of workload scores. However, cognitive workload measures based on questionnaires are prone to biases. Questionnaires lack real-time capabilities since they are employed *after* an experimental condition. Hence, these measures aggregate their workload measure of an experimental condition into a single score, making it challenging to measure fluctuations of cognitive workload *during* the experimental condition or user interface interaction. Furthermore, questionnaires are highly susceptible to subjective perception: Users assess questionnaires depending on their perception, so it becomes difficult for the experimenter to interpret the results if high variations of workload scores exist within the same experimental condition. Despite the disadvantages above, questionnaires resemble the most used metric for cognitive workload among HCI researchers and practitioners (see Figure 2). We describe the design of the questionnaires that emerged as relevant throughout the literature review in the following subsections.

4.1.1 NASA Task Load Index. The NASA Task Load Index (NASA-TLX) [90, 91] is a multi-scale questionnaire to assess participants' perceived subjective workload. Its origins date back to the

[6] Including conferences that switched to a journal format.

[7] https://www.thomaskosch.com/cl-paper-library - last access 2023-01-24.

[8] We decided to omit less frequent modalities for visual clarity.



(a) An approximate graphical representation of which measurement modalities were used simultaneously in the corpus. Note how physiological measures were often used together.

	Questionnaire	NASA-TLX	Cerebral (EEG)	Cerebral (fNIRS)	Ocular	Heart rate	Thermal	Dermal
NASA-TLX	13							
Cerebral (EEG)	3	9						
Cerebral (fNIRS)	1	6	1					
Ocular	9	16	5	2				
Heart rate	7	9	7	1	10			
Thermal	3	4	3	0	5	8		
Dermal	4	5	5	1	9	17	8	
Task performance	6	9	3	0	6	5	1	2

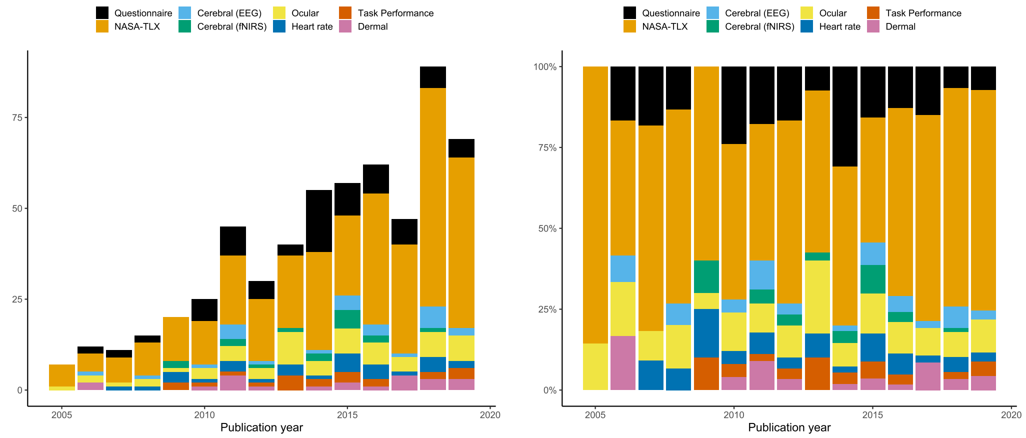
(b) The table contains counts of all papers which concurrently used two or more measurement modalities.

(c) The graph visualizes the total number of times each modality was employed in the corpus. Questionnaire-based methods were most prevalent, with a big emphasis on using the NASA-TLX.

Fig. 2. An overview of the different modalities used to estimate cognitive load in the corpus.

1980s when NASA researchers introduced it to study task load of pilots. Although it was not originally designed to evaluate cognitive workload based on the interaction with a computer, it is the most frequently used questionnaire to measure cognitive workload in our literature corpus (cf. Figures 2 and 3). For the questionnaire, the overall workload is split into six sub-scales: (1) Mental Demand, (2) Physical Demand, (3) Temporal Demand, (4) Performance, (5) Effort, and (6) Frustration. Each sub-scale has an additional description that participants need to read before rating the specific sub-scale to minimize misunderstandings (e.g., "How mentally demanding was the task?" for the *Mental Demand* sub-scale). Sub-scales are horizontal lines divided into 20 intervals with bipolar descriptions on the left and right (i.e., Low/High). Tick marks with increments of 5 allow participants to have ratings between 0 and 100. Researchers can then analyze the overall perceived workload in two ways: (a) with an individual sub-scale weighting, or (b) without an individual sub-scale weighting, often referred to as "Raw NASA-TLX".

4.1.2 Dundee Stress State Questionnaire. The Dundee Stress State Questionnaire (DSSQ) is a questionnaire measuring task-induced stress [167, 168]. Past research used this questionnaire to measure cognitive workload, assuming that stressful situations are associated with negative emotions that lead to higher cognitive demand. The DSSQ employs 11 state factors, which are associated with the three dimensions of *Task Engagement* (factors: Energetic Arousal, Tense Arousal, Hedonic Tone), *Distress* (Intrinsic Motivation), and *Worry* (Self-focus of Attention, Self-Esteem,



(a) Absolute number of papers given publication year for selected modalities.

(b) Individual share of selected modalities normalized per publication year.

Fig. 3. An overview of the eight most prominent modalities broken down by publication year. Note how the overall number of modalities increases over time, yet the distribution of individual shares stays rather constant over time. The metrics are highlighted in bold font in Table 1 excluding questionnaires.

Confidence, and Control, Concentration, Task-related Cognitive Interference, Task-irrelevant Cognitive Interference) of which each scale consists of eight items.

4.1.3 Instantaneous Self-Assessment. The **Instantaneous Self-Assessment (ISA)** [237] is a uni-dimensional subjective workload assessment technique. Participants rate their perceived workload in predefined intervals during the execution of a task – allowing for real-time assessment of cognitive workload. Participants are continually asked to self-rate their perceived workload during a task on a scale of 1 (‘underutilized’) to 5 (‘excessive’). While this can be done using pen and paper, traditionally, an additional keypad consisting of five buttons (each button representing a workload level) and a flashlight (flashing when a rating is requested) is required.

4.1.4 Bedford Workload Scale. The Bedford Workload Scale [205] is a uni-dimensional rating scale to determine free mental capacity. It consists of a hierarchical decision tree, and participants are asked whether (1) it was possible to complete the task, (2) the workload was tolerable, and (3) the workload was satisfactory without reduction. Depending on the decision, participants then have to rank their perceived workload on a rating scale with endpoints from *workload insignificant* (1) to *task abandoned* (10). This value represents the perceived workload of a participant.

4.1.5 Workload Profile. The “Workload Profile” questionnaire [241] combines a *Psychophysical Scale*, the *Bedford Workload Scale*, and a *Workload Profile* to measure cognitive workload. The psychophysical scaling procedure assigns an arbitrary workload score to a reference task. After completing the reference task, the participants are asked to rate the cognitive demand of the other conditions relative to the reference task. There are no numeric restrictions in the workload rating. Hence, each rating is subjectively proportional regarding the difficulty assigned to the reference task. Conceptually similar to the Bedford Workload Scale, the *Workload Profile* collects the proportion of remaining attentional resources after the subject has experienced all tasks. Eight workload dimensions are rated, summed up, and compared to the other two scales.

4.1.6 System Usability Scale. The **System Usability Scale (SUS)** [35] is a ten-item questionnaire for the subjective assessment of usability. Each item is represented as a statement (e.g., "I found the system unnecessarily complex"), and participants use five-point Likert scales to indicate the extent to which they agree with the specific statement (i.e., strongly disagree (1) to strongly agree (5)). The SUS has a predefined procedure to calculate the overall SUS score between 0 and 100. Although the SUS represents a classical subjective assessment method for usability, previous research has argued that cognitive workload can be "calculated indirectly through some of the questions in the SUS" [204] (e.g., "I thought the system was easy to use").

4.1.7 Rating Scale Mental Effort. The **Rating Scale Mental Effort (RSME)** [267] is a single-scale questionnaire with nine workload levels ranging from "not at all hard to do" to "tremendously hard to do." The RSME scale consists of a 150-millimeter vertical line with tick marks from 0 to 150 on the left-hand side and nine workload levels as labels on the right-hand side. Participants rate their workload on this scale. It is noteworthy that the lowest workload classification ("Not at all hard to do") is not represented as value 0, and neither is the highest workload classification ("Tremendously hard to do") represented as value 150. This allows participants to rate their workload beyond the predefined classification.

4.1.8 Driving Activity Load Index. The **Driving Activity Load Index (DALI)** [190] is a multi-scale questionnaire to assess participants' perceived subjective workload. The DALI is an adapted version of the NASA-TLX, yet tailored to the driving task. Similar to the NASA-TLX, the DALI consists of six scales: (1) Effort of Attention, (2) Visual Demand, (3) Auditory Demand, (4) Temporal Demand, (5) Interference, and (6) Situational Stress. Participants rate each sub-scale on a 100-point range with 5-point steps. In line with the NASA-TLX, an individual sub-scale weighting is needed to compose a global workload score.

4.1.9 Cognitive Task Load Model Questionnaire. The "Cognitive Task Load Model" resembles a theoretical model to analyze cognitive workload [179], which was developed explicitly for naval ship control centers. Cognitive workload is formalized as a metric to refine allocated mental resources, which a user interface can utilize to provide adaptive in-situ support. Users are observed and classified into an abstract model that monitors workload utilization rather than assigning scores to conditions. In contrast to the other workload-related questionnaires, the cognitive task load model does not result in a quantitative score but a classification into the framework.

4.1.10 Customized Questionnaires. In addition to the established questionnaires, previous research has incorporated various custom questionnaires to subjectively measure cognitive workload – tailored to their specific use cases or tasks. In addition, they often ask informally about the perceived mental effort (e.g., using Likert scales).

4.2 Cerebral System

The human brain is the information and control unit of a human being. Thereby, approximately 100 billion neurons underlie human cognition [93]. Electrical activity in several brain regions can be measured by placing electrodes on the scalp. Neurons communicate by exchanging electrical activity using neurotransmitters, a chemical substances transferred between neurons. Measuring cerebral activity has a rich history in which scientists experimented with different modalities to capture cortical activity. In 1929, Hans Berger published his first report about cortical electrical measurements that can be captured by placing electrodes on the outer scalp of the brain [144]. Since then, Berger became recognized as a pioneer in the quantification of cerebral activity. Other modalities, such as Functional Near-Infrared Spectroscopy (fNIRS) that investigate the oxygen levels in the blood, or Functional Magnetic Resonance Imaging (fMRI) that measures the electromagnetic responses of the brain, was introduced in the 70s [69] and 90s [229], respectively, to obtain more accurate brain imaging measurements at the cost of temporal resolution. We present the functionality of the brain measurements and their features in the following.

4.2.1 Electroencephalography (EEG). EEG is commonly leveraged in clinical applications and yields a non-invasive method to estimate cortical activity [166, 180, 254]. Electrical potentials between $1\mu v$ and $100\mu v$ (microvolts) are measured by placing conductive electrodes on a scalp. An additional electrode serves as a reference electrode, which can be placed on the earlobe or scalp [163]. The measured EEG potentials allow the processing and extraction of different features. Machine learning has been used on these features to retrieve insights into cognitive processes [155, 156]. EEG has been used to discriminate between cognitive states [71, 83, 148], as a measurement for user experience evaluation [70], and as an input method [14, 238]. For example, changes in electrical potentials are observed by analyzing frequency bands. Previous work found a drop in frequencies of alpha (8–12 Hz) and an increase of theta (4–8 Hz) [122, 129, 130, 135] when subjects had to raise mental capacities. An alternative approach is the assessment of **Event-Related Potentials (ERPs)** to infer mental workload [36, 206]. To complement this, real-time brain visualizations enable deeper insights into sequences of neuronal activity [132].

However, EEG measurements are prone to noise. Head movements, muscle contractions, eye movements, or even eye blinks cause changes in the electrical field on the scalp. Researchers are concerned about this and invest significant effort to reduce the number of measurement artifacts [53, 73, 127, 181]. Since artifacts cannot be completely avoided, EEG often requires a controlled environment comprising minimal body movements by the user. Such conditions are often impractical for end-users due to their high experimental control. However, recent technical advances ameliorate these disadvantages [176], which allows artifact corrections for EEG during mobile settings [29]. The barrier to using EEG in the real world has been lowered by making EEG headsets accessible to the consumer market. These are usually priced between $249[9] and $1600.[10] More affordable open source solutions can be acquired within the OpenEEG Project.[11]

Frequencies. Different oscillations in EEG signals are attributed to several cognitive states. The six bandwidths delta (1 Hertz (Hz)–3 Hz), theta (4 Hz–7 Hz), alpha (8 Hz–12 Hz), lower beta (13 Hz–20 Hz), upper beta (21 Hz–30 Hz), and gamma (31 Hz–100 Hz) power are predominantly employed when analyzing frequencies. Delta power is associated with varying levels of attention, salience, detection, and subliminal perception [125]. According to our literature survey, delta power has not been widely used to measure cognitive workload of computing systems. In contrast, changes in alpha and theta frequencies are correlated with the mental demand placed on working memory and increase or decrease in task engagement [77]. Theta oscillations correlate with the frontal anterior cingulate cortex regarding task engagement the users are experiencing [213]. Increased theta power is associated with higher task engagement, while lower theta power indicates low task engagement. The alpha power is associated with changes in brain resting states. Modulating alpha waves is achieved if the brain is resting, decreasing when users perform tasks requiring their memory. Since it is unknown if correlations between the demand of working memory (i.e., alpha power) and task engagement (i.e., theta power) might elicit a flow state, researchers use the theta-alpha ratio as a metric for cognitive demand [213]. Similarly, beta power is widely used as a metric for mental workload, including alertness, arousal, frustration, engagement, and workload states [141]. In contrast, gamma power indicates high mental activity at the somatosensory cortex [141]. Overall, theta, alpha, and beta power are dominantly used for workload assessments.

[9] https://www.choosemuse.com - last access 2023-01-24.

[10] https://www.futurehealth.org/bm_at1.htm - last access 2023-01-24.

[11] https://www.openeeg.sourceforge.net - last access 2023-01-24.

Event-Related Potentials. Event-Related Potential (ERP)s are specific amplitudes that characteristically appear after users have perceived a stimulus [158]. Different amplitudes may be measured depending on the type of provided stimulus. For example, if a stimulus does not match a user's expectation, a negative peak in the electroencephalography (EEG) signal is measured after 400ms. In contrast, the processing of stimuli is visible after 300ms in the EEG signal. The increase in amplitude is known as P300, a specific ERP component that is suspected to correlate with cognitive workload [36]. Therefore, smaller ERP amplitudes are expected when users perceive stimuli that diversely demand their working memory. Tasks that require a large amount of working memory are more difficult to process, thus resulting in a smaller P300 amplitude. This provides an indicator for working memory placed on users when potentially interacting with a user interface.

