Should we use the NASA-TLX in HCI?

A review of theoretical and methodological issues around Mental Workload Measurement

Babaei, Ebrahim; Dingler, Tilman; Tag, Benjamin; Velloso, Eduardo

DOI
10.1016/j.ijhcs.2025.103515

Publication date
2025

Document Version
Final published version

Published in
International Journal of Human Computer Studies

Int. J. Human–Computer Studies 201 (2025) 103515

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Should we use the NASA-TLX in HCI? A review of theoretical and methodological issues around Mental Workload Measurement

Ebrahim Babaei [a,*], Tilman Dingler [b], Benjamin Tag [c], Eduardo Velloso [d]

[a] The University of Melbourne, Melbourne, Australia
[b] TU Delft, Delft, Netherlands
[c] University of New South Wales, Sydney, Australia
[d] The University of Sydney, Sydney, Australia

ARTICLE INFO

CCS Concepts:
Human-centered computing: HCI theory, concepts and models

Keywords:
Mental workload
NASA-TLX
Multiple resource questionnaire
Cognitive load

ABSTRACT

Mental Workload (MWL) is a construct widely used in HCI to assess the cognitive demand users must exert to perform a task. Research in human factors, however, has suggested several issues regarding its definitions, scales, and applications. This paper, first, introduces debates surrounding the MWL concept and its most popular measure, the NASA-TLX. We present a systematic review of CHI papers involving MWL and highlight severe issues in its application. Finally, through a validation experiment, we assess the convergent validity and sensitivity of two MWL instruments—NASA-TLX and MRQ. Our findings reveal disagreements in the definitions of MWL and severe drawbacks in NASA-TLX and its applications. Our validation study also presents evidence for a lack of convergent validity and sensitivity of MWL subjective scales in HCI tasks. Our findings recommend caution when employing NASA-TLX in user studies and highlight the need for an MWL definition that is agreed upon within the HCI community.

1. Introduction

The tension between modern technology's growing demand for cognitive resources (Young et al., 2015) and the limits of users' cognitive capacities (Brown, 1997; Kahneman, 1973) has made Mental Workload (MWL) measurement a key research topic in Human–Computer Interaction (HCI). HCI researchers and practitioners use MWL measurements both during the early stages of the design process as formative feedback for interaction refinement and at its conclusion to summatively compare alternatives (Kosch et al., 2023). This interest also extends to other disciplines, such as human factors (Stanton et al., 2004), ergonomics (Young et al., 2015), and aerospace engineering (Dismukes, 2017). There are many approaches for measuring MWL, including subjective, psychophysiological, analytical, and performance measures (Xie and Salvendy, 2000; Thorpe et al., 2020; Kosch et al., 2023). Nevertheless, subjective measures administered through questionnaires are the most popular due to their ease of use and face validity (Estes, 2015). Among the many instruments available in the literature, NASA-TLX (Hart and Staveland, 1988) — a multidimensional scale (Hart, 2006) developed in the '70s and '80s for the subjective assessment of workload — arose as the dominant tool for measuring MWL (de Winter, 2014) with over 19,000 citations at the time of writing (Hancock et al., 2021). In HCI, NASA-TLX has become the de facto gold-standard MWL scale and is now a staple instrument in the UX toolkit, alongside other usability measures (Romero, 2017).

Though measuring MWL with NASA-TLX is prevalent in the HCI literature, it relies on an instrument not initially developed for studying interactions with modern digital systems. Further, its popularity and ease of use create the risk that researchers and practitioners might use the tool uncritically without an appropriate understanding of what it measures and awareness of its theoretical background. For example, in HCI, mental workload is often used interchangeably with a variety of constructs, such as cognitive load and cognitive workload, despite their substantially different meanings and disciplinary backgrounds (Wilson, 2023) (see Section 4.3). A 2023 review of the HCI literature by Kosch et al. (2023) also treats these three concepts as synonymous. Recent calls for increased methodological rigour in HCI (Babaei et al., 2021; Cockburn et al., 2018, 2020; Wacharamanotham et al., 2020) create a timely opportunity to take a step back and re-assess community practices surrounding the measurement of MWL and the administration of the NASA-TLX. As such, this paper fills a research gap identified by Kosch et al. (2023) and Hollender et al. (2010) in reassessing the validity of the NASA-TLX in HCI. In contrast to Kosch et al.

* Corresponding author.
E-mail addresses: e.babaei92@gmail.com (E. Babaei), t.dingler@tudelft.nl (T. Dingler), benjamin.tag@unsw.edu.au (B. Tag), evelloso@sydney.edu.au (E. Velloso).

https://doi.org/10.1016/j.ijhcs.2025.103515
Received 16 June 2024; Received in revised form 10 March 2025; Accepted 10 April 2025
Available online 5 May 2025

(2023), who catalogued several modalities for measuring cognitive load, cognitive workload and mental workload, we have a narrower focus on subjective ratings and the NASA-TLX specifically. As well as examining the theoretical foundations of HCI's understanding of MWL, we investigate psychometric properties of the NASA-TLX in an HCI context, contrasting it with another MWL questionnaire, the MRQ.

In particular, we are interested in three questions: (1) what is Mental Workload? (2) how do we measure it? and (3) are these measures valid in HCI? We explored these questions in three stages. In the first stage, we reviewed the theoretical underpinnings of the conceptualisation and measurement of MWL. We found that due to its complex and multidisciplinary history (Hancock et al., 2021), MWL is still an amorphous construct (Hart and Staveland, 1988; Xie and Salvendy, 2000) without a universally agreed-upon definition (Galy et al., 2018). We found a variety of scales for measuring it, but evidence for drawbacks in many of them, particularly NASA-TLX. Notably, we found little evidence for the validity of these scales in an HCI context.

In the second stage, we conducted a systematic methodological review of the CHI literature involving MWL measurement. We found that despite the availability of more modern instruments, NASA-TLX is the most widely adopted instrument by our community. Despite this popularity, there is a lack of methodological standards. We found considerable variability in the constructs operationalised with it, widespread issues in the administration of the instrument, and problems in the analyses and reporting of results.

Given the lack of validation of MWL scales in HCI, in the third stage, we conducted a controlled experiment to assess the convergent validity, reliability, and sensitivity of two MWL scales — NASA-TLX, due to its popularity; and Multiple Resources Questionnaire (MRQ), as a more modern alternative — in a representative HCI task. Our findings paint a dire picture of the applicability of both scales in HCI, finding little evidence for their convergent validity and sensitivity in questions that matter to HCI researchers.

Our findings highlight the need for a deeper consensus in the HCI community about what MWL is, why it is relevant to our work, and how we measure it. They also emphasise the need for methodological rigour and caution when incorporating subjective MWL measurements in user studies. We conclude with advice for how to measure MWL in future work.

2. Open questions about mental workload

This section summarises recent debates in the mental workload literature with implications for HCI. Specifically, we discuss seven critical debates around MWL: (1) the variety of theories used to explain mental workload, (2) the lack of a unified definition of MWL, (3) the opacity in what subjective MWL ratings reflect, (4) the questions of whether MWL is a hypothetical construct or intervening variable, (5) whether it is a unitary or multivariate construct, (6) the two-way relationship between MWL and performance, and, finally, (7) the appropriateness of the NASA-TLX as an instrument for measuring MWL.

2.1. Which theories explain mental workload?

The concept of mental workload (MWL) emerged from human factors research in the 1970s and 1980s, concurrent with the development of the NASA-TLX instrument in an aerospace context (Matthews et al., 2020). This work has been motivated by practical, real-world problems ever since its inception (Hancock et al., 2021). Thorpe et al. (2020)'s review suggests MWL has mostly been used in applied fields, such as engineering and computer science. It also found that MWL is the most frequently used term related to workload capacity in these domains. In its early days, the focus was on assessing the mental load a pilot experiences when flying an automated aircraft (Moray, 1979). Later, its scope expanded to other fields in which people interact with machines (Moray, 2013), such as ergonomics, human factors, and, eventually, HCI.

Mental workload has been related to several constructs, including flow (Keller et al., 2011), arousal (Staal, 2004), and cognitive load (Albers, 2011; Gavas et al., 2017; Byrne et al., 2014). In a recent review of the MWL literature, Longo et al. (2022) identified 22 different theories referenced when explaining mental workload, among which the most popular were multiple resource theory (Wickens, 2002), cognitive load theory (Sweller, 2011), event perception theory (Johansson et al., 1980), and activity theory (Vygotsky and Cole, 1978). In an HCI-focused review, Kosch et al. (2023), treated MWL as interacting with cognitive load, and framed their paper around "cognitive workload". This variety of theoretical backgrounds creates the risk that different works on the same topic might use similar terms for very different concepts.

Early reviews (Kramer, 2020) on MWL draw upon *attentional resource theories* to explain MWL. The earliest model in this category — the *bottleneck model of attention* (Anthony Deutsch and Deutsch, 1963; Broadbent, 1957, 1956) — considers humans as a channel that can transmit information at a rate influenced by arousal. From this perspective, the extent to which a task demands resources from this channel is called *mental effort*. According to this model, failure in performance happens when a task demands more than the capacity of a channel or when a channel is already occupied by another task, in which case the second task will be neglected. A limitation of this model is its inability to explain how people can perform multiple tasks concurrently.