4.2.2 Functional Near-Infrared Spectroscopy. Another non-invasive modality to measure brain activity is fNIRS [69]. fNIRS utilizes near-infrared light within a range of 650 *nm* and 1000 *nm* to measure changes in the concentration of **Oxygenated Hemoglobin (HBO)** and **Deoxygenated Hemoglobin (HbR)** in the human brain. The light emitters are placed on the human scalp and measure the outcoming light to affect the oxygen used. While fNIRS is famous for brain sensing in HCI research due to its easy setup [177], latencies exist between stimulus onset and the exact measure. This can make fNIRS challenging to be used with applications that require immediate feedback.

4.2.3 Functional Magnetic Resonance Imaging. fMRI is another non-invasive measure for cortical activity. In contrast to measures of electrical activity, fMRI uses magnetic fields to measure the level of blood oxygen in the brain [66]. fMRI uses electromagnetic waves instead of infrared light to find changes in blood oxygenation. Similar to fNIRS, it takes several seconds to measure changes in cortical activity after stimulus onset when using fMRI. While providing very accurate results, fMRI is a stationary and expensive measurement modality, making it rather unsuited to interactive systems.

4.3 Ocular System

The eyes are part of the human visual system. The eye collects light from the environment, focuses and tracks visual stimuli, and transports these signals to the human brain for further interpretation. In addition, the eye constantly moves – voluntarily or involuntarily – to adjust the light intensity by dilating, contracting the pupil, or focusing and following objects of interest. These eye movements can be recorded in real-time using eye-tracking devices. While stationary eye trackers provide high precision and are usually attached to a display, mobile eye trackers (e.g., as a standalone device or integrated into a head-mounted display) are worn like glasses, providing the possibility to move in an environment freely. Both eye-tracking devices allow for (e.g., gaze-based) interaction with a system or evaluation purposes. For example, websites or interfaces are potential use cases for evaluation with eye trackers. Here, areas of interest and focus sequences can be visualized as, e.g., heat maps or scan paths. However, previous research has shown that various eye movements can also indicate increased cognitive demand [63, 131, 133] or derive the user proficiency during interaction [151].

4.3.1 Pupil Diameter. Changes in pupil diameter are referred to as pupil dilation: This is an involuntary eye movement (i.e., a reflex), and the diameter can range from 1.5mm up to 8mm. Previous research has shown that pupils dilate with increasing task difficulty [47] across tasks like reading, problem-solving, or visual search activities [208], where the cognitive demand can be estimated in real-time using machine learning [131]. However, a significant influence on pupil diameter is the environment. In a darker environment, pupils dilate to acquire more light – pupils contract to reduce the amount of light in a brighter environment. Thus, different methods and models (e.g., [195]) were developed to ensure a valid measurement of cognitive workload using pupil dilation: Here, the total pupil dilation is seen as the sum of dilation due to light intensity and cognitive workload. Knowing the pupil diameter for a given light intensity allows us to calculate the effect due to cognitive workload. The Index of Cognitive Activity (ICA) [164] and the Index of Pupillary Activity (IPA) [62, 63] follow a similar principle to separate light-invoked and cognitive workload-related changes in the pupil diameter.

4.3.2 Saccades. The eye movement that allows for shifting between two fixations (i.e., areas of interest) is called saccades. Saccades are voluntary eye movements that take 30 to 80 ms to complete and are usually visualized as scan paths to, for example, show a sequence of areas of interest. Previous research has shown that the size and speed of saccades are highly discriminatory parameters for an increased cognitive workload: Here, an increased cognitive workload is indicated by larger or faster saccades [47].

4.3.3 Eye Blinks. Eye blinks indicate the perceived workload of a user, whereas repeated involuntary eye blinks are a sign of cognitive fatigue. Previous research has shown that the rate and latency of blinks are related to cognitive workload: Lower blink rates and higher blink latencies indicate an increased cognitive workload [47].

4.3.4 Fixations. Fixations are voluntary eye movements that focus on an area of interest. Previous research has shown that the fixation rate and duration can indicate an increased cognitive workload. The fixation rate of an area of interest describes the number of times an object was looked at – this can be interpreted as a repeated interest in an area and relevant for the current cognitive activity [208]. The fixation duration was found to indicate an increased strain on cognitive workload. Previous research shows that the rate and duration of fixation are indicators of an attention shift due to increasing task complexity [47].

4.3.5 Smooth Pursuit and Ocular Movements. Smooth pursuit eye movements are necessary to follow a moving target closely. Smooth pursuits are voluntary and can be seen as a conscious decision to track a target. Previous research investigated how these movements can be used as an input technique (e.g., target selection) and real-time cognitive workload measurement. Previous work has also shown that smooth pursuit eye movements deviate from a given trajectory in situations with a higher cognitive demand [133] while being less prone to changes in light intensity (cf. pupil diameter).

4.4 Cardiovascular System

Within our autonomic nervous systems (ANS), two[12] opposing systems – the sympathetic (activating) and the parasympathetic (inhibiting) nervous system [145] – control bodily functions, such as heart rate (HR). These changes are reflected in the cardiac cycle, and related metrics, e.g., heart rate variability (HRV) [212]. Among others, external influences such as stress, emotion, and work impact these metrics and can be used to infer those [162]. This section highlights cardiovascular responses evoked by the ANS. Other related responses, such as dermal activity or pupillary response, are covered in their respective sections.

4.4.1 Heart Rate. Techniques such as electrocardiography (ECG) and photoplethysmogram (PPG) are used to monitor the user's heart rate and derived metrics in the time and frequency

[12] Dependent on literature, the enteric nervous system is also included.

domains unobtrusively. Scenarios include, for example, office-work [49], elementary cognitive tasks [87], and automotive environments [244].

While heart rate-related metrics have been shown to indicate a user's cognitive workload, the abundance of external factors that additionally influence our ANS can make this modality ambiguous and warrant close control of confounding variables. Normalization techniques, such as establishing baselines for the sympathetic and parasympathetic nervous system responses [49], as well as relying on multiple modalities, may allow building more robust models [106].

4.4.2 Respiration. Research showed various links between respiratory activities and an increased cognitive workload: Here, respiration belts or thermistor-based flow sensors underneath the participant's nose were used to measure the respiration rate, its variance, and depth [106]. Increased cognitive workload can lead to higher and lower respiration rate variability [99].

4.4.3 Temperature. Here, recording facial and skin temperature via thermal imaging and temperature sensors has been investigated by previous research. A challenge is that temperature measurements significantly lag between stimulus onset and physiological response. Likewise, our temperature is influenced by the ANS. Research using thermal imaging found promising results in distinguishing cognitive load levels using standardized cognitive load tests [2]. However, in lower constraint scenarios, e.g., in a driving context, classification is more difficult [8]. Furthermore, using skin temperature sensors is often accompanied by other on-skin sensors (e.g., ECG, PPG). The availability of intelligent wearables makes these measurements readily available. Hence, models for cognitive workload estimation include multiple metrics derived from ANS responses. Here, temperature-related measures such as heat flux (heat transfer rate) are more robust [87, 217], most likely due to the less noise-prone method to measure them.

We classify temperature measurement as related to the cardiovascular systems as most works in HCI assumed the correlation between increased heart activity, skin temperature, arousal, and cognitive load. This reflects assumptions and partial evidence present in other fields. In physiological computing, Ikehara and Crosby [101] deemed temperature relevant for cognitive load measurement. In the human factors engineering field, temperature was linked to stress [256] and heart rate variability [74]. Thus, our review shows that HCI operates under the premise that skin temperature is correlated with cardiovascular processes induced by cognitive workload.

4.5 Dermal Activity

Another physiological response linked to the autonomic nervous systems is electrodermal activity (EDA), covering all electrical phenomena in skin [31]. There are various recording techniques, such as applying currents to measure the conductance or resistance of the skin. A significant use case in engineering psychology includes detecting different levels of arousal and stress [31].

Measurements of EDA and derived metrics have been successfully applied, including, e.g., user experience evaluation tools [76], detecting office workload [216] and driver-related tasks [202, 217, 224]. Similar to metrics related to heart rate, EDA suffers from ambiguity. External factors, such as body position and emotional stress [31] impact accuracy. Reported results in research fluctuate in terms of the discriminative power of EDA. While some work showed feasible accuracy [182, 202], other work argues that other modalities are superior [87, 201].

4.6 Task Performance

Traditional usability measurements for efficiency or effectiveness, such as time and error, could increase cognitive workload. Here, various studies report on task completion time (e.g., [45, 61]) and reaction time (e.g., [48, 82]) or the number of errors (e.g., [13, 61]) and task accuracy (e.g., [45, 153]) – all highly related to changes in cognitive workload. Additionally, keystroke dynamics and linguistic markers of typed text (i.e., language production) were highly related to cognitive workload [34, 245].

4.7 Haptic Interaction

Different modalities, such as input via pen or touch, proved to be valid discriminatory indicators of changes in cognitive workload. During handwriting with a stylus, the velocity [211, 260], frequency [210], trajectory duration [209], and pressure of strokes [260] were studied as indicators for cognitive workload. Similarly, touch input [39] – especially finger trajectories [173] – were used to measure cognitive workload.

4.8 Speech

Following a think-aloud protocol in a user, the study allows researchers to analyze reasons for participants' behavior, mood, or expectations. Additionally, previous research has shown that participants' spoken words during an evaluation can be used to analyze the cognitive workload. Here, speech features such as the flow of words (e.g., tempo [42] or pauses [117, 118]), but also more nuanced features such as the lexical density [115], the spectral domain [259], or the pitch contour [39] while talking were shown to be indicators of changes in participants' cognitive workload. Additionally, linguistic features were analyzed to study participants' cognitive workload [116, 246].

4.9 Body Movements

So far, previous research has investigated (1) postural behavior and gestures to assess phases of cognitive underload or overload [95], (2) mouse movements such as traveled distances [12] or pause/break activities [119], and (3) differences in glance duration at input modalities and output devices to solve a given task [61]. In addition, specialized measurements techniques, such as Electromyography (EMG), which records the electrical activity of muscles [110, 111, 171], can assess co-occurring physical demand to decompose cognitive load factors, such as the impact of motor memory [139].

4.10 Biomarkers

Biomarkers (or biological markers) are measurable indicators of a medical state [230]. More generally, biomarkers can be used to indicate the physiological state. These markers play a significant role in medicine, e.g., stress biomarkers in behavioral therapy. Typical markers include cortisol, alpha-amylase, and pro-inflammatory cytokines for different biological stress systems[13] [178]. However, despite their distinctiveness, biomarkers have not been widely researched within the area of HCI. One primary reason is the intricate process of obtaining and analyzing biomarkers, often requiring lab work and physicians to draw samples. There is potential for accurate ground-truth labeling. A careful evaluation of the actual measurement is inevitable, as external stressors, e.g., the social-evaluative threat of performing well in a study, may impact biomarkers [49].

Salivary cortisol, usually used as a biological marker for stress [114], has recently been utilized for cognitive workload measurements [49]. Cortisol levels are measured from salivary probes that require laboratory analysis. While the study of salivary cortisol is a reliable indicator for stress and workload [120], the need for a laboratory analysis restricts the real-time capabilities of interactive systems. However, cortisol is a reliable measure of cognitive workload and represents a well-grounded measure in clinical research.

[13] Hypothalamic-pituitary-adrenal (HPA) axis, autonomic nervous system (ANS), and immune systems, respectively.

5 USING EVALUATION METHODS FOR COGNITIVE WORKLOAD IN HCI STUDIES

While our review demonstrates current challenges in understanding cognitive workload aspects in HCI, it also surfaced examples where the concept of cognitive workload was used effectively. In our review, we found 579 papers that used cognitive workload measurements in an HCI context. Most importantly, past work can guide if, how, and when to use cognitive workload measures. This section aims to provide a compiled step-by-step procedure (cf. Figure 4) on choosing suitable metrics, complemented through Table 1 categorizing the papers in the corpus in terms of measurement modalities and metrics used.[14] We suggest that researchers and practitioners deciding on the cognitive workload metrics for their studies first consult the procedure (cf. Figure 4) in Section 5.1 to get an overview of specific methods from related work in Table 1, and details on the particular applicability, advantages, and limitations of metrics in Section 4. The referenced papers in Table 1 and our interactive paper library can then be investigated to get more details on a selected metric.

5.1 Choosing Suitable Metrics for HCI Studies

Based on our understanding of the corpus of work included in this review, we propose a four-step procedure (illustrated in Figure 4), which can help researchers to choose optimal metrics for measuring cognitive workload in future HCI studies. We note that our procedure should only be used to select candidate modalities for use in studies of interfaces and concerns primarily the design goals of the system to be studied. The procedure serves primarily as a thinking tool for understanding the possibilities for measuring cognitive workload induced by an interactive system. The final choice of modality should be based on good research practice and the requirements for ecological validity in conducting studies using a particular modality.

5.1.1 Step One: Are you Actually Evaluating Cognitive Workload? When considering the measurement of cognitive workload in a user study, the first step is verifying if cognitive workload is the right concept. We observed that many papers equated cognitive workload to usability, flow, or design quality throughout the corpus. Thus, we propose double-checking if the expected difference between the experimental conditions in cognitive workload would not be a product of a different underlying concept. For instance, if the interface is frustrating due to low usability in one of the conditions, measuring cognitive workload may obscure the underlying cause of the difference. In other words, when researchers decide to measure cognitive workload, they must ensure that possible sources of cognitive workload can be reliably identified.

5.1.2 Step Two: Do you Expect Cognitive Workload to Vary During System Usage? If a researcher decides to measure cognitive workload, the next step is to choose the metrics and measurement instruments. The corpus in this review shows that researchers most likely prefer questionnaires due to their practical benefits – they are quick to administer and require a single interaction. In addition, the high number of past papers using questionnaires implies that using a questionnaire offers the advantage of comparing one's work with a large body of past research. There are, however, established drawbacks to this method. Questionnaires usually provide only one measurement point per condition and rely on retrospection. Further, they can distract the user from interacting with the system. However, the wide use of questionnaires shows that, in many cases, the benefits outweigh the disadvantages. Thus, if a single-point, summative assessment of cognitive workload is enough to verify the hypotheses in the study, a questionnaire is an appropriate choice. If richer,

[14] We do not list papers that exclusively use questionnaires to assess cognitive workload in their studies. For a complete list of papers, including those that use questionnaires, consult our interactive paper library: www.thomaskosch.com/cl-paper-library - last access 2023-01-24.