Later, Kahneman (1973) proposed the *capacity model of attention*. Kahneman (1973) considers attention a limited resource that can be freely allocated to concurrent tasks. In this model, mental effort refers to the attention allocated to tasks. A performance failure happens when the demands of the task(s) exceed the limited attention resource.

Wickens (2008) extended these ideas through *Multiple Resource Theory*. This theory proposes that there are multiple limited attention resources instead of a single one (Isreal et al., 1980) and performance improves when concurrent tasks demand different resources compared to when they demand the same resource (Wickens, 2002). For instance, it is easier for a driver to listen to instructions while driving as these activities require separate resources (visual processing vs. auditory processing) than it is to read them because both require the same resource (visual processing). This theory describes mental workload as the relationship between a task's demands and an operator's ability to supply them. Wickens developed a model that describes these resources in four dimensions, namely (a) stages of processing, (b) codes of processing, (c) modalities, and (d) visual channels (see Table 1). Two tasks that both require the same level of a given dimension (e.g. two tasks requiring auditory perception) will interfere with each other more than two tasks that demand separate levels of the dimension (e.g. one visual, one auditory task) (Wickens, 2002). The Multiple Resources Questionnaire (MRQ) is a subjective assessment instrument that measures MWL based on the Multiple Resource Theory developed by Boles and Adair (2001). This questionnaire models MWL in terms of the demand for different attentional resources, such as short-term memory, auditory, and visual processes.

2.2. What is mental workload?

Despite decades of work on the topic and the importance of the concept for HCI, there is still no agreed-upon definition of MWL. In a review of definitions in the literature, Longo et al. found 68 different ones (Longo et al., 2022). One of the earliest models of MWL attributes was proposed by Jahns (1973). This model characterises MWL as involving input load, operator effort, and performance. More recent definitions usually focus on one of these three factors or their interactions, so they serve as a useful way to categorise definitions.

First, MWL can be defined according to the *input load*, that is, to factors externals to the user. For example, Beevis et al. define

Table 1
Dimensions of multiple resource theory.

Dimension	Description
Stages of processing	This dimension indicates that perceptual and cognitive (e.g. working memory) tasks use different resources from those underlying the selection and execution of an action. (Isreal et al., 1980)
Code of processing	This dimension indicates that spatial activity uses different resources than does verbal/linguistic activity (Wickens, 2008)
Modalities	This dimension indicates that in the perception stage of processing, auditory perception uses different resources than does visual perception.
Visual Channels	This dimension indicates if the visual resources are focal (related to object recognition and high acuity perception) or ambient (related to peripheral vision and perception of orientation and movement).

workload as "the task demands placed on an operator" (Beevis, 1999) and Hancock and Caird define it as "a multi-dimensional concept that is largely driven by the characteristics of local task demands" (Hancock and Caird, 1993). These definitions focus exclusively on the task as the source of MWL. Definitions in this category mostly use attentional resource theories as the basis for their description (Stanton et al., 2004). The central assumption in these theories is that an individual has a limited capacity of attention resources (Stanton et al., 2004) and MWL is the percentage of these resources required to meet the demands of a task (Welford, 1978).

Second, MWL can be defined in terms of the *operator effort*, encompassing factors or conceptualising events internal to the user. These definitions use a human-centred rather than a task-centred definition, describing it as the load an operator experiences when performing a task. For instance, Curry et al. define it as "the mental effort that the human operator devotes to control or supervision relative to his capacity to expend mental effort" (Curry et al., 1979). Paas and Van Merrienboer define it as "the total amount of controlled cognitive processing in which a subject is engaged" (Paas and Van Merriënboer, 1993) and Thorpe et al. (2020) define MWL as the amount of resources required to *complete* a task.

The third and least popular type of definition for MWL focuses on *performance*, that is, the output of the task resulting from the effort exerted by the user. As an example of these definitions, Gopher and Donchin (1986) describe MWL as "the difference between the capacities of the information processing system that are required for task performance to satisfy performance expectations and the capacity available at any given time". Wickens (2002) defines workload as the ratio of the time required to perform a task and the time available.

However, most of the highly cited definitions of MWL describe it in terms of the *interactions* between input load, operator effort, and performance. For instance, Moray (2013) framed MWL as the interaction of the input load (which is mainly caused by the task and the environment), the operator's effort (which depends on their personality, background, experience, etc.), and the operator's performance. Hart and Staveland (1988) suggested a model which is similar to Moray (2013)'s three variables — namely imposed workload, operator behaviour, and performance — and added a new variable measuring the operator's perception of the task's goals and structure, performance, and biases. Kosch et al. (2023) define MWL as "workload imposed through the instructional system design of user interface visualisations (e.g., extraneous load) or cognitive demand of users who process information", using a definition more closely aligned with Cognitive Load Theory (Sweller et al., 2011). Most of the well-known MWL scales, such as NASA-TLX and SWAT use this type of definition to describe MWL.

Synthesising various definitions from the literature, Longo et al. proposed that MWL represents "the degree of activation of a finite pool of resources, limited in capacity, while cognitively processing a primary task over time, mediated by external dynamic environmental and situational factors, as well as affected by static definite internal characteristics of a human operator, for coping with static task demands, by devoted effort and attention" (Longo et al., 2022). The sheer number of concepts involved in this definition is evidence of the complexity of the construct.

The disparity in MWL definitions challenges a shared understanding of the concept and makes it difficult to define it precisely. As Xie and Salvendy bluntly put it, *"the simple fact is that nobody seems to know what mental workload is"* (Xie and Salvendy, 2000). Definitions vary depending on the disciplinary background of the research team and the requirements of their research without commonly accepted formal definitions (Cain, 2007; Stanton et al., 2004).

2.3. Hypothetical construct or intervening variable?

Apart from the lack of a unified definition, the very existence of MWL is still under debate. Researchers typically belong to one of two camps—some consider it to be a hypothetical construct (Hart and Staveland, 1988), while others see it as an intervening variable (United States. National Aeronautics and Space Administration, 1988; Kantowitz, 2000). A *hypothetical construct* refers to an explanatory variable that cannot be observed directly, and that cannot be described by a single behaviour, attitude, process, or experience (e.g. intelligence, motivation, creativity). For example, we cannot directly observe creativity; instead, we must infer whether someone is creative from their behaviour, creative production, etc. In contrast, an *intervening variable* is a more restrictive concept that attempts to explain causal relationships between independent and dependent variables, summarising empirical findings (Hyland, 1981). For example, "hunger" can be seen as an intervening variable that summarises several relationships between independent variables such as the time without eating, the amount of food eaten (independent variables), and the behaviour of eating (the dependent variable). To sum up, hypothetical constructs are considered to exist but cannot be observed directly. In contrast, intervening variables do not exist, and researchers use them to better present the relationship between independent and dependent variables.

If we consider MWL an intervening variable, it works as an abstraction that describes what a scale (e.g. NASA-TLX) measures, and it exists as long as the scale exists. In this case, the outcome of each scale has a unique meaning and should not be used interchangeably. For instance, we would consider the score measured by NASA-TLX as having no meaning other than an abstraction to present questionnaire results and not being comparable to other scales (e.g. other questionnaires, physiological sensors) measured. On the other hand, if we consider MWL a hypothetical construct, it must also correspond to other representations (such as physiological measures), so any given scale is just one way of measuring it. In other words, the results of different scales represent the same construct with an identical meaning and must be highly correlated. The view that assumes MWL is an intervening variable infers MWL from changes in performance (Kantowitz, 2000; United States. National Aeronautics and Space Administration, 1988) assuming that poor performance is the result of MWL that is too high or too low (Kantowitz, 2000). On the other hand, those who consider MWL as a hypothetical construct believe it can be measured with different measures such as physiological measures, subjective assessments, and secondary task performance.

Most recent research on MWL tends to consider it a hypothetical construct; however, evidence suggests otherwise. One of the most controversial disputes is the divergence between different MWL scales. If MWL is indeed a hypothetical construct, changing the task demand should lead to a similar effect on different scales that measure this construct. However, several studies show divergence in the measurements of different scales, which suggests that they do not index the same

construct (Matthews et al., 2020, 2015b, 2014; Tsang and Vidulich, 2006). These results point to MWL being an intervening variable. Alternatively, the definitions researchers use for MWL might not be comprehensive enough, each covering a different aspect of a hypothetical construct. We further explain the former view in the coming section.

2.4. Unitary or multivariate construct?

In most definitions, MWL is considered a unitary construct describing the interaction between task demands and an operator's resources to accomplish the task. This unitary definition helps designers compare MWL in different situations with a single number and decide on the design alternative that induces lower MWL. Many scales reflect this view, summarising multiple dimensions of MWL into a single number. For instance, NASA-TLX aggregates different scales into a single rating that characterises the overall MWL.

Despite the practical applications of MWL as a unitary construct, the validity of this view has been questioned. Many researchers believe there is no single scalar that can describe it (Leplat, 1978), proposing that MWL should be considered a multidimensional construct (Galy et al., 2018). The first objection against unitary MWL models is the lack of studies validating this construct with modern psychometric techniques (Matthews et al., 2015b). Though MWL measures show sensitivity to task manipulations and seem individually reliable, they fail to converge (Matthews et al., 2015a). One explanation for this issue is that an increase in MWL can evoke different neurocognitive responses, each indicating different underlying cognitive processes (Matthews et al., 2015a). Therefore, to correctly represent MWL, these authors believe one must consider this construct a multidimensional factor without aggregating them into a single number.

2.5. Do MWL ratings reflect performance?

In an HCI context, maximising users' performances in a given task is a major success criterion for an interactive system. As such, the goal of measuring MWL is often to predict user performance (Cain, 2007; Butmee et al., 2019). In this section, we expound on the general understanding of the relationship between MWL and performance in HCI and clarify deficiencies in that view.

Early research on the relationship between MWL and performance was mainly focused on *overload* (Emerson et al., 1987; Stager, 1991)—situations in which task demands exceed the operator's capacity and hinder performance by causing errors. The aim was to design systems to avoid overload and consequent errors by minimising MWL. The application domain from which this work emerged — automatic aircraft navigation — involved an extremely high error cost, so much weight was put on minimising it. However, as researchers in other areas appropriated MWL, they found that errors were not the only cause of performance decrements as underload — not requiring a certain level of MWL by the system to achieve a certain level of performance — can also be detrimental to performance. De Waard and Brookhuis (1996) argued that there is an optimal level of MWL in which one can guarantee performance by avoiding MWL underload and overload. This view is similar to Yerkes–Dodson's law from psychology, which explains the relationship between cognitive arousal and performance. According to this law, cognitive arousal and performance have an inverted U-shaped relationship, meaning that performance level increases by cognitive arousal to a certain point, and beyond that point, increases in cognitive arousal are detrimental to performance. Therefore, to maximise performance, one must keep cognitive arousal in its optimal range (Hanoch and Vitouch, 2004). Based on our review, an inverted U-shaped relationship between MWL and performance is the assumption behind most publications that use MWL in HCI.

However, despite the ostensible simplicity of this model, one must be aware of three main issues while using it. First, MWL and performance seem to have a two-way causal relationship, which means that at the same time that high MWL can hinder performance, performance failure can increase the perception of MWL (Hancock, 1989b). Then, interpreting overload as the cause of performance decrements may not necessarily be true, especially when measuring MWL through subjective ratings.

Second, research has shown that users monitor task demands and adopt different strategies to cope with overload and underload in order to keep their performance at the desired level (Hancock and Matthews, 2019). These strategies can involve investing additional resources at the cost of individual strain (Hancock, 1989a) or changing their goals (Sperandio, 1978). In this case, overload and underload do not necessarily result in performance decline. In addition, Howard et al. (2020) found that the cognitive mechanism underpinning MWL varies in different task manipulations. For instance, the strategy an individual adopts when facing an increase in MWL due to adding a new task is different from the one they adopt when the difficulty of the same task increases. Howard et al. (2020) believes the changes in the MWL ratings due to different manipulations are not comparable.

Third, the traditional view considers MWL to be a unitary construct; however, research on Multiple Resource Theory suggests that MWL is not a unitary construct, so not only the level of task demands is important, but also which resources it demands and how they overlap. As such, one may observe significant drops in performance in cases where the demanded loads are optimal just because there is an overlap in the resources the task demands.

2.6. What do subjective ratings measure?

The literature contains numerous examples of MWL ratings describing the task as a whole. However, Xie and Salvendy argue that a single number cannot fully characterise the workload experience (Xie and Salvendy, 2000). The authors proposed a conceptual framework to paint a more fine-grained picture of how mental workload fluctuates within a task. In their framework (see Fig. 1(a)), the *instantaneous workload* dynamically varies over time. The *peak workload* is the maximum value it takes during the task. The *accumulated workload* is the area under the curve, representing the total amount of information processed during the task. Dividing the accumulated workload by the total amount of time, we obtain the *average workload*, which corresponds to the intensity of the workload. The authors argue that subjective MWL ratings do not exclusively capture any of these but are influenced by all of them. However, it is unclear how much influence each has on these subjective ratings. Nevertheless, instruments like the Instantaneous Self-Assessment (ISA) are designed to be administered multiple times during a task to build an instantaneous workload profile (Tattersall and Foord, 1996). This distinction matters in HCI because minimising peak, average, or accumulated workload implies different application re-design solutions.

A further issue in the measurement of MWL is how researchers compare ratings. For example, if two tasks result in NASA-TLX scores of 80 and 60, respectively, one might be inclined to say that one task leads to a 33% higher workload than the other. However, Estes (2015) showed evidence that the relationship between MWL and its subjective ratings is not linear but instead best represented by an S-shaped curve (see Fig. 1(b)). This S-shaped curve, which is also called a psychometric function, is a common mathematical function in psychology and psychophysics to represent the relationship between stimuli and response levels (Wichmann and Hill, 2001; Klein, 2001). MWL and its subjective ratings have a relationship that is mathematically described by psychometric function. This means that the distances between different points on subjective MWL scales are not equal—the increase in workload required to move a rating from 3 to 4 is not the same as the one required to move a rating from 4 to 5. This has implications for how researchers compare and average ratings.

Hertzum (2021) reviewed NASA-TLX values in 556 papers to determine reference values for NASA-TLX. His findings showed that the values were symmetrically distributed around an average of 42; however,

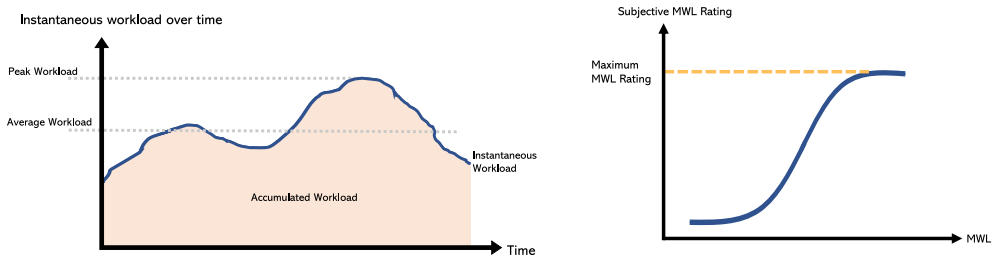


Fig. 1. (a) Mental Workload Parameters Framework (Xie and Salvendy, 2000) (b) Subjective MWL curve (Estes, 2015).

this number varied depending on application domains, technologies, regions, and experimental setups (in-lab vs. in-the-wild). In a similar attempt, Grier (2015) reviewed 237 papers (1173 overall MWL scores) to provide the basis for interpreting NASA-TLX ratings. Grier found scores to range between 6.21 and 88.5; however, these values depend on the task type, operator expertise, and stressors. Grier also found two samples in the literature where participants had to do nothing but wait, while their NASA-TLX scores were significantly higher than the minimum (12.0 and 14.8) and concluded demands are not the only component of MWL (Grier, 2015). Therefore, interpreting subjective MWL ratings is difficult as it is still unclear which MWL attributes subjective MWL ratings represent, the variation in subjective MWL is not linear, and reference values vary from case to case.

2.7. Is the NASA-TLX a good instrument for measuring MWL?

Developed by Hart and Staveland (1988) over three years of research and validated through 16 studies (de Winter, 2014), NASA-TLX is the most well-known and widely used when researchers measure MWL. Similar to other early attempts to operationalise workload, such as Cooper-Harper (Cooper and Harper, 1969) and SWAT (Reid and Nygren, 1988; Anon, 2021b), NASA-TLX was developed to assess pilots' experiences in automated aircraft. Unlike other short-lived scales, NASA-TLX became popular to the extent of being almost synonymous with MWL (de Winter, 2014). However, through an exploratory search in literature, de Winter (2014) suggested that this popularity might be attributed to the "Matthew effect" (Merton, 1968) (the higher likelihood for NASA-TLX to be used in correlation with its initial popularity level due to being published by NASA) rather than to the scale's validity and sensitivity: because NASA-TLX could reach a wider audience, it became the obvious choice for measuring MWL. This fame resulted in further popularity, even if it may not be the most reliable MWL scale.

NASA-TLX uses six subscales to subjectively measure workload, namely *mental demand*, *physical demand*, *temporal demand*, *performance*, *effort*, and *frustration level* (Anon, 2021a). A description of the subscales can be found in Table 2. These six subscales were selected as the factors most relevant to workload from a larger set of factors previously found in the literature (Hart and Staveland, 1988). Later, through several trials, researchers modified these factors for different applications and developed customised versions of the NASA-TLX, such as the SURG-TLX (Wilson et al., 2011) and the SIM-TLX (Harris et al., 2019) for measuring workload in surgery and VR environments, respectively.

NASA-TLX aims to measure *overall subjective workload*, but it is debatable whether this is a different construct from MWL. Some might consider them to be different because the NASA-TLX incorporates the physical demands of the task. However, we consider them to be the same for several reasons. First, the definitions of MWL we discussed so far do not exclude physical demands. For example, we would expect that any principally cognitive task would yield higher MWL if the same task was performed during physical exertion. Second, it is unclear whether Hart and Staveland considered subjective workload and MWL to be different. For example, their justification for measuring workload using a questionnaire was that "subjective ratings may come closest to tapping the essence of *mental workload* and provide the most generally valid, sensitive and practically useful indicator" (our emphasis). Third, they consider physical fatigue to be separate from their concept of workload. When discussing their findings, they claimed that "it appeared that subjects regarded fatigue as a separate phenomenon from workload". Hart and Staveland (1988).