Table 1. Papers in the Review Corpus Classified in Terms of the Cognitive Workload Metrics used in the Reported Studies

Type	Metric	References
Cerebral	**EEG** Frequencies	[16, 17, 26, 52, 67, 70, 83, 86, 87, 100, 105, 106, 124, 130, 135, 138, 148, 150, 184, 197, 198, 207, 235, 243, 244, 268]
	EEG ERPs	[23, 26, 59, 81, 184, 214, 227, 243]
	Hemodynamic Response (fNIRS)	[3, 5, 32, 80, 96, 97, 104, 159, 161, 191, 192, 196, 221–223, 227, 261, 262]
	Electromagnetic Changes (fMRI)	[220]
Ocular	Pupil Diameter	[4, 5, 9–11, 18–22, 27, 28, 39, 44–47, 55, 58, 62, 63, 72, 87, 92, 94, 99, 103, 107, 108, 112, 142, 143, 153, 154, 157, 165, 175, 187–189, 195, 199–201, 207, 215, 227, 236, 239, 248, 249, 255, 257, 266]
	Saccades	[10, 21, 27, 44–47, 76, 87, 99, 165, 184, 193, 228, 236, 258]
	Eye Blinks	[4, 5, 27, 43–47, 76, 87, 99, 175, 182, 184, 236, 268]
	Fixations	[1, 10, 54, 68, 99, 140, 153, 184, 219, 228]
	Smooth Pursuits and Ocular Movements	[133]
Cardiovascular	**Heart Rate**	[4, 10, 25, 27, 40, 41, 49, 50, 76, 78, 82, 85, 87, 99, 106, 146, 152, 160, 169, 184, 185, 194, 201, 202, 214, 216, 217, 224, 225, 227, 240, 244, 264, 264, 266, 268]
	Respiration	[87, 99, 106, 184, 185, 194]
	Temperature	[1, 2, 8, 87, 184, 185, 194, 216, 217, 268]
Dermal	Skin Conductance (EDA)	[25, 50, 51, 76, 78, 87, 99, 106, 134, 182–185, 194, 201, 202, 215–217, 224, 225, 227, 268]
Task Performance	Number of Errors	[13, 30, 61, 113, 266]
	Task Completion Time	[7, 45, 61, 68, 153, 184, 242]
	Task Accuracy	[45, 153]
	Reaction Time	[48, 79, 82]
	Keystroke Dynamics	[34, 227, 245]
	Language Production	[34, 245]
	Real-Time Self-Report	[226]
Haptic	Pen Input	[209–211, 260]
	Touch Input	[39, 102, 173, 247]
Speech	Lexical Density	[115]
	Tempo	[42]
	Pauses	[117, 118]
	Spectral Domain	[259]
	Pitch Contour	[39]
	Linguistic Features	[116, 246]
Body Movement	Muscle Contractions (EMG)	[41, 201]
	Glance Duration	[61]
	Mouse Movement	[12, 119]
	Postural Behavior and Gestures	[95, 227]
Biomarkers	Salivary Cortisol	[49]

The bold values refer to the most used and prominent cognitive workload measurement modalities highlighted in Figures 3 and 4. We did not include questionnaires in the table due to a large number of entries.

formative data is needed, researchers should use other methods. The procedure ends with this step if a questionnaire is chosen.

5.1.3 Step Three: Does your System Need to Adapt to Changes in the User's Cognitive Workload? Having decided that the study will use more granular measures than questionnaires for cognitive workload, we suggest that researchers inquire if the system will use cognitive workload to learn about the implications and improve the interface design or as user input itself, i.e., whether the study will involve elements of the studied systems changing their parameters based on measured cognitive workload. Our corpus showed that such systems are increasingly present in HCI literature. In cognitive workload-based adaptation, the metric considerations may primarily stem from how the system is implemented. In such systems, the recommendation is to **Use modalities that offer robust data that can be processed quickly**. For example, cognitive workload data may need to be processed rapidly for the system to adapt interactively. The input should also be easily interpreted so the system can adapt algorithmically. In contrast, if the system does not need to react to the current cognitive workload, researchers can choose complex, data-intensive metrics that can be analyzed and interpreted after completing the study. Thus, when cognitive workload is not used, for adaptation, we recommend to **Use modalities that offer rich data for detailed analysis**. This may offer additional insight into the workload induced by the system.

5.1.4 Step Four: Choose a Modality Based on the Considerations. The final step in the procedure is choosing a metric of the required richness and complexity. The selected metric must provide a volume of data that can be effectively processed and offer the necessary richness to understand a given system. Researchers can also employ multiple metrics to build hybrid assessments of cognitive workload. This, however, appears to be complex as the fraction of papers using such procedures in our corpus is low (see Figure 2). To facilitate this difficult choice, we provide a classification for the most common measurement modalities in terms of increasing complexity (i.e., data volume) and decreasing speed for the last step of this procedure (see Figure 4). An informed choice of modalities can best be made by studying the examples provided for each modality in Table 1.

5.2 A Note on Selecting Questionnaires

If a researcher decides to end the procedure in step two and use a questionnaire, they can choose which questionnaire to use. Our corpus shows that the NASA-TLX is significantly more prevalent in HCI research and may often appear to be the optimal choice. However, throughout this paper, we noted that the reasons for the scale's success are unclear. The primary reason behind the use of NASA-TLX appears to be community convention (cf. Section 6.2). Consequently, we would like to encourage HCI authors to consider alternative scales that may offer advantages over the NASA-TLX. For example, Zijlstra's [267] RSME is a single-item rating and may enable rapid evaluations or be used in experience sampling; the ISA scale [237] can provide near real-time assessment; while the Bedford Workload Scale [205] can provide insights into potential mental overload situations for the user—all roles for which the NASA-TLX is not well suited. Further, one should not forget that the use of certain metrics places strict requirements on the experimental design, e.g., block designs are usually recommended for cerebral modalities. Thus, the final design of a user study that uses cognitive workload measurement is almost always a trade-off between the ecological validity of the task and the requirements posed by the cognitive workload measurement modality.

6 FUTURE DIRECTIONS AND CHALLENGES OF COGNITIVE WORKLOAD RESEARCH IN HCI

Research in HCI has a broad and, at times, vague understanding of the concept of cognitive workload, as evident from the results of our review. We presented current practices, related

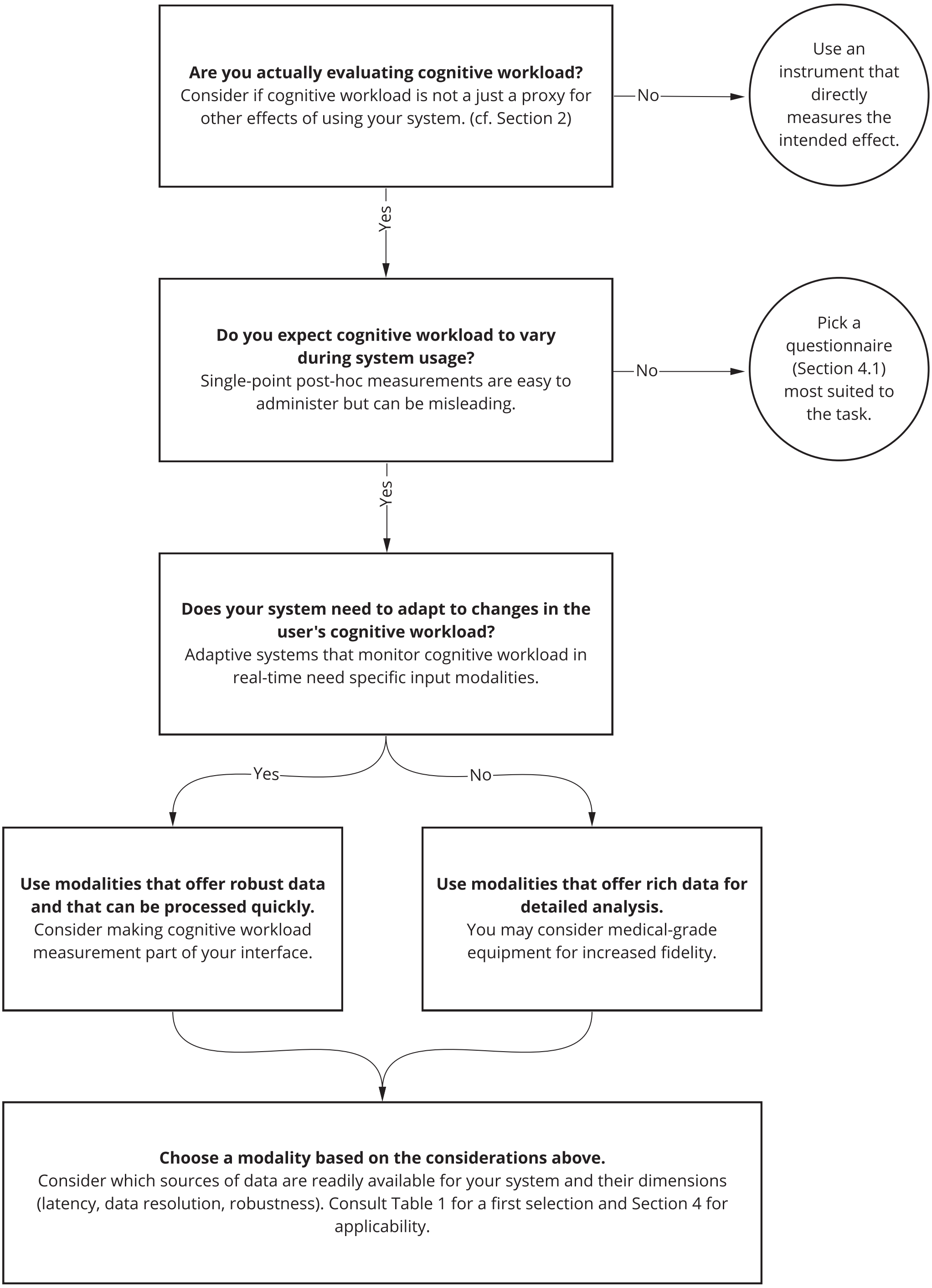

Are you actually evaluating cognitive workload?
Consider if cognitive workload is not a just a proxy for other effects of using your system. (cf. Section 2)
No
Use an instrument that directly measures the intended effect.
Yes
Do you expect cognitive workload to vary during system usage?
Single-point post-hoc measurements are easy to administer but can be misleading.
No
Pick a questionnaire (Section 4.1) most suited to the task.
Yes
Does your system need to adapt to changes in the user's cognitive workload?
Adaptive systems that monitor cognitive workload in real-time need specific input modalities.
Yes
No
Use modalities that offer robust data and that can be processed quickly.
Consider making cognitive workload measurement part of your interface.
Use modalities that offer rich data for detailed analysis.
You may consider medical-grade equipment for increased fidelity.
Choose a modality based on the considerations above.
Consider which sources of data are readily available for your system and their dimensions (latency, data resolution, robustness). Consult Table 1 for a first selection and Section 4 for applicability.

Fig. 4. A step-by-step procedure to select appropriate cognitive workload measurements. HCI researchers can use this procedure as a reference for future studies.

measurement modalities, and metrics often used in contemporary HCI research (cf. Section 4). Additionally, our step-by-step guide on choosing suitable metrics (cf. Section 5) provides a functional categorization for researchers to choose a suitable cognitive workload measurement modality. However, our review has also showcased open research gaps, where choosing the correct metric can still be suboptimal depending on the research question. Choosing the right assessment for the right task has been recently discussed in various research publications (cf. [15, 60, 64, 65, 126, 186, 265]), but especially HCI research needs to deepen and conceptualize our understanding of cognitive workload in interacting with computer systems.

6.1 Defining and Interpreting Cognitive Workload in HCI

The results of our review show that cognitive workload is perceived as a qualitative component of an interactive system. The presented interfaces are designed to handle the user's cognitive resources economically. This is communicated as an explicit design goal or, more commonly, implied while evaluating user interfaces. In most reviewed papers, cognitive workload is used as an implicit concept. It is assumed that the reader is familiar with it. Our review showed that our search terms initially resulted in many papers that mention cognitive workload as a justification for specific system design choices (see Section 3.4). Consequently, low cognitive workload is often treated as equivalent to high system quality without necessarily determining what features or properties of the system's design lead to a perception of quality.

The current concept of cognitive workload in HCI is generalized without taking the individual, temporally changing, cognitive resources of the user into account. Further, as there is no definition of workload widely recognized by the community, researchers should define their understanding of cognitive workload, mainly if the concept is used to explain study results:

> **Research Gap 1:** HCI implicitly uses cognitive workload without discussing which workload components are measured (e.g., working memory, visual imagery). HCI research should engage with the underlying theories of workload to improve how empirical results are interpreted.

We suggest that researchers define their understanding of cognitive workload in cases where cognitive workload levels are reported. Consequently, researchers must elaborate on how they expect different interfaces or interface adaptations to affect cognitive workload. **Thus, it remains a challenge to study how different usability components and usability-related qualities of interactive systems are linked to different types of cognitive workload measured by various methods.** There is a need to evaluate novel systems in terms of both cognitive workload and usability using robust, multi-dimensional measures. In that case, **the HCI field can gradually build a meta-understanding of the relation between usability and cognitive workload**.

6.2 The Hidden Cost of the NASA-TLX—A Legacy Issue?

A key finding of our literature review is that the HCI field prominently relies on using questionnaires and particularly on the NASA-TLX questionnaire for assessing cognitive workload. Previous work suggests combining the NASA-TLX questionnaire with other measures to increase the meaningfulness while improving the interpretation of cognitive workload measures [64]. One could risk stating that the NASA-TLX questionnaire has become a local standard for the HCI research community. This is likely due to historical reasons since many HCI pioneers had roots in human factors engineering. The NASA-TLX was prominently used in human factors engineering in the nascent days of HCI. Since then, the HCI community has applied the scale in various modalities and contexts, leading to individual deviations in the subjective interpretation of the NASA-TLX. Although the NASA-TLX questionnaire is affected by several drawbacks, it has remained a frequently used workload measure throughout the past decades. Potential reasons for the continued widespread use of the NASA-TLX include its simple usage and analysis of the questionnaire data. In addition, the fact that NASA-TLX results can be rapidly obtained facilitates its use in comparative studies. Despite these benefits, reliability and replicability are limited when sampling individuals. Our results show that using the NASA-TLX could be interpreted as an academic tradition, despite being introduced to the HCI field without an extensive explanation [33]. However, given the ever-growing number of research contributions in the HCI field, it is worthwhile to critically reflect on how the HCI field uses and interprets the results of the NASA-TLX questionnaire.

Our review shows that convenience and speed are key advantages of the NASA-TLX and questionnaires in general. In particular, the unweighted variant of the NASA-TLX, commonly known as raw NASA-TLX, provides a quick assessment of perceived task load, including cognitive workload. Researchers often use it in repeated-measures study designs where participants must complete the questionnaire multiple times. If practical considerations are the primary motivation for choosing a cognitive workload evaluation method, one might wonder if convenience does not negatively impact how the evaluation is conceptualized; mainly since in most studies in our corpus, only the unweighted scores (raw NASA-TLX) are analyzed. Such a choice offers the convenience of not hypothesizing which components examined in the NASA-TLX (cf. Section 4.1.1) are crucial in a given system. Instead, **the scale is employed as an accurate universal measure of cognitive workload *per se***. Furthermore, the NASA-TLX questionnaire is susceptible to individual biases through user expectations, leading to improved assessments of user interfaces by belief and placebos [136]. It remains a challenge for the HCI community to further reflect on *if* and *how* the workload induced by a novel interactive system can be accurately measured with a scale that is more than 30 years old and not designed for HCI research:

> **Research Gap 2:** HCI still relies on scales for measuring cognitive workload adapted from another field in the early days of HCI. The validity of these scales should be reassessed. We must develop new measurement instruments for cognitive workload specific to computer interaction.