NASA-TLX's original procedure recommends a two-part evaluation process, called *weighting* and *rating*. In the weighting process, after pairwise comparisons, each subscale receives a weight from highest (5) to lowest (0), depending on how relevant they are to the task. The highest weight (5) indicates the most relevant subscale and the lowest (0) indicates the least relevant subscale to that specific task from the operator's point of view. NASA-TLX adopted this weighting process to reduce between-subject variability (Hart and Staveland, 1988). Later, in the rating process, operators rate their perceived demands of each subscale on a 21-point scale, and finally, the weighted average of these scores is taken as the overall workload score.

Multiple studies have evaluated the reliability of NASA-TLX for measuring MWL, identifying major drawbacks in its scales and measurements (Bustamante and Spain, 2008; Moroney et al., 1995; Nygren, 1991; Galy et al., 2018; Hayashi and Kishi, 2014). However, to date, no attempt to improve NASA-TLX and address these issues has been carried out. In the remainder of this section, we overview existing criticisms of NASA-TLX.

De Waard and Lewis-Evans (de Waard and Lewis-Evans, 2014) argued that MWL is a dynamic, complex concept, and it is too simplistic to believe that a questionnaire can fully characterise it. They show the self-regulation of operators in overload and underload conditions as an example of the inability of a questionnaire to capture all facets involved in MWL measurements. Performance is one of the subscales of NASA-TLX, and it assumes that high performance indicates low MWL. However, individuals may maintain performance for periods of time by self-regulation, e.g. by investing more effort in high-MWL conditions—a fact NASA-TLX is incapable of capturing. De Ward and Lewis-Evans (de Waard and Lewis-Evans, 2014) suggest that to properly measure MWL; one must incorporate multiple measures such as subjective scales, task performance, and physiological signals.

Although NASA-TLX is a multidimensional scale, it considers workload as a unitary construct because it aggregates multiple measures into a single one. Boles et al. argue that NASA-TLX considers attention to be a global undifferentiated pool of resources that can be assigned to the tasks as demanded (Boles et al., 2007), as described by Kahneman (1973)'s capacity model. Hart and Staveland (1988) explained in their report on the development of NASA-TLX that they studied correlations between MWL and various factors that had been found significant in measuring MWL in the literature beforehand, aiming to pick a maximum of six items. MWL measured in the original version of NASA-TLX is the weighted average of these items. Therefore, the resulting subscales are simply the six factors most correlated with

Table 2
Subscales of NASA-TLX and MRQ and their definitions.

Questionnaire	Subscale	Definition
NASA-TLX	Mental demand	How much mental and perceptual activity was required?
	Physical demand	How much physical activity was required?
	Temporal demand	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?
	Performance	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)?
	Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?
	Frustration level	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?
MRQ	Auditory emotional process	Required judgments of emotion (e.g., tone of voice or musical mood) presented through the sense of hearing.
	Auditory linguistic process	Required recognition of words, syllables, or other verbal parts of speech presented through the sense of hearing.
	Facial figural process	Required recognition of faces, or of the emotions shown on faces, presented through the sense of vision.
	Facial motive process	Required movement of your own face muscles, unconnected to speech or the expression of emotion.
	Manual process	Required movement of the arms, hands, and/or fingers.
	Short-term memory process	Required remembering of information for a period of time ranging from a couple of seconds to half a minute.
	Spatial attentive process	Required focusing of attention on a location, using the sense of vision.
	Spatial categorical process	Required judgment of simple left-versus-right or up-versus down relationships, without consideration of precise location, using the sense of vision.
	Spatial concentrative process	Required judgment of how tightly spaced are numerous visual objects or forms.
	Spatial emergent process	Required "picking out" of a form or object from a highly cluttered or confusing background, using the sense of vision.
	Spatial positional process	Required recognition of a precise location as differing from other locations, using the sense of vision.
	Spatial quantitative process	Required judgment of numerical quantity based on a nonverbal, nondigital representation, using the sense of vision.
	Tactile figural process	Required recognition or judgment of shapes (figures), using the sense of touch.
	Visual lexical process	Required recognition of words, letters, or digits, using the sense of vision.
	Visual phonetic process	Required detailed analysis of the sound of words, letters, or digits, presented using the sense of vision.
	Visual temporal process	Required judgment of time intervals, or of the timing of events, using the sense of vision.
	Vocal process	Required use of your voice.

MWL among a pool of factors rather than being chosen based on theoretical considerations. However, as proposed in MRT, multitasking studies have demonstrated that attention is not a unitary resource and pairing certain tasks causes more performance loss due to the overlap of demands (Finomore et al., 2013, 2009). This implies that NASA-TLX assumptions are based on outdated theories of attention.

The subscales in the NASA-TLX were selected with limited elaboration on the rationale behind the choice of factors (de Waard and Lewis-Evans, 2014). As such, the main criterion for deciding whether a factor influences MWL seems to be whether it had been considered in the literature of that period. This issue is not limited to NASA-TLX. As Van Acker et al. (2018) point out, the selection of defining variables in MWL scales tends to be arbitrary. Further, the descriptions of these variables come in various levels of abstraction (e.g. some definitions operationalise it as the more general concept of "performance", while others are more specific in differentiating "subjective" from "objective" performance).

Another issue with the subscales in the NASA-TLX is that the number of subscales was determined a priori to keep the questionnaire short enough to be completed in operational environments, as opposed to including all possible factors that might influence MWL. Hart and Staveland (1988) conducted multiple studies to select a maximum of six subscales to represent MWL. This means that many factors potentially significant for measuring MWL were ignored in the design process of NASA-TLX.

A source of confusion in administering NASA-TLX is its obsolete weighting process. While this process was adopted to adjust each subscale's contribution to the overall MWL and decrease between-subject variability, further studies showed that it performs similarly without using these weights (Nygren, 1991). This is mainly due to the pairwise comparison used to weight scales. In this weighting process, respondents pick the most relevant subscale between each pair of the six subscales (15 pairs in total). The number of times a subscale is selected is taken as its weight. Therefore, if an individual rates the subscales consistently, they will be ranked from 5 (the most important) to 0 (the least important). Consequently, the subscale with weight zero will be removed, while the most important subscale will receive at most 33.33% of the total weight—an arguably arbitrary range. Moreover, the difference between the most and least important subscales among the remaining 5 is at most 26% of the total weight, which may not represent the contribution of each factor correctly. Nygren (1991), through an analysis of NASA-TLX scores, showed that the improvements identified during the NASA-TLX development in between-subject variability were simply because of linear averaging, not its weighting process. The Raw-TLX (also known as Raw NASA-TLX and R-TLX) attempts to address this issue by ignoring the weighting step of NASA-TLX and simply averaging all subscales. The R-TLX is now the recommended approach for administering the instrument.

Further issues with NASA-TLX stem from the inconsistencies across different questionnaire versions. NASA published two inconsistent versions on their webpage (Anon, 2021c), one of which marked the extreme ends of each subscale with *very low* and *very high* and the other with *low* and *high*. Further, while each of the six NASA-TLX subscales ranges from 0 to 100 in 5-unit increments (21 increments), a published version describes its subscales as *"five 7-point scales"*. Last but not least, five of the subscales in the NASA-TLX range from "low" to "high", while the performance subscale ranges from "good" to "poor"—not only different values but also a different conceptual order. This creates a potential risk of accidental errors in ratings (de Waard and Lewis-Evans, 2014).

3. MWL measurement at CHI

The review of the broader MWL literature revealed various pitfalls in the conceptualisation and measurement of this construct. Given its prevalence and importance in the CHI community, it is critical to assess how we as a community use and measure MWL. For this purpose, we conducted a methodological review of CHI papers involving the measurement of MWL. Our goal was to understand what CHI researchers measure when they collect MWL data, how they measure it, and whether these measurements are analysed and reported appropriately.