Consequently, future researchers should be wary that the NASA-TLX is an inherited rather than an efficient tool. Our review found no empirical evidence to determine the range of artifacts or experiences to which the NASA-TLX would be applicable. This implies that **we currently do not know how correct the NASA-TLX is in estimating cognitive workload in HCI**. Hence, **future researchers should be cautious in ascribing explanatory properties to the scale's dimensions**. Interestingly, the NASA-TLX often serves as a *fallback* or baseline measurement when assessing cognitive workload with other metrics (e.g., pupil diameter) to see how reliable a novel measurement is. Instead, future research should revert this methodology by first studying cognitive workload via robust physiological measurements to (1) validate the dimensions of the NASA-TLX, and (2) identify how user studies have to be designed to minimize the biases of the NASA-TLX and questionnaires in general. We envision this direction ultimately leading to an adapted variant of the NASA-TLX that better fits the HCI domain as a pacing and evolving research area.

6.3 HCI Research Has to Become a Catalyst for Creating Workload-Aware Systems

The presented literature review reveals that the concept of cognitive workload is inherently relevant to HCI research. However, HCI researchers still largely rely on suboptimal evaluation methods adapted from adjacent research domains, often failing to capture the multi-faceted nature of interactive systems in HCI research. We argue that HCI researchers are equipped with the tools to explore further the role of cognitive workload in interacting with computers beyond measuring performance and usability. Integrating cognitive workload as an input parameter for workload-aware systems should be at the heart of HCI's expertise. Understanding and using cognitive workload in interactive systems can effectively support users in real-time when performing their tasks. How the design of such support (e.g., through adaptive workload-aware mixed reality interfaces [123, 137]) would look like remains a question for future research. Thus, HCI should not be content with a static evaluation of interfaces but look at how we can create workload-aware systems that make best use of the user's current cognitive resources:

> **Research Gap 3:** Research in HCI needs to capitalize on its opportunities and create interactive systems utilizing cognitive workload as an input parameter for workload-aware systems. The field should innovate by combining established methods (i.e., questionnaires) with novel measures (i.e., physiological sensing) to gain detailed insights into the user's cognitive demand.

Our literature review showed that **physiological measures can provide a detailed, real-time understanding of the user's cognitive processing**. Consequently, HCI should lead in creating interactive systems capable of supporting an individual's optimal point for cognitive engagement. We strongly believe that real-time cognitive workload sensing capabilities will be essential to achieve this goal.

7 CONCLUSION

This paper presents a literature review of current practices measuring cognitive workload in Human-Computer Interaction (HCI) studies. Our survey found 579 relevant papers that assess cognitive workload using questionnaires, physiological sensing, and human behavior. Our results show that the contextualization and selecting suitable cognitive workload measures is a challenging subject in HCI research. We present a step-by-step procedure for selecting a suitable cognitive workload measurement modality for their studies to guide researchers. Our survey reveals three research gaps in the current landscape of HCI research on cognitive workload: improving the definition, broadening the choice of measures, and developing more workload-aware systems. The gaps intend to stimulate future research on cognitive workload assessments in HCI. Future researchers and user experience designers will determine how cognitive workload measurements can be integrated into their interaction paradigms and evaluation settings. Our review shows that questionnaires are a popular method to assess cognitive workload due to their easy deployment, use, and analysis. However, the insights into cognitive processes obtained using questionnaires could be more extensive. In contrast, more advanced cognitive workload metrics (e.g., through physiological sensing or user behavior) require user-friendly methods, interfaces, and deployment strategies to be adopted in future research. Our work provides a structured overview of current cognitive workload measurements and the means for future researchers and designers to choose an appropriate metric for their particular study or design. Yet, this literature survey makes evident that HCI lacks a proper understanding of the concepts underlying cognitive workload. Thus, research towards tailored frameworks and theories is necessary to consolidate the understanding of cognitive workload in HCI.

REFERENCES

[1] Yomna Abdelrahman, Anam Ahmad Khan, Joshua Newn, Eduardo Velloso, Sherine Ashraf Safwat, James Bailey, Andreas Bulling, Frank Vetere, and Albrecht Schmidt. 2019. Classifying attention types with thermal imaging and eye tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 69 (Sept. 2019). https://doi.org/10.1145/3351227

[2] 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, Article 33 (Sept. 2017). https://doi.org/10.1145/3130898

[3] Daniel Afergan, Evan M. Peck, Erin T. Solovey, Andrew Jenkins, Samuel W. Hincks, Eli T. Brown, Remco Chang, and Robert J. K. Jacob. 2014. Dynamic difficulty using brain metrics of workload. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*. Association for Computing Machinery, Toronto, Ontario, Canada, 3797–3806. https://doi.org/10.1145/2556288.2557230

[4] Muneeb Imtiaz Ahmad, Ingo Keller, David A. Robb, and Katrin S. Lohan. 2020. A framework to estimate cognitive load using physiological data. *Personal and Ubiquitous Computing* (2020), 1–15. https://doi.org/10.1007/s00779-020-01455-7

[5] Oliver Amft, Florian Wahl, Shoya Ishimaru, and Kai Kunze. 2015. Making regular eyeglasses smart. *IEEE Pervasive Computing* 14, 3 (July 2015), 32–43. https://doi.org/10.1109/MPRV.2015.60

[6] E. W. Anderson, K. C. Potter, L. E. Matzen, J. F. Shepherd, G. A. Preston, and C. T. Silva. 2011. A user study of visualization effectiveness using EEG and cognitive load. *Computer Graphics Forum* 30, 3 (2011), 791–800. https://doi.org/10.1111/j.1467-8659.2011.01928.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2011.01928.x.

[7] Vicki Antrobus, Gary Burnett, and Lee Skrypchuk. 2016. "Turn Left at the Fairham Pub" Using navigational guidance to reconnect drivers with their environment. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (Automotive'UI 16)*. ACM, Ann Arbor, MI, USA, 35–42. https://doi.org/10.1145/3003715.3005392

[8] Bernhard Anzengruber and Andreas Riener. 2012. "FaceLight": Potentials and drawbacks of thermal imaging to infer driver stress. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'12)*. ACM, Portsmouth, New Hampshire, USA, 209–216. https://doi.org/10.1145/2390256.2390292

[9] Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci. 2018. Cross-subject workload classification using pupil-related Measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA'18)*. ACM, New York, NY, USA, 4:1–4:8. https://doi.org/10.1145/3204493.3204531 event-place: Warsaw, Poland.

[10] Tobias Appel, Natalia Sevcenko, Franz Wortha, Katerina Tsarava, Korbinian Moeller, Manuel Ninaus, Enkelejda Kasneci, and Peter Gerjets. 2019. Predicting cognitive load in an emergency simulation based on behavioral and physiological measures. In *2019 International Conference on Multimodal Interaction (ICMI'19)*. Association for Computing Machinery, New York, NY, USA, 154–163. https://doi.org/10.1145/3340555.3353735

[11] Stephanie Arevalo, Stanislaw Miller, Martha Janka, and Jens Gerken. 2019. What's behind a choice? Understanding modality choices under changing environmental conditions. In *2019 International Conference on Multimodal Interaction (ICMI'19)*. Association for Computing Machinery, New York, NY, USA, 291–301. https://doi.org/10.1145/3340555.3353717

[12] Syed Arshad, Yang Wang, and Fang Chen. 2013. Analysing mouse activity for cognitive load detection. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration (OzCHI'13)*. ACM, Adelaide, Australia, 115–118. https://doi.org/10.1145/2541016.2541083

[13] Amna Asif and Susanne Boll. 2010. Where to turn my car?: Comparison of a tactile display and a conventional car navigation system under high load condition. In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'10)*. ACM, Pittsburgh, Pennsylvania, 64–71. https://doi.org/10.1145/1969773.1969786

[14] Jonas Auda, Uwe Gruenefeld, Thomas Kosch, and Stefan Schneegass. 2022. The butterfly effect: Novel opportunities for steady-state visually-evoked potential stimuli in virtual reality. In *Proceedings of the 3rd Augmented Humans International Conference (AHs'22)*. ACM, New York, NY, USA. https://doi.org/10.1145/3519391.3519397

[15] Paul Ayres, Joy Yeonjoo Lee, Fred Paas, and Jeroen van Merriënboer. 2021. The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology* 12 (2021). https://doi.org/10.3389/fpsyg.2021.702538

[16] Ashwin Ramesh Babu, Akilesh Rajavenkatanarayanan, James Robert Brady, and Fillia Makedon. 2018. Multimodal approach for cognitive task performance prediction from body postures, facial expressions and EEG signal. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data (MCPMD'18)*. Association for Computing Machinery, New York, NY, USA, Article 14. https://doi.org/10.1145/3279810.3279849

[17] Sarune Baceviciute, Aske Mottelson, Thomas Terkildsen, and Guido Makransky. 2020. Investigating representation of text and audio in educational VR using learning outcomes and EEG. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376872

[18] Brian P. Bailey and Chris W. Busbey. 2006. TAPRAV: A tool for exploring workload aligned to models of task execution. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI'06)*. ACM, Venezia, Italy, 467–470. https://doi.org/10.1145/1133265.1133360

[19] Brian P. Bailey, Chris W. Busbey, and Shamsi T. Iqbal. 2007. TAPRAV: An interactive analysis tool for exploring workload aligned to models of task execution. *Interacting with Computers* 19, 3 (01 2007), 314–329. https://doi.org/10.1016/j.intcom.2007.01.004 arXiv:http://oup.prod.sis.lan/iwc/article-pdf/19/3/314/2392213/iwc19-0314.pdf.

[20] Brian P. Bailey and Shamsi T. Iqbal. 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Trans. Comput.-Hum. Interact.* 14, 4, Article 21 (Jan. 2008), 28 pages. https://doi.org/10.1145/1314683.1314689

[21] Mike Bartels and Sandra P. Marshall. 2006. Eye tracking insights into cognitive modeling. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications (ETRA'06).* ACM, New York, NY, USA, 141–147. https://doi.org/10.1145/1117309.1117358

[22] Roman Bednarik, Piotr Bartczak, Hana Vrzakova, Jani Koskinen, Antti-Pekka Elomaa, Antti Huotarinen, David Gil de Gómez Pérez, and Mikael von und zu Fraunberg. 2018. Pupil size as an indicator of visual-motor workload and expertise in microsurgical training tasks. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA'18).* ACM, New York, NY, USA, 60:1–60:5. https://doi.org/10.1145/3204493.3204577 event-place: Warsaw, Poland.

[23] Chris Berka, Daniel J. Levendowski, Milenko M. Cvetinovic, Miroslav M. Petrovic, Gene Davis, Michelle N. Lumicao, Vladimir T. Zivkovic, Miodrag V. Popovic, and Richard Olmstead. 2004. Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction* 17, 2 (June 2004), 151–170. https://doi.org/10.1207/s15327590ijhc1702_3

[24] Nigel Bevan and Miles MacLeod. 1994. Usability measurement in context. *Behaviour & Information Technology* 13, 1-2 (Jan. 1994), 132–145. https://doi.org/10.1080/01449299408914592

[25] Rhushabh Bhandari, Avinash Parnandi, Eva Shipp, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2015. Music-based respiratory biofeedback in visually-demanding tasks. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Edgar Berdahl and Jesse Allison (Eds.). Louisiana State University, Baton Rouge, Louisiana, USA, 78–82.

[26] Maneesh Bilalpur, Mohan Kankanhalli, Stefan Winkler, and Ramanathan Subramanian. 2018. EEG-based evaluation of cognitive workload induced by acoustic parameters for data sonification. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI'18).* ACM, New York, NY, USA, 315–323. https://doi.org/10.1145/3242969.3243016 event-place: Boulder, CO, USA.

[27] Pradipta Biswas, Varun Dutt, and Pat Langdon. 2016. Comparing ocular parameters for cognitive load measurement in eye-gaze-controlled interfaces for automotive and desktop computing environments. *International Journal of Human-Computer Interaction* 32, 1 (Jan. 2016), 23–38. https://doi.org/10.1080/10447318.2015.1084112

[28] Pradipta Biswas and Pat Langdon. 2015. Multimodal intelligent eye-gaze tracking system. *International Journal of Human-Computer Interaction* 31, 4 (April 2015), 277–294. https://doi.org/10.1080/10447318.2014.1001301

[29] Sarah Blum, Stefan Debener, Reiner Emkes, Nils Volkening, Sebastian Fudickar, and Martin G. Bleichner. 2017. EEG Recording and Online Signal Processing on Android: A Multiapp Framework for Brain-Computer Interfaces on Smartphone. https://doi.org/10.1155/2017/3072870

[30] Susanne Boll, Amna Asif, and Wilko Heuten. 2011. Feel your route: A tactile display for car navigation. *IEEE Pervasive Computing* 10, 3 (July 2011), 35–42. https://doi.org/10.1109/MPRV.2011.39

[31] Wolfram Boucsein. 2012. *Electrodermal Activity, (2nd ed.).* Springer Science + Business Media, New York, NY, USA. Pages: xviii, 618. https://doi.org/10.1007/978-1-4614-1126-0.

[32] Mark Boyer, M. L. Cummings, Lee B. Spence, and Erin T. Solovey. 2015. Investigating mental workload changes in a long duration supervisory control task. *Interacting with Computers* 27, 5 (05 2015), 512–520. https://doi.org/10.1093/iwc/iwv012 arXiv:http://oup.prod.sis.lan/iwc/article-pdf/27/5/512/5112816/iwv012.pdf.

[33] Stephen A. Brewster, Peter C. Wright, and Alistair D. N. Edwards. 1994. The design and evaluation of an auditory-enhanced scrollbar. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'94).* Association for Computing Machinery, Boston, Massachusetts, USA, 173–179. https://doi.org/10.1145/191666.191733

[34] David Guy Brizan, Adam Goodkind, Patrick Koch, Kiran Balagani, Vir V. Phoha, and Andrew Rosenberg. 2015. Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. *International Journal of Human-Computer Studies* 82 (2015), 57–68. https://doi.org/10.1016/j.ijhcs.2015.04.005

[35] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

[36] Anne-Marie Brouwer, Maarten A. Hogervorst, Jan B. F. van Erp, Tobias Heffelaar, Patrick H. Zimmerman, and Robert Oostenveld. 2012. Estimating workload using EEG spectral power and ERPs in the n-back task. 14 pages. https://doi.org/10.1088/1741-2560/9/4/045008

[37] Andreas Bulling and Thorsten O. Zander. 2014. Cognition-aware computing. *IEEE Pervasive Computing* 13, 3 (2014), 80–83. https://doi.org/10.1109/MPRV.2014.42

[38] John T. Cacioppo, Louis G. Tassinary, and Gary Berntson. 2007. *Handbook of Psychophysiology.* Cambridge University Press.