3.0.1. Source selection

Analogous to previously conducted reviews of HCI research (Kjærup et al., 2021; Bowman et al., 2023; Mack et al., 2021; Babaei et al., 2021), we reviewed full papers published in the proceedings of the ACM CHI Conference on Human Factors in Computing Systems as representatives of high-quality and mature research practices in the field of HCI. As such, we limit our claims to HCI work published at CHI and acknowledge that this excludes other venues interested in HCI research (e.g. Ubicomp, IMWUT, UIST, ISWC, etc.). We conducted our search in the ACM Digital Library, selecting papers with the terms "Mental Workload", "Subjective Workload", "NASA Task Load Index", "NASA TLX", "NASA-TLX", "Subjective Workload Assessment Technique", "SWAT", "Cooper Harper", "Cooper-Harper", "Subjective Workload Dominance", "Bedford Workload Scale", "Bedford", "Multiple Resource Questionnaire", "MRQ", or "Workload Profile" in any part of the paper. We restate that our focus in this paper is solely on the MWL construct, and these terms were selected to exclude other related constructs, such as cognitive load. However, we acknowledge that these terms are not inclusive of all HCI research employing MWL since terms related to other constructs may have been used instead of MWL in some articles as it is proved to be common in HCI (Aeschbach et al., 2021).Our search spanned all CHI proceedings from 1981 to 2020 and was performed on 11/26/2020. Specifically, we used the following search string:

```
"query": { AllField: ("Mental Workload" OR "Subjective
  Workload" OR "NASA Task Load Index"
 OR "NASA TLX" OR "NASA-TLX" OR "Subjective Workload
  Assessment Technique" OR "SWAT" OR
  "Cooper Harper" OR "Cooper-Harper" OR "Subjective
  Workload Dominance" OR "Bedford Workload Scale"
 OR "Bedford" OR "Multiple Resource Questionnaire" OR
  "MRQ" OR "Workload Profile") }
"filter": { Conference Collections: CHI: Conference
  on Human Factors in Computing Systems }
```

3.0.2. Screening criteria

This search resulted in 605 papers. We further restricted this sample to full research papers, resulting in 442 papers. The first author inspected each paper in the sample and inspected the method section, removing those that did not measure MWL, resulting in 311. Among these, 270 were published from 2011 to 2020, and only two used instruments other than NASA-TLX to measure MWL. We randomly sampled 75 papers listed in the supplementary material for an in-depth evaluation of their use of MWL. We chose this sample size based on the recommended sample size for a population size of 275 with a confidence level of 95%, $p = 0.05$ and precision of $\pm 10\%$ (Israel, 1992).

3.1. Results

As mentioned before, only two papers (out of 270) used instruments other than NASA-TLX to measure MWL in our sample before random sampling. Braun et al. (2019) used DALI (Driving Activity Load Index (Pauzie', 2008)), which is a different version of NASA-TLX developed specifically for driving and Züger and Fritz (2015) used their own scale to measure MWL. None of these papers was included in our sample (75 papers); therefore, in the remainder of this section, we focus on methodological practices involving NASA-TLX. One paper in our sample did not provide any details or analysis on their NASA-TLX results, which we excluded from this discussion.

3.1.1. Measured construct

Although NASA-TLX was developed to measure the overall subjective perception of the workload of a task, the studies sampled used it for measuring various constructs. The terms used to describe the construct being measured by NASA-TLX varied across the sample, suggesting little agreement in terminology. Terms related to workload appeared in 32 papers (subjective workload (11), perceived workload (7), workload (10), perceived overall workload (1), overall workload (2), perceived subjective workload (1)) and seven papers used terms related to MWL (MWL (4), subjective MWL (1), mental demand (1), mental load (1)). Cognitive load is another construct that has been operationalised using NASA-TLX in five papers (cognitive workload (2), cognitive load (2), perceived cognitive load (1)) of our sample. Task-related terms were used in 12 papers (task load (2), task workload (3), task difficulty (1), perceived task difficulty (1), perceived task load (2), difficulty (2), subjective task load (1)). Nine papers used the subscales of NASA-TLX independently without referring them to any construct, and three papers used effort-related terms (effort (2), subjective effort (1)) to refer to NASA-TLX measurements. Five papers in our sample did not mention any construct operationalised by NASA-TLX results.

3.1.2. Questionnaire

As discussed in previous sections, the original version of NASA-TLX has a weighting procedure for reducing between-subject variability; however, further research showed that NASA-TLX is similarly valid without this weighting process. Although researchers can choose to use or skip this weighting process, e.g. by using the R-TLX, it is important to clearly describe the version used as the range of results differs. In general, most analyses were not described in sufficient detail to understand whether any weighting had been performed. Given the lack of a description, we assume that no weighting had been performed unless explicitly stated. Only 14 papers explicitly reported using R-TLX, and only two reported using the weighted version.

NASA-TLX questionnaires can be administered in many formats, including pen and paper, desktop/web software, and mobile applications. Previous research has shown that the medium through which the questionnaire is administered has a statistically significant effect on the ratings (Noyes and Bruneau, 2007). In an experiment comparing the same task being rated with a computer-based and a paper-based version of the questionnaire, Noyes and Bruneau found that participants reported higher workload ratings when using the computer version (Noyes and Bruneau, 2007). Therefore, it is essential to describe how the questionnaire was administered. Only five papers in our sample explicitly mentioned the form of the questionnaire they used.

Modifying a psychometric test can have significant implications for its validity and reliability. In 14 papers in our sample, we found descriptions of modified scales. Seven reported using 7-point subscales—though it is unclear whether they used a 7-point subscale or copied the incorrect description from NASA's website. Four used 10-point subscales, two used 20-point subscales and one used subscales 11-point subscales. Five papers in our sample also reported using subscales not in the standard NASA-TLX (satisfaction (Vashistha et al., 2017), hard work and success (Malacria et al., 2013), annoyance (Di Campli San Vito et al., 2019; Brehmer et al., 2012), liked (Kosch et al., 2018) and fatigue (Brehmer et al., 2012)).

3.1.3. Analysis

Hart and Staveland (1988) selected six factors as subscales of MWL based on five criteria: sensitivity to task and task manipulations, independence from other factors, correlation with the overall workload, and their importance for participants. As such, the relationship between these factors and MWL is more correlational than causal. Further, they suggested that each subscale can contribute unequally to MWL, so a weighted average of these subscales is warranted. Therefore, when using NASA-TLX, it is essential to calculate the overall MWL score and use that as the basis of the comparisons for MWL. Many papers analyse the subscales themselves to understand what may cause differences in overall workload scores. In 20 papers in our sample, authors skipped the calculation of the overall MWL score and compared subscales independently across conditions. Six papers used only a subset of the six subscales for comparing MWL. Two papers used only the *Mental Demand* subscale. It is possible that the similarity between the terms "mental workload" and "mental demand" led to this misunderstanding.

4. Discussion of reviews

So far, we reviewed the literature outside HCI regarding current debates around MWL and highlighted the challenges around MWL and its most well-known scale, the NASA-TLX. In addition, through a review of 75 CHI papers, we exposed critical issues in the applications of NASA-TLX in CHI papers. In this section, we highlight the most critical issues and discuss potential consequences if they are left unaddressed. These include ambiguity in the definition and concepts of MWL, lack of reliable measures, and limited understanding of the concept and scales. We acknowledge that the challenges we describe were identified in the CHI literature. Analysing MWL practices in other HCI venues such as Ubicomp, IMWUT, UIST, and ISWC may lead to different outcomes. In addition, we highlight that this review aims to identify existing issues rather than quantify their extent. The precision of our sampling method is $\pm 10\%$, which may not be enough for a precise quantitative estimate of the prevalence of issues in the literature; however, it can give us a qualitative understanding of improvement areas in the use of MWL in the CHI community.

4.1. Ambiguity in definitions and concept

Although the concept of MWL has existed for many years, there is no agreement on its definitions, models, or even existence. Though

resolving these issues might be outside of HCI's scope, it is nevertheless essential to be aware of these debates and agree on why, when, and how to define and measure MWL for our purposes. Current norms cannot satisfy our community's needs as they lack consistency, and we cannot rely on empirical research as the definitions vary from case to case. Our review of CHI papers emphasised the issues this lack of a standard definition may cause. For example, 22 different terms were used to describe what NASA-TLX measures, resulting in uncertainty and ambiguity in the theoretical underpinnings of MWL research in CHI papers. In addition, theories of MWL are not fully embraced and accepted by the community and are barely used to discuss or frame the results.

4.2. Lack of reliable measurement instruments

Another critical issue that limits the use of MWL is the lack of a reliable measurement instrument. Our analysis reveals that NASA-TLX is the most widely used instrument to measure MWL, and CHI papers rarely have used other standard subjective MWL questionnaires. The keywords representing other standard subjective MWL scales did not have any hit in our review. NASA-TLX dominated this area so much that any alternative appears incorrect (de Waard and Lewis-Evans, 2014)—even though newer scales showed better performance in other fields (Rubio et al., 2004; Moustafa et al., 2017; Longo and Orru, 2019). Further, despite much evidence of the limitations of the NASA-TLX, there has been little work on improving it (de Waard and Lewis-Evans, 2014). Inconsistent versions of NASA-TLX make interpreting and comparing MWL scores in different papers troublesome. Values of NASA-TLX score depend on which grade researchers use for NASA-TLX subscales (7-point grade, or 21-grade), the range of these subscales (1 to 7 or 0 to 100) and the inclusion/exclusion of the weighting process. In addition, we found instances in our review of researchers using arbitrary subscales (e.g. 11-point grade ranging from 0 to 10), which lead to MWL rates in various ranges. Given that these decisions are not reported in most papers, one cannot confidently compare or aggregate results. Finally, the NASA-TLX was initially developed for a context substantially different from the common use cases in which it is applied in HCI. All points above point to the need to rethink the choice of MWL gold standard in our field. Newly refined questionnaires can be developed or customised specifically for HCI needs, or recently developed questionnaires should be validated and assessed in HCI tasks.

4.3. Limited understanding of the concept and scales

Our review of CHI papers revealed a limited understanding of the concept of MWL in CHI papers. NASA-TLX has been used to measure various constructs such as cognitive load, difficulty, and mental demand despite their different meanings. For example, cognitive load is a construct that is frequently misused as a synonym for MWL. Developed by Sweller (2011), Cognitive Load Theory stems from educational psychology with the primary concern of adapting instructions to the constraints of the learner's cognitive system (Schnotz and Kürschner, 2007). According to Sweller (2011), cognitive load consists of three components: intrinsic, extraneous, and germane cognitive load. While intrinsic and extraneous cognitive load are affected by the presented information or material, germane cognitive is defined by the users' skill and ability levels (Sweller, 2010). Klepsch et al. (2017) developed and validated a questionnaire aimed at identifying how the innate complexity of the information presented (**intrinsic**), instructions given (**extraneous**), and the effort used to develop correctly functioning mental models through learning (**germane**) (Sweller et al., 1998; Sweller, 2010; Paas and van Gog, 2006). Cognitive load builds on different theories, is measured through subjective scales, and involves applications mostly limited to learning. Nevertheless, it is often confused with MWL due to the similarity in their nomenclature. Furthermore, some cognitive load theory researchers have used modified versions of NASA-TLX subscales to measure various components of cognitive load (Sweller et al., 2011). For instance, Gerjets et al. used three subscales of NASA-TLX and modified them to measure different components of cognitive load (Gerjets et al., 2004, 2006). In a review of the HCI literature on MWL measurement conducted in parallel to ours, Kosch et al. (2023) treated "mental workload", "cognitive workload", and "cognitive load" interchangeably and provided a common definition that mixes cognitive load theory with multiple resource theory. The authors acknowledged that these terms stem from different theoretical backgrounds, but given that the community uses them in a mixed way and their review aimed to be descriptive of common practices, they left a more normative approach out of their scope.

However, we find this terminology choice unfortunate. We advise against using MWL and cognitive load interchangeably as it is not yet apparent to what degree these two constructs refer to the same concept (Galy and Mélan, 2015), and current research on their relationship shows it is far more complicated than their being identical (Galy et al., 2012; Galy and Mélan, 2015). In addition, a study on the validity of NASA-TLX in measuring cognitive load shows its sensitivity depends on the single components of cognitive load and each of their intensities (Wiebe et al., 2010). Consequently, for subjective measurements of cognitive load, we recommend using scales that have been developed and validated specifically for cognitive load measurements, such as the mental-effort rating scale developed by Paas (1992), informed and naive rating questionnaires designed by Klepsch et al. (2017), and the 10-item questionnaire proposed by Leppink et al. (2013).

Several papers used NASA-TLX without stating any research questions related to MWL but rather as part of the usability evaluation of a system. These papers lacked any reasoning as to why or how MWL would impact usability. Regardless of whether MWL is relevant to the study hypothesis, this haphazard use seems to be accepted — and often *expected* by reviewers — as a means to evaluate systems. However, arguments both against (Longo, 2018, 2017; Longo and Dondio, 2015) and in favour (Kokini et al., 2012) of the association between usability and MWL exist in the literature. Longo et al. investigated this relationship in the context of web design in three different studies (Longo, 2018, 2017; Longo and Dondio, 2015). They concluded that these two concepts are independent of each other and the traditional view that considers increments in required MWL to correspond with decrements in usability perception and vice-versa may not be valid in most cases.

Our review also found examples that used a modified version of NASA-TLX with no validation. In some cases, the subscales were changed, only a subset was selected, or the range and number of points on the scales were modified. These changes affect the validity of NASA-TLX measurements, and it is recommended that only validated versions (NASA-TLX and R-TLX) be used. In some cases, the *mental demand* subscale of NASA-TLX was used to operationalise MWL. While the overall score of NASA-TLX is an accepted measure of MWL, having a subscale named *mental demand* may induce an incorrect impression that this subscale alone can measure MWL. To be clear, mental demand refers to the characteristics of the task, whereas MWL refers to its effect on the user/operator. Another recurring problem in interpretations of NASA-TLX scores was the consideration of subscales as dimensions of MWL. These subscales were selected during the instrument development as they were highly correlated with workload and independent from each other. Therefore, we cannot view them as dimensions of MWL as their relationship with overall MWL is solely correlational.

4.4. Summary

MWL has broad applicability and rich potential for HCI researchers and practitioners to manage cognitive demands in interaction design. However, to avoid pitfalls and best leverage this potential, it is necessary to adopt reliable measures backed by established theories and validate them in the HCI context. This enhances the interpretability of results alongside their replicability and validity. NASA-TLX has already

shown weaknesses in other disciplines (Bustamante and Spain, 2008; Moroney et al., 1995; Nygren, 1991; Galy et al., 2018; McKendrick and Cherry, 2018) and has yet to be validated in an HCI context. Therefore, in the remainder of this paper, we present a validation study of NASA-TLX and a more recent alternative — the MRQ — assessing their convergent validity, sensitivity, and reliability. This validation experiment aims to foster a movement to clarify the concept of MWL and validate the scales that measure it in various contexts of HCI. We invite the community to extend this line of research by conducting similar experiments in their area of interest.

5. Validation of two MWL instruments on an HCI task

Previous sections highlighted concerns around the reliability of findings related to MWL in HCI studies due to ambiguities in its conceptualisation and deficiencies in its measurement. In addition to these drawbacks, the lack of validation studies of these instruments on tasks relevant to HCI exacerbates this problem. Though studies on the validity of MWL instruments in other disciplines are abundant in the literature (Bustamante and Spain, 2008; Moroney et al., 1995; Nygren, 1991; Galy et al., 2018; McKendrick and Cherry, 2018) and some research exists on the suitability of NASA-TLX for different applications relevant to HCI (Ramkumar et al., 2017; Afridi and Mengash, 2020; Fréard et al., 2007; Hayashi and Kishi, 2014), to the best of our knowledge, no study tested the psychometric properties of MWL scales on canonical HCI tasks. To address this issue, we designed an experiment to assess the most frequently used MWL scale in HCI, the NASA-TLX, against the Multiple Resources Questionnaire (MRQ), a more recent instrument.

We chose the MRQ for several reasons. First, unlike prior scales, the MRQ operationalises a resource-based definition backed by a more robust theoretical background, namely, Multiple Resource Theory (Wickens et al., 1984). By splitting items into specific perceptual/cognitive resource domains, the MRQ has the potential to not only identify differences in MWL but also explain why they occur. This is particularly valuable in an HCI context, where tasks commonly involve multiple sensory modalities. Second, previous studies have shown it to be more effective than the NASA-TLX in capturing workload differences in multitasking vs. single-task conditions—which is a common feature in HCI tasks (Finomore et al., 2013).

In this experiment, we aimed to test these two questionnaires' sensitivity, internal consistency, and convergent validity to changes in the difficulty of a canonical HCI task. To preclude any ambiguity, in the rest of this section, the term *difficulty* refers to literal changes in the mechanisms of a task. For instance, by difficulty level in the N-back task, we are referring to the 1-back (Easy) versus 3-back (hard) versions of this task. In addition, as resource-based questionnaires rely on the participants' fair judgment of the required resources for a task, we investigate whether these scales have test-retest stability regardless of the mental workload induced through a separate previous task. Further, we assess whether there is a correlation between the subscales of these questionnaires and between performance and overall MWL. In summary, this experiment aims to shed light on the following questions:

- **RQ1:** Do MRQ and NASA-TLX converge in measuring MWL in HCI tasks? If both scales measure the same construct, their scores should be strongly correlated.
- **RQ2:** Are MRQ and NASA-TLX sensitive enough to capture the changes in MWL caused by differences in an HCI task's difficulty? We hypothesise that if they are sensitive enough to variations in task difficulty, tasks with statistically significant differences in performance measures should also present statistically significant differences in scores on these scales.
- **RQ3:** Are MRQ and NASA-TLX reliable in measuring MWL in HCI tasks? We test two types of reliability for this RQ, namely internal consistency and test-retest reliability.
  - **Internal Consistency:** Do all subscales within NASA-TLX or MRQ measure the same construct? We hypothesise that if a scale is internally consistent, its subscales should be closely related as a group (Cronbach's $\alpha > 0.7$).
  - **Test-retest Stability:** Are MRQ and NASA-TLX stable enough to produce similar rates for separate rounds of the same HCI task? We hypothesise that if NASA-TLX and MRQ are stable, they will not produce statistically significant different scores for two rounds of an HCI task with identical difficulty.
- **RQ4:** Do MWL scales allow us to make predictions about performance? If so, we hypothesise that subjective ratings should be correlated with task performance.
- **RQ5:** Could the MRQ and NASA-TLX complement each other? If so, we should be able to observe a low Variance Inflation Factor (VIF) in some subscales of both questionnaires as we combine them.

5.1. Method

To address the RQs mentioned above, we developed an online experiment using two instruments (MRQ, NASA-TLX), and two tasks (SAK and Dual N-back), each with two difficulty levels, and recruited 100 participants. To induce higher or lower MWL, we manipulated the task difficulty. We used a mixed design with SAK difficulty as the between-subjects independent variable and Dual N-back difficulty as the within-subjects independent variable. Participants completed the questionnaires only for SAK tasks (not for the Dual N-back tasks) and upon completion. We elaborate on our method below. In the remainder of the paper, we refer to "easy/hard difficulty" as a property of the task and "high/low MWL" as a property of the participant.