[39] Davide Maria Calandra, Antonio Caso, Francesco Cutugno, Antonio Origlia, and Silvia Rossi. 2013. CoWME: A general framework to evaluate cognitive workload during multimodal interaction. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI'13)*. ACM, Sydney, Australia, 111–118. https://doi.org/10.1145/2522848.2522867

[40] Yujia Cao, Mariët Theune, and Anton Nijholt. 2009. Modality effects on cognitive load and performance in high-load information presentation. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI'09)*. ACM, New York, NY, USA, 335–344. https://doi.org/10.1145/1502650.1502697

[41] Daniel Chen, Jamie Hart, and Roel Vertegaal. 2007. Towards a physiological model of user interruptability. In *Human-Computer Interaction – INTERACT 2007*, Cécilia Baranauskas, Philippe Palanque, Julio Abascal, and Simone Diniz Junqueira Barbosa (Eds.), Vol. 4663. Springer Berlin, Berlin, 439–451. https://doi.org/10.1007/978-3-540-74800-7_39

[42] Rui Chen, Tiantian Xie, Yingtao Xie, Tao Lin, and Ningjiu Tang. 2016. Do speech features for detecting cognitive load depend on specific languages?. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*. ACM, Tokyo, Japan, 76–83. https://doi.org/10.1145/2993148.2993149

[43] Siyuan Chen and Julien Epps. 2013. Blinking: Toward wearable computing that understands your current task. *IEEE Pervasive Computing* 12, 3 (July 2013), 56–65. https://doi.org/10.1109/MPRV.2013.45

[44] Siyuan Chen and Julien Epps. 2014. Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction* 29, 4 (July 2014), 390–413. https://doi.org/10.1080/07370024.2014.892428

[45] Siyuan Chen, Julien Epps, and Fang Chen. 2011. A comparison of four methods for cognitive load measurement. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference (OzCHI'11)*. ACM, Canberra, Australia, 76–79. https://doi.org/10.1145/2071536.2071547

[46] Siyuan Chen, Julien Epps, and Fang Chen. 2013. Automatic and continuous user task analysis via eye activity. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI'13)*. ACM, New York, NY, USA, 57–66. https://doi.org/10.1145/2449396.2449406

[47] Siyuan Chen, Julien Epps, Natalie Ruiz, and Fang Chen. 2011. Eye activity as a measure of human mental effort in HCI. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI'11)*. ACM, New York, NY, USA, 315–318. https://doi.org/10.1145/1943403.1943454

[48] 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

[49] Burcu Cinaz, Bert Arnrich, Roberto La Marca, and Gerhard Tröster. 2013. Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and Ubiquitous Computing* 17, 2 (Feb. 2013), 229–239. https://doi.org/10.1007/s00779-011-0466-1

[50] Johnny Collins, Holger Regenbrecht, Tobias Langlotz, Yekta Said Said Can, Cam Ersoy, and Russel Butson. 2019. Measuring cognitive load and insight: A methodology exemplified in a virtual reality learning context. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 351–362.

[51] Dan Conway, Ian Dick, Zhidong Li, Yang Wang, and Fang Chen. 2013. The effect of stress on cognitive load measurement. In *Human-Computer Interaction – INTERACT 2013*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.), Vol. 8120. Springer Berlin, Berlin, 659–666. https://doi.org/10.1007/978-3-642-40498-6_58

[52] Igor Crk, Timothy Kluthe, and Andreas Stefik. 2016. Understanding programming expertise: An empirical study of phasic brain wave changes. *ACM Trans. Comput.-Hum. Interact.* 23, 1, Article 2 (Dec. 2016). https://doi.org/10.1145/2829945

[53] Tim R. H. Cutmore and Daniel A. James. 1999. Identifying and reducing noise in psychophysiological recordings. *International Journal of Psychophysiology* 32, 2 (1999), 129–150. https://doi.org/10.1016/S0167-8760(99)00014-8

[54] Laura Dabbish and Robert E. Kraut. 2004. Controlling interruptions: Awareness displays and social motivation for coordination. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW'04)*. ACM, New York, NY, USA, 182–191. https://doi.org/10.1145/1031607.1031638

[55] Vagner Figueredo de Santana, Juliana Jansen Ferreira, Rogério Abreu de Paula, and Renato Fontoura de Gusmão Cerqueira. 2018. An eye gaze model for seismic interpretation support. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA'18)*. Association for Computing Machinery, New York, NY, USA, Article 44. https://doi.org/10.1145/3204493.3204554

[56] E. Debie, R. Fernandez Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, and H. A. Abbass. 2021. Multimodal fusion for objective assessment of cognitive workload: A review. *IEEE Transactions on Cybernetics* 51, 3 (2021), 1542–1555. https://doi.org/10.1109/TCYB.2019.2939399

[57] Nicolas Debue and Cécile van de Leemput. 2014. What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology* 5 (2014), 1099. https://doi.org/10.3389/fpsyg.2014.01099

[58] Vera Demberg, Asad Sayeed, Angela Mahr, and Christian Müller. 2013. Measuring linguistically-induced cognitive load during driving using the ConTRe Task. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'13)*. ACM, Eindhoven, Netherlands, 176–183. https://doi.org/10.1145/2516540.2516546

[59] Yi Ding, Brandon Huynh, Aiwen Xu, Tom Bullock, Hubert Cecotti, Matthew Turk, Barry Giesbrecht, and Tobias Höllerer. 2019. Multimodal classification of EEG during physical activity. In *2019 International Conference on Multimodal Interaction (ICMI'19)*. Association for Computing Machinery, New York, NY, USA, 185–194. https://doi.org/10.1145/3340555.3353759

[60] Onur Dönmez, Yavuz Akbulut, Esra Telli, Miray Kaptan, İbrahim H. Özdemir, and Mukaddes Erdem. 2022. In search of a measure to address different sources of cognitive load in computer-based learning environments. *Education and Information Technologies* (2022), 1–22. https://doi.org/10.1007/s10639-022-11035-2

[61] Wendy Doubé and Jeanie Beh. 2012. Typing over autocomplete: Cognitive load in website use by older adults. In *Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI'12)*. ACM, Melbourne, Australia, 97–106. https://doi.org/10.1145/2414536.2414553

[62] Andrew T. Duchowski, Krzysztof Krejtz, Nina A. Gehrer, Tanya Bafna, and Per Bækgaard. 2020. The Low/High index of pupillary activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–12. https://doi.org/10.1145/3313831.3376394

[63] Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, New York, NY, USA, 282:1–282:13. https://doi.org/10.1145/3173574.3173856

[64] Rodrigo Duran, Albina Zavgorodniaia, and Juha Sorva. 2022. Cognitive load theory in computing education research: A review. *ACM Trans. Comput. Educ.* 22, 4, Article 40 (Sep. 2022), 27 pages. https://doi.org/10.1145/3483843

[65] John Fitzgerald Ehrich, Steven J. Howard, Sahar Bokosmaty, and Stuart Woodcock. 2021. An item response modeling approach to cognitive load measurement. *Frontiers in Education* 6 (2021). https://doi.org/10.3389/feduc.2021.648324

[66] Stephen A. Engel, David E. Rumelhart, Brian A. Wandell, Adrian T. Lee, Gary H. Glover, Eduardo-Jose Chichilnisky, and Michael N. Shadlen. 1994. fMRI of human visual cortex. *Nature* (1994).

[67] Andéol Evain, Ferran Argelaguet, Nicolas Roussel, Géry Casiez, and Anatole Lécuyer. 2017. Can I think of something else when using a BCI?: Cognitive demand of an SSVEP-based BCI. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI'17)*. ACM, New York, NY, USA, 5120–5125. https://doi.org/10.1145/3025453.3026037

[68] Dakota C. Evans and Mary Fendley. 2017. A multi-measure approach for connecting cognitive workload and automation. *International Journal of Human-Computer Studies* 97 (2017), 182–189. https://doi.org/10.1016/j.ijhcs.2016.05.008

[69] Marco Ferrari and Valentina Quaresima. 2012. A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage* 63, 2 (2012), 921–935. https://doi.org/10.1016/j.neuroimage.2012.03.049

[70] Jérémy Frey, Maxime Daniel, Julien Castet, Martin Hachet, and Fabien Lotte. 2016. Framework for electroencephalography-based evaluation of user experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, Santa Clara, California, USA, 2283–2294. https://doi.org/10.1145/2858036.2858525

[71] Jérémy Frey, Christian Mühl, Fabien Lotte, and Martin Hachet. 2013. Review of the use of electroencephalography as an evaluation method for human-computer interaction. *arXiv:1311.2222 [cs]* (Nov. 2013). http://arxiv.org/abs/1311.2222. arXiv: 1311.2222.

[72] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T. Freeman. 2018. Cognitive load estimation in the wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, New York, NY, USA, 652:1–652:9. https://doi.org/10.1145/3173574.3174226

[73] Bruce H. Friedman and Julian F. Thayer. 1991. Facial muscle activity and EEG recordings: Redundancy analysis. *Electroencephalography and Clinical Neurophysiology* 79, 5 (1991), 358–360. https://doi.org/10.1016/0013-4694(91)90200-N

[74] Catherine Gabaude, Bruno Baracat, Christophe Jallais, Marion Bonniaud, and Alexandra Fort. 2012. Cognitive load measurement while driving. In *Human Factors: a view from an integrative perspective*. Human Factors and Ergonomics Society, 67–80. https://hal.archives-ouvertes.fr/hal-01027475/file/doc00014683.pdf.

[75] Edith Galy, Magali Cariou, and Claudine Mélan. 2012. What is the relationship between mental workload factors and cognitive load types? *International Journal of Psychophysiology* 83, 3 (2012), 269–275. https://doi.org/10.1016/j.ijpsycho.2011.09.023

[76] Vanessa Georges, Francois Courtemanche, Sylvain Senecal, Thierry Baccino, Marc Fredette, and Pierre-Majorique Leger. 2016. UX heatmaps: Mapping user experience on visual interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, Santa Clara, California, USA, 4850–4860. https://doi.org/10.1145/2858036.2858271

[77] Alan Gevins, Michael E. Smith, Linda McEvoy, and Daphne Yu. 1997. High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex* 7, 4 (1997), 374–385. https://doi.org/10.1093/cercor/7.4.374

[78] Dimitris Giakoumis, Dimitrios Tzovaras, and George Hassapis. 2013. Subject-dependent biosignal features for increased accuracy in psychological stress detection. *International Journal of Human-Computer Studies* 71, 4 (2013), 425–439. https://doi.org/10.1016/j.ijhcs.2012.10.016

[79] Stéphanie Giraud, Pierre Thérouanne, and Dirk D. Steiner. 2018. Web accessibility: Filtering redundant and irrelevant information improves website usability for blind users. *International Journal of Human-Computer Studies* 111 (2018), 23–35. https://doi.org/10.1016/j.ijhcs.2017.10.011

[80] Audrey Girouard, Erin T. Solovey, Leanne M. Hirshfield, Krysta Chauncey, Angelo Sassaroli, Sergio Fantini, and Robert J. K. Jacob. 2009. Distinguishing difficulty levels with non-invasive brain activity measurements. In *Human-Computer Interaction – INTERACT 2009*, Tom Gross, Jan Gulliksen, Paula Kotzé, Lars Oestreicher, Philippe Palanque, Raquel Oliveira Prates, and Marco Winckler (Eds.), Vol. 5726. Springer Berlin, Berlin, 440–452. https://doi.org/10.1007/978-3-642-03655-2_50

[81] Christiane Glatz, Stas S. Krupenia, Heinrich H. Bülthoff, and Lewis L. Chuang. 2018. Use the right sound for the right job: Verbal commands and auditory icons for a task-management system favor different information processes in the brain. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, New York, NY, USA, 472:1–472:13. https://doi.org/10.1145/3173574.3174046

[82] Christopher Griffiths, Judy Bowen, and Annika Hinze. 2017. Investigating wearable technology for fatigue identification in the workplace. In *Human-Computer Interaction - INTERACT 2017*, Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler (Eds.), Vol. 10514. Springer International Publishing, Cham, 370–380. https://doi.org/10.1007/978-3-319-67684-5_22

[83] David Grimes, Desney S. Tan, Scott E. Hudson, Pradeep Shenoy, and Rajesh P. N. Rao. 2008. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*. ACM, Florence, Italy, 835–844. https://doi.org/10.1145/1357054.1357187

[84] Jonathan Grudin. 2017. From tool to partner: The evolution of human-computer interaction. *Synthesis Lectures on Human-Centered Interaction* 10, 1 (2017), i–183. https://doi.org/10.2200/S00745ED1V01Y201612HCI035

[85] Wei Guo and Jingtao Wang. 2018. Towards attentive speed reading on small screen wearable devices. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI'18)*. ACM, New York, NY, USA, 278–287. https://doi.org/10.1145/3242969.3243009 event-place: Boulder, CO, USA.

[86] Surabhi Gupta, Tim Coles, Cedric Dumas, Simon J. McBride, and DanaKai Bradford. 2016. Gamer style: Performance factors in gamified simulation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, Santa Clara, California, USA, 2014–2025. https://doi.org/10.1145/2858036.2858461

[87] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-Physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp'10)*. ACM, Copenhagen, Denmark, 301–310. https://doi.org/10.1145/1864349.1864395

[88] S. G. Hart, M. E. Childress, and J. R. Hauser. 1982. Individual definitions of the term "workload". In *Eighth Symposium on Psychology in the Department of Defense*. 478–485.

[89] Sandra G. Hart. 1986. Theory and measurement of human workload. *Human Productivity Enhancement* 1 (1986), 396–456.

[90] Sandra G. Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (Oct. 2006), 904–908. https://doi.org/10.1177/154193120605000909

[91] 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*. North-Holland, Oxford, England, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[92] Peter A. Heeman, Tomer Meshorer, Andrew L. Kun, Oskar Palinko, and Zeljko Medenica. 2013. Estimating cognitive load using pupil diameter during a spoken dialogue task. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'13)*. ACM, Eindhoven, Netherlands, 242–245. https://doi.org/10.1145/2516540.2516570

[93] Suzana Herculano-Houzel. 2009. The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience* 3 (2009). https://doi.org/10.3389/neuro.09.031.2009

[94] Morten Hertzum and Kristin Due Holmegaard. 2013. Perceived time as a measure of mental workload: Effects of time constraints and task success. *International Journal of Human-Computer Interaction* 29, 1 (Jan. 2013), 26–39. https://doi.org/10.1080/10447318.2012.676538

[95] Heinke Hihn, Sascha Meudt, and Friedhelm Schwenker. 2016. Inferring mental overload based on postural behavior and gestures. In *Proceedings of the 2nd Workshop on Emotion Representations and Modelling for Companion Systems (ERM4CT'16)*. Association for Computing Machinery, New York, NY, USA, Article 3. https://doi.org/10.1145/3009960.3009961

[96] Leanne M. Hirshfield, Rebecca Gulotta, Stuart Hirshfield, Sam Hincks, Matthew Russell, Rachel Ward, Tom Williams, and Robert Jacob. 2011. This is your brain on interfaces: Enhancing usability testing with functional near-infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, Vancouver, BC, Canada, 373–382. https://doi.org/10.1145/1978942.1978996

[97] Leanne M. Hirshfield, Erin T. Solovey, Audrey Girouard, James Kebinger, Robert J. K. Jacob, Angelo Sassaroli, and Sergio Fantini. 2009. Brain measurement for usability testing and adaptive interfaces: An example of uncovering syntactic workload with functional near infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*. ACM, Boston, MA, USA, 2185–2194. https://doi.org/10.1145/1518701.1519035

[98] Nina Hollender, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. 2010. Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior* 26, 6 (Nov. 2010), 1278–1288. https://doi.org/10.1016/j.chb.2010.05.031

[99] M. Sazzad Hussain, Rafael A. Calvo, and Fang Chen. 2013. Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. *Interacting with Computers* 26, 3 (06 2013), 256–268. https://doi.org/10.1093/iwc/iwt032 arXiv:http://oup.prod.sis.lan/iwc/article-pdf/26/3/256/2054117/iwt032.pdf.

[100] Youjin Hwang, Siyoung Lee, Hyeong Seok Jeon, Jung Han Yoon Park, Ki Won Lee, and Joonhwan Lee. 2018. "Eat what you want and be healthy!": Comfort food effects: Human-food interaction in view of celebratory technology. In *Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction (MHFI'18)*. ACM, New York, NY, USA, 4:1–4:8. https://doi.org/10.1145/3279954.3279958 event-place: Boulder, CO, USA.

[101] Curtis S. Ikehara and Martha E. Crosby. 2005. Assessing cognitive load with physiological sensors. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE, 295a.

[102] Jittrapol Intarasirisawat, Chee Siang Ang, Christos Efstratiou, Luke William Feidhlim Dickens, and Rupert Page. 2019. Exploring the touch and motion features in game-based cognitive assessments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 87 (Sept. 2019). https://doi.org/10.1145/3351245

[103] Shamsi T. Iqbal, Piotr D. Adamczyk, Xianjun Sam Zheng, and Brian P. Bailey. 2005. Towards an index of opportunity: Understanding changes in mental workload during task execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'05)*. ACM, Portland, Oregon, USA, 311–320. https://doi.org/10.1145/1054972.1055016

[104] Kurtulus Izzetoglu, Scott Bunce, Banu Onaral, Kambiz Pourrezaei, and Britton Chance. 2004. Functional optical brain imaging using near-infrared during cognitive tasks. *International Journal of Human-Computer Interaction* 17, 2 (June 2004), 211–227. https://doi.org/10.1207/s15327590ijhc1702_6

[105] Julie A. Jacko, Armando Barreto, Ingrid U. Scott, Robert H. Rosa, and Charles J. Pappas. 2000. Using electroencephalogram to investigate stages of visual search in visually impaired computer users: Preattention and focal attention. *International Journal of Human-Computer Interaction* 12, 1 (May 2000), 135–150. https://doi.org/10.1207/S15327590IJHC1201_6

[106] Jan Jarvis, Felix Putze, Dominic Heger, and Tanja Schultz. 2011. Multimodal person independent recognition of workload related biosignal patterns. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI'11)*. ACM, Alicante, Spain, 205–208. https://doi.org/10.1145/2070481.2070516

[107] Xianta Jiang, M. Stella Atkins, Geoffrey Tien, Roman Bednarik, and Bin Zheng. 2014. Pupil responses during discrete goal-directed movements. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI'14)*. ACM, Toronto, Ontario, Canada, 2075–2084. https://doi.org/10.1145/2556288.2557086

[108] Xianta Jiang, Bin Zheng, Roman Bednarik, and M. Stella Atkins. 2015. Pupil responses to continuous aiming movements. *International Journal of Human-Computer Studies* 83 (2015), 1–11. https://doi.org/10.1016/j.ijhcs.2015.05.006

[109] Slava Kalyuga. 2011. Cognitive load theory: How many types of load does it really need? *Educational Psychology Review* 23, 1 (2011), 1–19. https://doi.org/10.1007/s10648-010-9150-7

[110] Jakob Karolus, Felix Bachmann, Thomas Kosch, Albrecht Schmidt, and Paweł W. Woźniak. 2021. Facilitating bodily insights using electromyography-based biofeedback during physical activity. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction* (Toulouse & Virtual, France) *(MobileHCI'21)*. Association for Computing Machinery, New York, NY, USA, Article 14, 15 pages. https://doi.org/10.1145/3447526.3472027

[111] Jakob Karolus, Simon Thanheiser, David Peterson, Nicolas Viot, Thomas Kosch, Albrecht Schmidt, and Paweł W. Wozniak. 2022. Imprecise but fun: Playful interaction using electromyography. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 190 (Sep. 2022), 21 pages. https://doi.org/10.1145/3546725

[112] Ioanna Katidioti. [n.d.]. Interrupted by your pupil: An interruption management system based on pupil dilation. ([n.d.]), 12.

[113] Caitlin Kelleher and Wint Hnin. 2019. Predicting cognitive load in future code puzzles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. Association for Computing Machinery, Glasgow, Scotland UK, 1–12. https://doi.org/10.1145/3290605.3300487

[114] Stephanie Khalfa, Simone Dalla Bella, Mathieu Roy, Isabelle Peretz, and Sonia J. Lupien. 2003. Effects of relaxing music on salivary cortisol level after psychological stress. *Annals of the New York Academy of Sciences* 999, 1 (2003), 374–376.

[115] M. Asif Khawaja, Fang Chen, and Nadine Marcus. 2010. Using language complexity to measure cognitive load for adaptive interaction design. In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI'10)*. ACM, New York, NY, USA, 333–336. https://doi.org/10.1145/1719970.1720024

[116] M. Asif Khawaja, Fang Chen, and Nadine Marcus. 2014. Measuring cognitive load using linguistic features: Implications for usability evaluation and adaptive interaction design. *International Journal of Human–Computer Interaction* 30, 5 (2014), 343–368. https://doi.org/10.1080/10447318.2013.860579 arXiv:https://doi.org/10.1080/10447318.2013.860579

[117] M. Asif Khawaja, Natalie Ruiz, and Fang Chen. 2007. Potential speech features for cognitive load measurement. In *Proceedings of the 19th Australasian Conference on Computer-Human Interaction: Entertaining User Interfaces (OZCHI'07)*. ACM, Adelaide, Australia, 57–60. https://doi.org/10.1145/1324892.1324902

[118] M. Asif Khawaja, Natalie Ruiz, and Fang Chen. 2008. Think before you talk: An empirical study of relationship between speech pauses and cognitive load. In *Proceedings of the 20th Australasian Conference on Computer-Human Interaction: Designing for Habitus and Habitat (OZCHI'08)*. ACM, Cairns, Australia, 335–338. https://doi.org/10.1145/1517744.1517814

[119] Ahmad Khawaji, Fang Chen, Jianlong Zhou, and Nadine Marcus. 2014. Trust and cognitive load in the text-chat environment: The role of mouse movement. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design (OzCHI'14)*. ACM, Sydney, New South Wales, Australia, 324–327. https://doi.org/10.1145/2686612.2686661

[120] Clemens Kirschbaum and Dirk H. Hellhammer. 1989. Salivary cortisol in psychobiological research: An overview. *Neuropsychobiology* 22, 3 (1989), 150–169. https://doi.org/10.1159/000118611

[121] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology* 51, 1 (Jan. 2009), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

[122] Wolfgang Klimesch, Hannes Schimke, and Gert Pfurtscheller. 1993. Alpha frequency, cognitive load and memory performance. *Brain Topography* 5, 3 (1993), 241–251.

[123] Pascal Knierim, Thomas Kosch, Johannes Groschopp, and Albrecht Schmidt. 2020. Opportunities and challenges of text input in portable virtual reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA'20)*. ACM, New York, NY, USA. https://doi.org/10.1145/3334480.3382920

[124] Avi Knoll, Yang Wang, Fang Chen, Jie Xu, Natalie Ruiz, Julien Epps, and Pega Zarjam. 2011. Measuring cognitive workload with low-cost electroencephalograph. In *Human-Computer Interaction – INTERACT 2011*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.), Vol. 6949. Springer Berlin, Berlin, 568–571. https://doi.org/10.1007/978-3-642-23768-3_84

[125] Gennady G. Knyazev. 2012. EEG delta oscillations as a correlate of basic homeostatic and motivational processes. *Neuroscience & Biobehavioral Reviews* 36, 1 (2012), 677–695. https://doi.org/10.1016/j.neubiorev.2011.10.002

[126] Andreas Korbach, Roland Brünken, and Babette Park. 2018. Differentiating different types of cognitive load: A comparison of different measures. *Educational Psychology Review* 30, 2 (2018), 503–529. https://doi.org/10.1007/s10648-017-9404-8

[127] A. V. Korshakov, A. A. Frolov, and P. D. Bobrov. 2010. On-line automatic suppression of artifacts in multi-dimensional signals using ICA. In *Abstr. 10th Europ. Conf. on Non-Destructive Testing*. 370.

[128] Thomas Kosch. 2020. Workload-Aware Systems and Interfaces for Cognitive Augmentation. arXiv: 2010.07703 [cs.HC]

[129] Thomas Kosch and Lewis L. Chuang. 2019. Investigating the influence of RSVP display parameters on working memory load using electroencephalography. In *2nd International Conference on Neuroadaptive Technology*.

[130] Thomas Kosch, Markus Funk, Albrecht Schmidt, and Lewis L. Chuang. 2018. Identifying cognitive assistance with mobile electroencephalography: A case study with in-situ projections for manual assembly. *Proceedings of the ACM on Human-Computer Interaction* 2, EICS (June 2018), 1–20. https://doi.org/10.1145/3229093

[131] Thomas Kosch, Mariam Hassib, Daniel Buschek, and Albrecht Schmidt. 2018. Look into my eyes: Using pupil dilation to estimate mental workload for task complexity adaptation. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA'18)*. ACM, New York, NY, USA, LBW617:1–LBW617:6. https://doi.org/10.1145/3170427.3188643

[132] Thomas Kosch, Mariam Hassib, and Albrecht Schmidt. 2016. The brain matters: A 3D real-time visualization to examine brain source activation leveraging neurofeedback. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Santa Clara, California, USA) *(CHI EA'16)*. ACM, New York, NY, USA, 1570–1576. https://doi.org/10.1145/2851581.2892484

[133] Thomas Kosch, Mariam Hassib, Paweł W. Woźniak, Daniel Buschek, and Florian Alt. 2018. Your eyes tell: Leveraging smooth pursuit for assessing cognitive workload. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, New York, NY, USA, 436:1–436:13. https://doi.org/10.1145/3173574.3174010

[134] Thomas Kosch, Jakob Karolus, Havy Ha, and Albrecht Schmidt. 2019. Your skin resists: Exploring electrodermal activity as workload indicator during manual assembly. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS'19)*. Association for Computing Machinery, New York, NY, USA, Article 8. https://doi.org/10.1145/3319499.3328230

[135] Thomas Kosch, Albrecht Schmidt, Simon Thanheiser, and Lewis L. Chuang. 2020. One does not simply RSVP: Mental workload to select speed reading parameters using electroencephalography. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376766

[136] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2022. The placebo effect of artificial intelligence in human-computer interaction. *ACM Trans. Comput.-Hum. Interact.* (June 2022). https://doi.org/10.1145/3529225 Just Accepted.

[137] Thomas Kosch, Kevin Wennrich, Daniel Topp, Marcel Muntzinger, and Albrecht Schmidt. 2019. The digital cooking coach: Using visual and auditory in-situ instructions to assist cognitively impaired during cooking. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Rhodes, Greece) *(PETRA'19)*. ACM, New York, NY, USA, 156–163. https://doi.org/10.1145/3316782.3321524

[138] Makrina Viola Kosti, Kostas Georgiadis, Dimitrios A. Adamos, Nikos Laskaris, Diomidis Spinellis, and Lefteris Angelis. 2018. Towards an affordable brain computer interface for the assessment of programmers' mental workload. *International Journal of Human-Computer Studies* 115 (2018), 52–66. https://doi.org/10.1016/j.ijhcs.2018.03.002

[139] John W. Krakauer and Reza Shadmehr. 2006. Consolidation of motor memory. *Trends in Neurosciences* 29, 1 (Jan. 2006), 58–64. https://doi.org/10.1016/j.tins.2005.10.003

[140] Tuomo Kujala. 2009. Efficiency of visual time-sharing behavior: The effects of menu structure on POI search tasks while driving. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'09)*. ACM, Essen, Germany, 63–70. https://doi.org/10.1145/1620509.1620522

[141] Naveen Kumar and Jyoti Kumar. 2016. Measurement of cognitive load in HCI systems using EEG power spectrum: An experimental study. *Procedia Computer Science* 84 (2016), 70–78. https://doi.org/10.1016/j.procs.2016.04.068

[142] Andrew L. Kun, Oskar Palinko, and Ivan Razumenić. 2012. Exploring the effects of size and luminance of visual targets on the pupillary light reflex. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'12)*. ACM, Portsmouth, New Hampshire, USA, 183–186. https://doi.org/10.1145/2390256.2390287

[143] Andrew L. Kun, Alexander Shyrokov, and Peter A. Heeman. 2010. Spoken tasks for human-human experiments: Towards in-car speech user interfaces for multi-threaded dialogue. In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Pittsburgh, Pennsylvania) *(AutomotiveUI'10)*. Association for Computing Machinery, New York, NY, USA, 57–63. https://doi.org/10.1145/1969773.1969784

[144] T. J. La Vaque. 1999. The history of EEG Hans Berger: Psychophysiologist. A historical vignette. *Journal of Neurotherapy* 3, 2 (1999), 1–9.

[145] J. N Langley. 1921. *The Autonomic Nervous System. Part 1.* W. Heffer, Cambridge, England.

[146] David R. Large, Gary Burnett, Ben Anyasodo, and Lee Skrypchuk. 2016. Assessing cognitive demand during natural language interactions with a digital driving assistant. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (Automotive'UI 16)*. ACM, Ann Arbor, MI, USA, 67–74. https://doi.org/10.1145/3003715.3005408

[147] Byung Lee, Kwanghun Chung, and Sung-Hee Kim. 2018. Interruption cost evaluation by cognitive workload and task performance in interruption coordination modes for human–computer interaction tasks. *Applied Sciences* 8, 10 (Sept. 2018), 1780. https://doi.org/10.3390/app8101780

[148] Johnny Chung Lee and Desney S. Tan. 2006. Using a low-cost electroencephalograph for task classification in HCI research. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology - UIST'06*. ACM Press, Montreux, Switzerland, 81. https://doi.org/10.1145/1166253.1166268

[149] Moira LeMay and Eric Hird. 1986. Operator work load: When is enough enough? *Commun. ACM* 29, 7 (July 1986), 638–642. https://doi.org/10.1145/6138.6147

[150] Grace Leslie and Tim Mullen. 2011. MoodMixer: EEG-based collaborative sonification. In *Proceedings of the International Conference on New Interfaces for Musical Expression.* Oslo, Norway, 296–299.

[151] Calvin Liang, Jakob Karolus, Thomas Kosch, and Albrecht Schmidt. 2018. On the suitability of real-time assessment of programming proficiency using gaze properties. In *Proceedings of the 7th ACM International Symposium on Pervasive Displays* (Munich, Germany) *(PerDis'18).* Association for Computing Machinery, New York, NY, USA, Article 31, 2 pages. https://doi.org/10.1145/3205873.3210702

[152] Tao Lin, Atsumi Imamiya, and Xiaoyang Mao. 2008. Using multiple data sources to get closer insights into user cost and task performance. *Interacting with Computers* 20, 3 (02 2008), 364–374. https://doi.org/10.1016/j.intcom.2007.12.002 arXiv:http://oup.prod.sis.lan/iwc/article-pdf/20/3/364/2007251/iwc20-0364.pdf.