5.1.1. Tasks

In the HCI literature, MWL questionnaires are often administered to evaluate or compare new interaction techniques. A typical study of this kind asks participants to complete the same task (e.g. selection, manipulation, text entry, etc.) with 2-5 alternative techniques and compares them in terms of objective performance (e.g. task completion time, error rate) and subjective measures (e.g. NASA-TLX, SUS, UEQ, etc.).

As a task representative of this approach, we chose a canonical HCI task—text entry. To be representative of studies that evaluate novel techniques, we needed a technique that participants would be unfamiliar with. In addition, to be suitable for an online experiment, the technique should only require a minimum apparatus, be deployable on a website, and be easy enough to learn without intervention from the research team. As such, we chose the *Scanning Ambiguous Keyboard* (SAK) technique developed by MacKenzie and Felzer (2010), which has been frequently used in HCI research (Waddington et al., 2017; MacKenzie, 2009; Jabeen et al., 2018; Jabeen and Tao, 2017) and for which a publicly accessible implementation (MacKenzie and Felzer (2022)) was available to ensure reproducibility. We developed a web version of this task, which we made publicly available.[1] Additionally, we used a Dual N-back task (Jaeggi et al., 2003; Blacker et al., 2017) with two levels of difficulty before the SAK task to assess the reliability of the questionnaires.

SAK: SAK is a text entry technique that enables users to type using a single key (the spacebar in our case). Its interface (see Fig. 2) contains a letter selection region and a word selection region. The letter selection region consists of four tiles, three of which represent a set of letters organised alphabetically and the fourth representing the SPACE character. In operation, the tiles sequentially illuminate ("scan"), and

[1] https://github.com/ebi-b/Experiment_Implementation

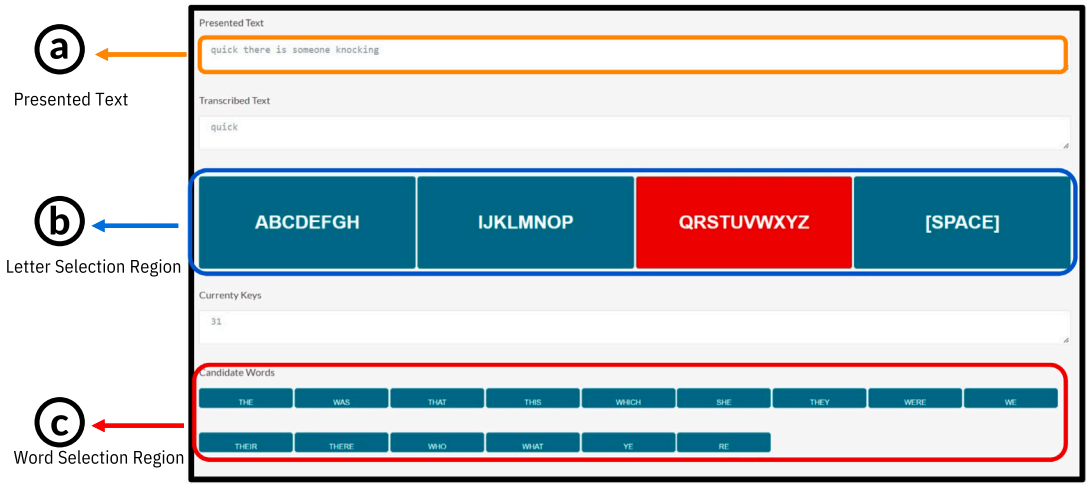


Fig. 2. Implemented SAK Task Interface—**(a)** Presented Text Region displays a sentence the participant must insert using the SAK interface. **(b)** The Letter Selection Region contains four tiles that sequentially illuminate. Participants can select a letter by pressing the spacebar when the tile bearing the desired letter is highlighted. **(c)** The Word Selection Region is a set of tiles, each representing one potential word that can be spelt with the selected letters.

participants can select a letter by pressing the spacebar when the tile bearing the desired letter is highlighted. After selecting each tile, its number will be added to the Current Keys section and form a sequence. At each point, the most frequent words start with the potential combinations of letters in the Current Key sequence is shown in the word selection region. For instance, in Fig. 2, tile number 3 (QRSTUVWXYZ) and tile number 1 (ABCDEFGH) are selected sequentially, and the most frequent words that start with one of the letters in tile QRSTUVWXYZ and have the second letter from ABCDEFGH are shown in the candidate words section. The letter selection continues until the desired word appears in the word selection region. By selecting the SPACE tile, participants can enter the word selection mode. The word selection region is a set of tiles, each representing one potential word that can be spelt with the selected letters, ordered by their usage frequency in English. Similar to the letter selection region, these tiles are sequentially highlighted ("scan") after a specific time, and participants can select a word by pressing the spacebar when the tile bearing the desired word is highlighted. Upon selection, the word is added to the transcribed text followed by a SPACE character and scanning reverts to the letter selection region for the input of the next word. The task is completed when the participant has fully entered the given sentence. The scanning interval is one of the parameters (MacKenzie and Felzer, 2010, 2022) used to manipulate SAK, and their analysis showed a statistically significant increase in error rates for scanning intervals equal to and below 800 ms. Following these authors, we developed two versions of this task with distinct difficulty levels: the scanning interval (i.e. how long each tile is highlighted before moving on to the next one) in the easy version was 1000 ms, and 500 ms in the hard version. In our pilot experiments, we ensured participants experienced different difficulty levels in these versions. We tested the significance of these manipulations by analysing the Scan per Character (SPC) values. We found statistically significant differences in SPC in these two conditions. The detailed analysis is presented in Section 5.2.2.

Dual N-back: To assess the reliability of the questionnaires (RQ3), we asked participants to complete a Dual N-back task (Jaeggi et al., 2003; Blacker et al., 2017) in two levels of difficulty before the SAK task. The idea was to understand whether the experience of different difficulty levels in a *different* task would leak into the subjective ratings of the *main* task. This simulates the effect of extraneous factors outside the task of interest on the ratings of this task. Ideally, an instrument that measures the MWL of a task should only capture reactions to that and only to that task—not to any other task around it. In our case, ideally, the dual N-back task should not have any effect on the MWL ratings of the SAK tasks.

Dual N-back is a dual-task paradigm in which two independent sequences of stimuli are presented simultaneously, one auditory and one visual. Participants must act when the current stimulus matches the one from N steps earlier in the sequence. For this task, we used Perry-Houts (2022)'s implementation. We chose N = 1 as the easy and N = 3 as the difficult level, informed by our pilot tests. In one round, participants did the 1-back (easy version) followed by SAK, while in the other, they did the 3-back (hard version) followed by SAK. The order of these rounds was randomised. The SAK task in both rounds had the same difficulty level for each participant, so we expected participants to rate questionnaires similarly in the two conditions despite the difference in difficulty on the dual N-back tasks.

In summary, we designed the study so that there should be a statistically significant difference in performance between the hard and easy SAK tasks (between participants), which should lead to a statistically significant difference in subjective ratings of MWL between them. However, if these ratings are specific to the SAK tasks and robust to the effects of other tasks (i.e. the N-back tasks), we should not observe a statistically significant difference in the subjective MWL ratings within each participant, because there is no difference in difficulty in the SAK task for the same participant.

5.1.2. Questionnaires

We chose MRQ and NASA-TLX as MWL measures for this experiment. The MRQ is a 17-item questionnaire based on Multiple Resource Theory (Wickens, 2008; Wickens et al., 1984) developed by Boles and Adair (2001). Each of its subscales assesses the demand on an attentional resource using a 5-level scale ranging from 0 to 4 (0: *"no usage"*, 1: *"light usage"*, 2: *"moderate usage"*, 3: *"heavy usage"*, 4:*"extreme usage"*). Table 2 illustrates the subscales of the MRQ. For this experiment, we developed a publicly available web version of this questionnaire.[2] We chose the R-TLX version of NASA-TLX since the weighting process of NASA-TLX was previously shown to be redundant (Nygren, 1991), as we discussed in Section 2.7.

5.1.3. Participants

We recruited our participants using the Prolific platform.[3] 188 participants started the experiment on the prolific platform and 100 of them managed to complete the experiment. We used these 100 completed trials as our sample, which included participants (50 men/50

[2] https://github.com/ebi-b/Experiment_Implementation
[3] https://www.prolific.co/

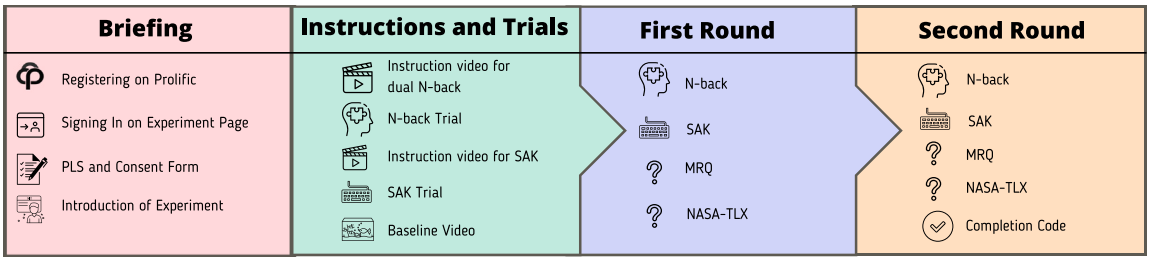


Fig. 3. Experiment Procedure.

women) aged between 18 and 62 ($M = 32.5$, $SD = 10.8$) whose first language was English and who used a laptop or desktop computer to complete the task. Participants completed the experiment on a website developed for that purpose and were reimbursed A$10 upon completion. The experiment took 30 to 40 min for each participant.