[153] Y. Lin, W. J. Zhang, and L. G. Watson. 2003. Using eye movement parameters for evaluating human–machine interface frameworks under normal control operation and fault detection situations. *International Journal of Human-Computer Studies* 59, 6 (2003), 837–873. https://doi.org/10.1016/S1071-5819(03)00122-8

[154] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST'19).* Association for Computing Machinery, New York, NY, USA, 147–160. https://doi.org/10.1145/3332165.3347945

[155] Fabien Lotte, Laurent Bougrain, Anatole Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. 2018. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. 28 pages. https://doi.org/10.1088/1741-2552/aab2f2

[156] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. 2007. A review of classification algorithms for EEG-based brain–computer interfaces. 13 pages. https://doi.org/10.1088/1741-2560/4/2/r01

[157] Shihan Lu, Meng Yuan Zhang, Tulga Ersal, and X. Jessie Yang. 2019. Workload management in teleoperation of unmanned ground vehicles: Effects of a delay compensation aid on human operators' workload and teleoperation performance. *International Journal of Human–Computer Interaction* 35, 19 (2019), 1820–1830. https://doi.org/10.1080/10447318.2019.1574059 arXiv:https://doi.org/10.1080/10447318.2019.1574059

[158] Steven J. Luck. 2014. *An Introduction to the Event-related Potential Technique.* MIT Press.

[159] Kristiyan Lukanov, Horia A. Maior, and Max L. Wilson. 2016. Using fNIRS in Usability Testing: Understanding the Effect of Web Form Layout on Mental Workload. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16).* ACM, Santa Clara, California, USA, 4011–4016. https://doi.org/10.1145/2858036.2858236

[160] Yongqiang Lyu, Xiaomin Luo, Jun Zhou, Chun Yu, Congcong Miao, Tong Wang, Yuanchun Shi, and Ken-ichi Kameyama. 2015. Measuring Photoplethysmogram-Based Stress-Induced Vascular Response Index to Assess Cognitive Load and Stress. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15).* ACM, Seoul, Republic of Korea, 857–866. https://doi.org/10.1145/2702123.2702399

[161] Horia A. Maior, Max L. Wilson, and Sarah Sharples. 2018. Workload alerts—using physiological measures of mental workload to provide feedback during tasks. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 9 (April 2018), 30 pages. https://doi.org/10.1145/3173380

[162] M. Malik and A. J. Camm. 1990. Heart rate variability. *Clinical Cardiology* 13, 8 (1990), 570–576. https://doi.org/10.1002/clc.4960130811 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/clc.4960130811.

[163] Plonsey Malmivuo, Jaakko Malmivuo, and Robert Plonsey. 1995. *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields.* Oxford University Press. Google-Books-ID: H9CFM0TqWwsC.

[164] Sandra P. Marshall. 2002. The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants.* IEEE, 7–7.

[165] Sebastian Marwecki, Andrew D. Wilson, Eyal Ofek, Mar Gonzalez Franco, and Christian Holz. 2019. Mise-Unseen: Using eye tracking to hide virtual reality scene changes in plain sight. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST'19).* Association for Computing Machinery, New York, NY, USA, 777–789. https://doi.org/10.1145/3332165.3347919

[166] S. G. Mason and G. E. Birch. 2003. A general framework for brain-computer interface design. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 11, 1 (March 2003), 70–85. https://doi.org/10.1109/TNSRE.2003.810426

[167] G. Matthews, L. Joyner, K. Gilliland, S. Campbell, S. Falconer, J. Huggings, and I. Mervielde. 1999. Validation of a comprehensive stress state questionnaire: Towards a state Big Three?. *8th European Conference on Personality.* 7 (1999), 335–350.

[168] Gerald Matthews, James Szalma, April Rose Panganiban, Catherine Neubauer, and Joel S. Warm. 2013. Profiling task stress with the Dundee Stress State Questionnaire. *Psychology of Stress: New Research* 1 (2013), 49–90.

[169] Daniel J. McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W. Picard. 2016. COGCAM: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16).* ACM, Santa Clara, California, USA, 4000–4004. https://doi.org/10.1145/2858036.2858247

[170] Samuel Melamed, Arie Shirom, Sharon Toker, Shlomo Berliner, and Itzhak Shapira. 2006. Burnout and risk of cardiovascular disease: Evidence, possible causal paths, and promising research directions. *Psychological Bulletin* 132, 3 (2006), 327. https://doi.org/doi/10.1037/0033-2909.132.3.327

[171] Roberto Merletti and Dario Farina. 2016. *Surface Electromyography: Physiology, Engineering and Applications*. John Wiley & Sons.

[172] Shinji Miyake. 2001. Multivariate workload evaluation combining physiological and subjective measures. *International Journal of Psychophysiology* 40, 3 (2001), 233–238. https://doi.org/10.1016/S0167-8760(00)00191-4 Psychophysiology in.

[173] Philipp Mock, Peter Gerjets, Maike Tibus, Ulrich Trautwein, Korbinian Möller, and Wolfgang Rosenstiel. 2016. Using touchscreen interaction data to predict cognitive workload. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*. ACM, Tokyo, Japan, 349–356. https://doi.org/10.1145/2993148.2993202

[174] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine* 6, 7 (July 2009), e1000097. https://doi.org/10.1371/journal.pmed.1000097

[175] Prithima R. Mosaly, Hua Guo, and Lukasz Mazur. 2019. Toward better understanding of task difficulty during physicians' interaction with electronic health record system (EHRs). *International Journal of Human–Computer Interaction* 35, 20 (2019), 1883–1891. https://doi.org/10.1080/10447318.2019.1575081 arXiv:https://doi.org/10.1080/10447318.2019.1575081

[176] T. R. Mullen, C. A. E. Kothe, Y. M. Chi, A. Ojeda, T. Kerth, S. Makeig, T. Jung, and G. Cauwenberghs. 2015. Real-time neuroimaging and cognitive monitoring using wearable dry EEG. 2553–2567. https://doi.org/10.1109/TBME.2015.2481482

[177] Noman Naseer and Keum-Shik Hong. 2015. fNIRS-based brain-computer interfaces: A review. *Frontiers in Human Neuroscience* 9 (2015), 3. https://doi.org/10.3389/fnhum.2015.00003

[178] Urs M. Nater, Nadine Skoluda, and Jana Strahler. 2013. Biomarkers of stress in behavioural medicine. *Current Opinion in Psychiatry* 26, 5 (Sept. 2013), 440–445. https://doi.org/10.1097/YCO.0b013e328363b4ed

[179] Mark Neerincx. 2003. Cognitive task load analysis: Allocating tasks and designing support. *Handbook of Cognitive Task Design. Chapter 13. Mahwah, NJ: Lawrence Erlbaum Associates* (01 2003), 283–305.

[180] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. 2012. Brain computer interfaces, a review. *Sensors* 12, 2 (Feb. 2012), 1211–1279. https://doi.org/10.3390/s120201211

[181] Vadim V. Nikulin, Guido Nolte, and Gabriel Curio. 2011. A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition. *NeuroImage* 55, 4 (April 2011), 1528–1535. https://doi.org/10.1016/j.neuroimage.2011.01.057

[182] Nargess Nourbakhsh, Yang Wang, and Fang Chen. 2013. GSR and blink features for cognitive load classification. In *Human-Computer Interaction – INTERACT 2013*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.), Vol. 8117. Springer Berlin, Berlin, 159–166. https://doi.org/10.1007/978-3-642-40483-2_11

[183] Nargess Nourbakhsh, Yang Wang, Fang Chen, and Rafael A. Calvo. 2012. Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI'12)*. ACM, Melbourne, Australia, 420–423. https://doi.org/10.1145/2414536.2414602

[184] Domen Novak, Benjamin Beyeler, Ximena Omlin, and Robert Riener. 2014. Workload estimation in physical human–robot interaction using physiological measurements. *Interacting with Computers* 27, 6 (05 2014), 616–629. https://doi.org/10.1093/iwc/iwu021 arXiv:http://oup.prod.sis.lan/iwc/article-pdf/27/6/616/2620736/iwu021.pdf.

[185] Domen Novak, Matjaž Mihelj, and Marko Munih. 2009. Using psychophysiological measurements in physically demanding virtual environments. In *Human-Computer Interaction – INTERACT 2009*, Tom Gross, Jan Gulliksen, Paula Kotzé, Lars Oestreicher, Philippe Palanque, Raquel Oliveira Prates, and Marco Winckler (Eds.), Vol. 5726. Springer Berlin, Berlin, 490–493. https://doi.org/10.1007/978-3-642-03655-2_55

[186] Kim Ouwehand, Avalon van der Kroef, Jacqueline Wong, and Fred Paas. 2021. Measuring cognitive load: Are there more valid alternatives to Likert rating scales? *Frontiers in Education* 6 (2021). https://doi.org/10.3389/feduc.2021.702616

[187] Oskar Palinko and Andrew L. Kun. 2012. Exploring the effects of visual cognitive load and illumination on pupil diameter in driving simulators. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA'12)*. ACM, New York, NY, USA, 413–416. https://doi.org/10.1145/2168556.2168650

[188] Oskar Palinko, Andrew L. Kun, Alexander Shyrokov, and Peter Heeman. 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications (ETRA'10)*. ACM, New York, NY, USA, 141–144. https://doi.org/10.1145/1743666.1743701

[189] Timo Partala and Veikko Surakka. 2003. Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies* 59, 1–2 (2003), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X Applications of Affective Computing in Human-Computer Interaction.

[190] Annie Pauzié. 2008. A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems* 2, 4 (2008), 315–322.

[191] E. M. Peck, E. Carlin, and R. Jacob. 2015. Designing brain-computer interfaces for attention-aware systems. *Computer* 48, 10 (Oct. 2015), 34–42. https://doi.org/10.1109/MC.2015.315

[192] Evan M. Peck, Beste F. Yuksel, Alvitta Ottley, Robert J. K. Jacob, and Remco Chang. 2013. Using fNIRS brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. ACM, New York, NY, USA, 473–482. https://doi.org/10.1145/2470654.2470723

[193] Tabitha C. Peck, Jessica J. Good, and Kimberly A. Bourne. 2020. Inducing and mitigating stereotype threat through gendered virtual body-swap illusions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376419

[194] V. Pejović, M. Gjoreski, C. Anderson, K. David, and M. Luštrek. 2020. Toward cognitive load inference for attention management in ubiquitous systems. *IEEE Pervasive Computing* 19, 2 (2020), 35–45.

[195] Bastian Pfleging, Drea K. Fekety, Albrecht Schmidt, and Andrew L. Kun. 2016. A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, Santa Clara, California, USA, 5776–5788. https://doi.org/10.1145/2858036.2858117

[196] Matthew F. Pike, Horia A. Maior, Martin Porcheron, Sarah C. Sharples, and Max L. Wilson. 2014. Measuring the effect of think aloud protocols on workload using fNIRS. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI'14)*. ACM, Toronto, Ontario, Canada, 3807–3816. https://doi.org/10.1145/2556288.2556974

[197] Kathrin Pollmann, Oliver Stefani, Amelie Bengsch, Matthias Peissner, and Mathias Vukelić. 2019. How to work in the car of the future? A neuroergonomical study assessing concentration, performance and workload based on subjective, behavioral and neurophysiological insights. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. Association for Computing Machinery, Glasgow, Scotland, UK, 1–14. https://doi.org/10.1145/3290605.3300284

[198] Felix Putze and Tanja Schultz. 2014. Investigating intrusiveness of workload adaptation. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI'14)*. ACM, Istanbul, Turkey, 275–281. https://doi.org/10.1145/2663204.2663279

[199] Sohail Rafiqi, Suku Nair, and Ephrem Fernandez. 2014. Cognitive and context-aware applications. In *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA'14)*. Association for Computing Machinery, New York, NY, USA, Article 23. https://doi.org/10.1145/2674396.2674445

[200] Sohail Rafiqi, Chatchai Wangwiwattana, Jasmine Kim, Ephrem Fernandez, Suku Nair, and Eric C. Larson. 2015. PupilWare: Towards pervasive cognitive load measurement using commodity devices. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA'15)*. Association for Computing Machinery, New York, NY, USA, Article 42. https://doi.org/10.1145/2769493.2769506

[201] Rahul Rajan, Ted Selker, and Ian Lane. 2016. Task load estimation and mediation using psycho-physiological measures. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI'16)*. ACM, New York, NY, USA, 48–59. https://doi.org/10.1145/2856767.2856769

[202] Bryan Reimer, Bruce Mehler, Joseph F. Coughlin, Kathryn M. Godfrey, and Chuanzhong Tan. 2009. An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'09)*. ACM, Essen, Germany, 115–118. https://doi.org/10.1145/1620509.1620531

[203] G. Robert J. Hockey. 1997. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology* 45, 1 (March 1997), 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

[204] Romisa Rohani Ghahari, Jennifer George-Palilonis, Hossain Gahangir, Lindsay N. Kaser, and Davide Bolchini. 2016. Semi-aural interfaces: Investigating voice-controlled aural flows. *Interacting with Computers* 28, 6 (10 2016), 826–842. https://doi.org/10.1093/iwc/iww004 arXiv:http://oup.prod.sis.lan/iwc/article-pdf/28/6/826/7920074/iww004.pdf.

[205] A. H. Roscoe and G. A. Ellis. 1990. A subjective rating scale for assessing pilot workload in flight: A decade of practical use. (1990), 18.

[206] R. N. Roy, S. Charbonnier, A. Campagne, and S. Bonnet. 2016. Efficient mental workload estimation using task-independent EEG features. 10 pages. https://doi.org/10.1088/1741-2560/13/2/026019

[207] D. Rozado and A. Dünser. 2015. Combining EEG with pupillometry to improve cognitive workload detection. *Computer* 48, 10 (Oct. 2015), 18–25. https://doi.org/10.1109/MC.2015.314

[208] Darrell S. Rudmann, George W. McConkie, and Xianjun Sam Zheng. 2003. Eyetracking in cognitive state detection for HCI. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (Vancouver, British Columbia, Canada) *(ICMI'03)*. ACM, New York, NY, USA, 159–163.

[209] Natalie Ruiz, Qian Qian Feng, Ronnie Taib, Tara Handke, and Fang Chen. 2010. Cognitive skills learning: Pen input patterns in computer-based athlete training. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI'10)*. ACM, Beijing, China, Article 41, 4 pages. https://doi.org/10.1145/1891903.1891955

[210] Natalie Ruiz, Ronnie Taib, and Fang Chen. 2011. Freeform pen-input as evidence of cognitive load and expertise. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI'11)*. ACM, Alicante, Spain, 185–188. https://doi.org/10.1145/2070481.2070511

[211] Natalie Ruiz, Ronnie Taib, Yu (David) Shi, Eric Choi, and Fang Chen. 2007. Using pen input features as indices of cognitive load. In *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI'07)*. ACM, Nagoya, Aichi, Japan, 315–318. https://doi.org/10.1145/1322192.1322246

[212] B. Sakyrs. 1973. Analysis of heart rate variability. *Ergonomics* 16, 1 (1973), 17–32. https://doi.org/10.1080/00140137308924479

[213] Paul Sauseng, Wolfgang Klimesch, Manuel Schabus, and Michael Doppelmayr. 2005. Fronto-parietal EEG coherence in theta and upper alpha reflect central executive functions of working memory. *International Journal of Psychophysiology* 57, 2 (2005), 97–103. https://doi.org/10.1016/j.ijpsycho.2005.03.018 EEG Coherence.

[214] Sergei A. Schapkin and Xenija Weißbecker-Klaus. 2015. Test battery for assessment of cognitive function in older employees: Performance, brain processes, and cardiovascular "Costs". In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA'15)*. Association for Computing Machinery, New York, NY, USA, Article 7. https://doi.org/10.1145/2769493.2769553

[215] Clemens Schartmüller, Klemens Weigl, Philipp Wintersberger, Andreas Riener, and Marco Steinhauser. 2019. Text comprehension: Heads-up vs. auditory displays: Implications for a productive work environment in SAE Level 3 automated vehicles. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Utrecht, Netherlands) *(AutomotiveUI'19)*. Association for Computing Machinery, New York, NY, USA, 342–354. https://doi.org/10.1145/3342197.3344547

[216] Florian Schaule, Jan Ole Johanssen, Bernd Bruegge, and Vivian Loftness. 2018. Employing consumer wearables to detect office workers' cognitive load for interruption management. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1 (March 2018). https://doi.org/10.1145/3191764

[217] Stefan Schneegass, Bastian Pfleging, Nora Broy, Frederik Heinrich, and Albrecht Schmidt. 2013. A data set of real world driving to assess driver workload. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'13)*. ACM, Eindhoven, Netherlands, 150–157. https://doi.org/10.1145/2516540.2516561

[218] Wolfgang Schnotz and Christian Kürschner. 2007. A reconsideration of cognitive load theory. *Educational Psychology Review* 19, 4 (2007), 469–508. https://doi.org/10.1007/s10648-007-9053-4

[219] Bobbie Seppelt, Sean Seaman, Linda Angell, Bruce Mehler, and Bryan Reimer. 2017. Differentiating cognitive load using a modified version of AttenD. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'17)*. ACM, Oldenburg, Germany, 114–122. https://doi.org/10.1145/3122986.3123019

[220] Daniel Sjölie, Kenneth Bodin, Eva Elgh, Johan Eriksson, Lars-Erik Janlert, and Lars Nyberg. 2010. Effects of interactivity and 3D-motion on mental rotation brain activity in an immersive virtual environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, Atlanta, Georgia, USA, 869–878. https://doi.org/10.1145/1753326.1753454

[221] Erin T. Solovey, Daniel Afergan, Evan M. Peck, Samuel W. Hincks, and Robert J. K. Jacob. 2015. Designing implicit interfaces for physiological computing: Guidelines and lessons learned using fNIRS. *ACM Trans. Comput.-Hum. Interact.* 21, 6, Article 35 (Jan. 2015), 27 pages. https://doi.org/10.1145/2687926

[222] Erin T. Solovey, Francine Lalooses, Krysta Chauncey, Douglas Weaver, Margarita Parasi, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, Paul Schermerhorn, Audrey Girouard, and Robert J. K. Jacob. 2011. Sensing cognitive multitasking for a brain-based adaptive user interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. Association for Computing Machinery, Vancouver, BC, Canada, 383–392. https://doi.org/10.1145/1978942.1978997

[223] Erin T. Solovey, Paul Schermerhorn, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, and Robert Jacob. 2012. Brainput: Enhancing interactive systems with streaming fNIRS brain input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. Association for Computing Machinery, Austin, Texas, USA, 2193–2202. https://doi.org/10.1145/2207676.2208372

[224] Erin T. Solovey, Marin Zec, Enrique Abdon Garcia Perez, Bryan Reimer, and Bruce Mehler. 2014. Classifying driver workload using physiological and driving performance data: Two field studies. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI'14)*. ACM, Toronto, Ontario, Canada, 4057–4066. https://doi.org/10.1145/2556288.2557068

[225] Priyashri K. Sridhar, Samantha W. T. Chan, and Suranga Nanayakkara. 2018. Going beyond performance scores: Understanding cognitive-affective states in kindergarteners. In *Proceedings of the 17th ACM Conference on Interaction Design and Children (IDC'18)*. ACM, New York, NY, USA, 253–265. https://doi.org/10.1145/3202185.3202739 event-place: Trondheim, Norway.

[226] Namrata Srivastava, Eduardo Velloso, Jason M. Lodge, Sarah Erfani, and James Bailey. 2019. Continuous evaluation of video lectures from real-time difficulty self-report. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. Association for Computing Machinery, Glasgow, Scotland, UK, 1–12. https://doi.org/10.1145/3290605.3300816

[227] Mark St. John, David A. Kobus, Jeffrey G. Morrison, and Dylan Schmorrow. 2004. Overview of the DARPA augmented cognition technical integration experiment. *International Journal of Human-Computer Interaction* 17, 2 (June 2004), 131–149. https://doi.org/10.1207/s15327590ijhc1702_2

[228] Ben Steichen, Giuseppe Carenini, and Cristina Conati. 2013. User-adaptive information visualization: Using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI'13)*. ACM, New York, NY, USA, 317–328. https://doi.org/10.1145/2449396.2449439

[229] Klaas Enno Stephan and Alard Roebroeck. 2012. A short history of causal modeling of fMRI data. *NeuroImage* 62, 2 (2012), 856–863. https://doi.org/10.1016/j.neuroimage.2012.01.034 20 YEARS OF fMRI.

[230] Kyle Strimbu and Jorge A. Tavel. 2010. What are biomarkers? *Current Opinion in HIV and AIDS* 5, 6 (Nov. 2010), 463–466. https://doi.org/10.1097/COH.0b013e32833ed177

[231] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12, 2 (April 1988), 257–285. https://doi.org/10.1207/s15516709cog1202_4

[232] John Sweller. 2008. Evolutionary bases of human cognitive architecture: Implications for computing education. In *Proceedings of the Fourth International Workshop on Computing Education Research (ICER'08)*. ACM, Sydney, Australia, 1–2. https://doi.org/10.1145/1404520.1404521

[233] John Sweller. 2016. Cognitive load theory and computer science education. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education (SIGCSE'16)*. ACM, Memphis, Tennessee, USA, 1–1. https://doi.org/10.1145/2839509.2844549

[234] John Sweller, Jeroen J. G. Van Merrienboer, and Fred G. W. C. Paas. 1998. Cognitive architecture and instructional design. *Educational Psychology Review* 10, 3 (1998), 251–296. https://doi.org/10.1023/A:1022193728205

[235] Daniel Szafir and Bilge Mutlu. 2012. Pay attention! Designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. Association for Computing Machinery, Austin, Texas, USA, 11–20. https://doi.org/10.1145/2207676.2207679

[236] Koji Takahashi, Minoru Nakayama, and Yasutaka Shimizu. 2000. The response of eye-movement and pupil size to audio instruction while viewing a moving target. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA'00)*. ACM, New York, NY, USA, 131–138. https://doi.org/10.1145/355017.355043

[237] Andrew J. Tattersall and Penelope S. Foord. 1996. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics* 39, 5 (May 1996), 740–748. https://doi.org/10.1080/00140139608964495

[238] M. Thulasidas, Cuntai Guan, and Jiankang Wu. 2006. Robust classification of EEG signal for brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14, 1 (2006), 24–29. https://doi.org/10.1109/TNSRE.2005.862695

[239] Dereck Toker, Sébastien Lallé, and Cristina Conati. 2017. Pupillometry and head distance to the screen to predict skill acquisition during information visualization tasks. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI'17)*. Association for Computing Machinery, New York, NY, USA, 221–231. https://doi.org/10.1145/3025171.3025187

[240] Udo Trutschel, Christian Heinze, Bill Sirois, Martin Golz, David Sommer, and David Edwards. 2012. Heart rate measures reflect the interaction of low mental workload and fatigue during driving simulation. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'12)*. ACM, Portsmouth, New Hampshire, USA, 261–264. https://doi.org/10.1145/2390256.2390299

[241] Pamela S. Tsang and Velma L. Velazquez. 1996. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics* 39, 3 (1996), 358–381. https://doi.org/10.1080/00140139608964470

[242] Rajan Vaish, Keith Wyngarden, Jingshu Chen, Brandon Cheung, and Michael S. Bernstein. 2014. Twitch crowdsourcing: Crowd contributions in short bursts of time. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI'14)*. ACM, Toronto, Ontario, Canada, 3645–3654. https://doi.org/10.1145/2556288.2556996

[243] K. D. Van Benthem, S. Cebulski, C. M. Herdman, and J. Keillor. 2018. An EEG brain–computer interface approach for classifying vigilance states in humans: A gamma band focus supports low misclassification rates. *International Journal of Human–Computer Interaction* 34, 3 (March 2018), 226–237. https://doi.org/10.1080/10447318.2017.1342942

[244] Akos Vetek and Saija Lemmelä. 2011. Could a dialog save your life?: Analyzing the effects of speech interaction strategies while driving. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI'11)*. ACM, Alicante, Spain, 145–152. https://doi.org/10.1145/2070481.2070506

[245] Lisa M. Vizer and Andrew Sears. 2017. Efficacy of personalized models in discriminating high cognitive demand conditions using text-based interactions. *International Journal of Human-Computer Studies* 104 (2017), 80–96. https://doi.org/10.1016/j.ijhcs.2017.03.001

[246] Maria Vukovic, Vidhyasaharan Sethu, Jessica Parker, Lawrence Cavedon, Margaret Lech, and John Thangarajah. 2019. Estimating cognitive load from speech gathered in a complex real-life training exercise. *International Journal of Human-Computer Studies* 124 (2019), 116–133. https://doi.org/10.1016/j.ijhcs.2018.12.003

[247] L. Wang, T. Gu, A. X. Liu, H. Yao, X. Tao, and J. Lu. 2019. Assessing user mental workload for smartphone applications with built-in sensors. *IEEE Pervasive Computing* 18, 1 (2019), 59–70.

[248] Weihong Wang, Zhidong Li, Yang Wang, and Fang Chen. 2013. Indexing cognitive workload based on pupillary response under luminance and emotional changes. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI'13)*. ACM, New York, NY, USA, 247–256. https://doi.org/10.1145/2449396.2449428

[249] Chatchai Wangwiwattana, Xinyi Ding, and Eric C. Larson. 2018. PupilNet, measuring task evoked pupillary response using commodity RGB tablet cameras: Comparison to mobile, infrared gaze trackers for inferring cognitive load. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4 (Jan. 2018). https://doi.org/10.1145/3161164

[250] Christopher D. Wickens. 1981. *Processing Resources in Attention, Dual Task Performance, and Workload Assessment.* Technical Report. Illinois University at Urbana Engineering-Psychology Research Lab.

[251] Christopher D. Wickens. 1984. Processing resources in attention. *Varieties of Attention* (1984).

[252] Christopher D. Wickens. 2008. Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 3 (June 2008), 449–455. https://doi.org/10.1518/001872008X288394

[253] John R. Wilson and Sarah Sharples. 2015. *Evaluation of Human Work.* CRC Press.

[254] Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. 2002. Brain–computer interfaces for communication and control. *Clinical Neurophysiology* 113, 6 (June 2002), 767–791. https://doi.org/10.1016/S1388-2457(02)00057-3

[255] Jie Xu, Yang Wang, Fang Chen, and Eric Choi. 2011. Pupillary response based cognitive workload measurement under luminance changes. In *Human-Computer Interaction – INTERACT 2011*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.), Vol. 6947. Springer Berlin, Berlin, 178–185. https://doi.org/10.1007/978-3-642-23771-3_14

[256] Takehiro Yamakoshi, Ken-ichi Yamakoshi, Shinobu Tanaka, Masamichi Nogawa, Mariko Shibata, Y. Sawada, P. Rolfe, and Yukio Hirose. 2007. A preliminary study on driver's stress index using a new method based on differential skin temperature measurement. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 722–725.

[257] Xiaonan Yang and Jung Hyup Kim. 2018. Measuring workload in a multitasking environment using fractal dimension of pupil dilation. *International Journal of Human–Computer Interaction* (Oct. 2018). https://www.tandfonline.com/doi/pdf/10.1080/10447318.2018.1525022?needAccess=true.

[258] Yan Yang, Bryan Reimer, Bruce Mehler, Alan Wong, and Mike McDonald. 2012. Exploring differences in the impact of auditory and visual demands on driver behavior. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'12)*. ACM, Portsmouth, New Hampshire, 173–177. https://doi.org/10.1145/2390256.2390285

[259] Bo Yin, Natalie Ruiz, Fang Chen, and M. Asif Khawaja. 2007. Automatic cognitive load detection from speech features. In *Proceedings of the 19th Australasian Conference on Computer-Human Interaction: Entertaining User Interfaces (OZCHI'07)*. ACM, Adelaide, Australia, 249–255. https://doi.org/10.1145/1324892.1324946

[260] Kun Yu, Julien Epps, and Fang Chen. 2011. Cognitive load evaluation of handwriting using stroke-level features. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI'11)*. ACM, New York, NY, USA, 423–426. https://doi.org/10.1145/1943403.1943481

[261] Beste Filiz Yuksel, Daniel Afergan, Evan Peck, Garth Griffin, Lane Harrison, Nick Chen, Remco Chang, and Robert Jacob. 2015. BRAAHMS: A novel adaptive musical interface based on users' cognitive state. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Edgar Berdahl and Jesse Allison (Eds.). Louisiana State University, Baton Rouge, Louisiana, USA, 136–139.

[262] Beste F. Yuksel, Kurt B. Oleson, Lane Harrison, Evan M. Peck, Daniel Afergan, Remco Chang, and Robert J. K. Jacob. 2016. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, Santa Clara, California, USA, 5372–5384. https://doi.org/10.1145/2858036.2858388

[263] N. Jane Zbrodoff and Gordon D. Logan. 2005. What everyone finds: The problem-size effect. In *Handbook of Mathematical Cognition*. Psychology Press, New York, NY, USA, 331–345.

[264] Xiao Zhang, Yongqiang Lyu, Xin Hu, Ziyue Hu, Yuanchun Shi, and Hao Yin. 2018. Evaluating photoplethysmogram as a real-time cognitive load assessment during game playing. *International Journal of Human–Computer Interaction* 34, 8 (Aug. 2018), 695–706. https://doi.org/10.1080/10447318.2018.1461763

[265] Robert Z. Zheng. 2017. *Cognitive Load Measurement and Application*. Routledge.

[266] Jens Ziegler, Markus Graube, Alexander Suhrbier, Niels Wessel, Hagen Malberg, and Leon Urbas. 2011. The influence of the spatial separation of control elements on the workload for mobile information systems. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'11)*. Association for Computing Machinery, New York, NY, USA, 191–200. https://doi.org/10.1145/2037373.2037403

[267] F. R. H. Zijlstra and L. Van Doorn. 1985. *The Construction of a Scale to Measure Perceived Effort*. University of Technology.

[268] Manuela Züger and Thomas Fritz. 2015. Interruptibility of software developers and its prediction using psychophysiological sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. Association for Computing Machinery, Seoul, Republic of Korea, 2981–2990. https://doi.org/10.1145/2702123.2702593

Received 20 July 2021; revised 11 January 2023; accepted 18 January 2023