5.1.4. Procedure

The experiment consisted of four stages, as illustrated in Fig. 3. Participants could decide to participate in this experiment on the Prolific website, from where they were redirected to the experiment's web page.

Stage 1 - Introduction: In the first stage, participants logged in using their Prolific ID. We used this ID to reimburse them upon completion. Afterwards, participants proceeded through the plain language statement, consent form, introduction to the experiment, and information on required time and devices.

Stage 2 - Tutorials: Participants then received instructions required to complete the tasks through two video tutorials and completed two practice trials, one for the dual N-back task and one for SAK. These instructions and trials ensured that participants were familiar with the tasks and could complete them. Participants could revisit the task instructions and repeat the trials at any point. Upon finishing the practice stage, participants watched a calming one-minute video of a fish tank to ensure that all started from a similar MWL baseline.

Stages 3 and 4 - Experimental tasks: The third and fourth stages of this experiment were two rounds of the dual N-back task, followed by SAK, MRQ and NASA-TLX. Fig. 4 illustrates the steps in Stages 3 and 4 and the RQs related to each part. Participants completed the easy version of dual N-back in one round and the hard version in the other. The order was balanced across the sample. The difficulty level of SAK in both rounds was the same for each participant but different across participants—participants were divided into two groups, one group doing the easy SAK and the other the hard SAK. The order of the MWL questionnaires in each stage, the N-back versions and the SAK difficulty level were randomly selected. To ensure the credibility of the participants' data in this experiment, we embedded two attention checks, ensuring participants remained focused for the whole experiment and did not enter random responses to the questionnaires. Failure in the attention checks terminated the experiment. Prolific flagged participants who failed the attention check, and we excluded those participants from the analysis. For our analysis, we selected 100 participants who did not fail any attention checks. We specified in the Stage 2 instructions and inside each questionnaire in Stages 3 and 4 that the MWL questionnaires had to be filled in only based on the demands of the SAK task and *not* based on the N-back task. Finally, upon completion of the experiment, participants received a code that they could enter in Prolific to receive their reimbursement.

5.2. Results

In this section, we present our results for each RQ. We collected responses to the MRQ and NASA-TLX questionnaires and the timestamped actions during the experiment from 100 participants. We discarded the data of two participants as it was found faulty due to an error in saving their data in the database, leaving us with N = 98 participants. We ensured that each participant only completed the tasks once. We calculated R-TLX as the overall NASA-TLX score.

5.2.1. RQ1 - convergent validity

Following our discussion in Section 2.6, we considered subjective MWL rates as ordinal values with a non-linear relationship with MWL, and to address **RQ1**, we computed Spearman's Rank Correlation coefficient between the overall MWL measured by each questionnaire. Spearman's Rank Correlation does not consider any linear relationship assumption for the measurements taken and is appropriate for ordinal scales. The results indicate a small statistically significant positive correlation between the overall R-TLX and MRQ ratings (Spearman' $\rho(194) = 0.2312, p = 0.0012$).

We further examined this relationship through a factor analysis of the subscales of the questionnaires. The Kaiser–Meyer–Olkin measure of sampling adequacy was 0.77, above the commonly recommended value of 0.6, and Bartlett's test of sphericity was statistically significant ($\chi 2(253) = 1489, p < 0.001$). To select the number of factors, we plotted the factors' eigenvalues scree plot (see Fig. 5(b)). The scree plot curve shows an elbow on the fifth factor, so we selected the first 5 factors for our factor analysis. We fitted the model with *MINRES* and orthogonal rotation (*varimax*). Fig. 5(a) shows the factor loadings of the subscales of the two questionnaires. For identifying statistically significant sample loadings, we set the threshold to 0.4 based on Hair et al. (1998)'s suggestion for a sample size of $\sim$ 200. As the results indicate, NASA-TLX subscales only influence the 3rd factor, while MRQ subscales influence all factors except the 3rd one.

In summary, the convergent validity between NASA-TLX and MRQ is questionable. While there is a statistically significant correlation between the two measures, its magnitude is small, and their subscales load onto different factors, suggesting they may assess distinct aspects of MWL.

5.2.2. RQ2 - sensitivity

If MWL captures information about the task's difficulty, a more difficult task should lead to a higher mental workload and, hence, higher scores on the scales. In our study, two groups of participants completed the SAK task with two difficulty levels. To validate the difference in difficulty, we computed each condition's average SPC (Scan per Character) score. SPC counts the average number of scan steps necessary to enter a text character using a given scanning keyboard in a given language (MacKenzie and Felzer, 2010). The higher the SPC, the more difficult the task. The average SPC in the easy and hard versions were 5.64 and 11.81, respectively. We tested whether this difference was statistically significant with a general linear mixed effects model, testing for the effect of difficulty on SPC, including random effects of participants. The test showed a significant effect of SAK difficulty ($\beta = 6.03 \pm 0.7$, p< 0.0001). This validates that our hard task was indeed more difficult than the easy task.

This difference, however, was not reflected in the subjective ratings of mental workload. A Wilcoxon rank sum test with continuity correction failed to find a statistically significant effect of the task difficulty on either the R-TLX or MRQ scores. The results of the test for R-TLX

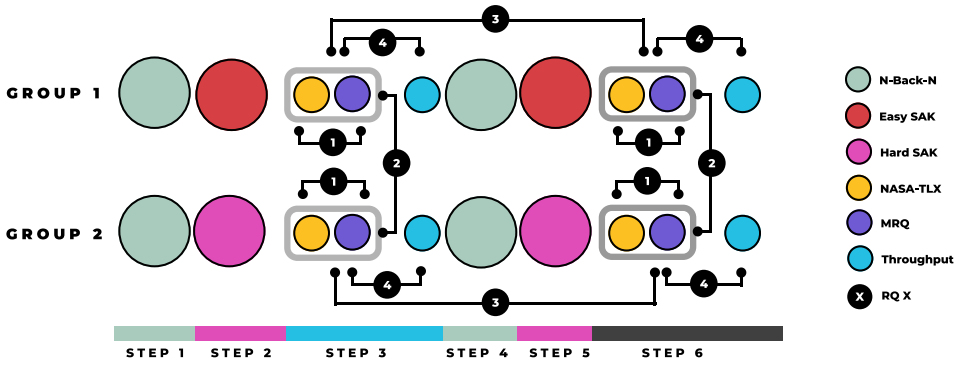
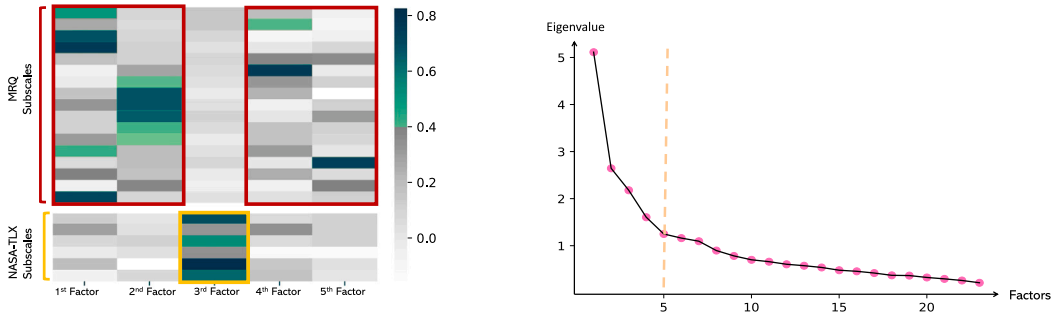


Fig. 4. Stage 3 and 4 of the experiment. The black circles illustrate the comparisons we make to answer each research question.



(a) Factor Analysis Results (b)

Fig. 5. (a) Factor Analysis Results (b) Scree Plot.

scores were $W = 5054, p = 0.19$ (hard: $mean = 48, sd = 21$ and easy: $mean = 52, sd = 14$) and for the MRQ scores were $W = 4250, p = 0.17$ (hard: $mean = 1.56, std = 0.62$ and easy: $mean = 1.42, sd = 0.54$).

In summary, neither questionnaire was sensitive enough to capture the differences in MWL caused by task difficulty. In fact, the MWL scores in the easier task were slightly higher than in the harder task (though this difference was not significant).

5.2.3. RQ3 - consistency and stability

A reliable scale should be internally consistent, i.e. its items should be closely related as a group. To measure such consistency in **RQ3**, we computed Cronbach's alpha for each questionnaire separately. Both questionnaires demonstrated good internal consistency: MRQ (17 items, $\alpha = 0.81$) and NASA-TLX (6 items, $\alpha = 0.74$).

To evaluate stability, we examined whether mental workload (MWL) scores remained consistent across repeated trials. Because participants completed the SAK task twice at the same difficulty level, their MWL scores should be similar despite prior exposure to N-back tasks of varying difficulties. In other words, the N-back task difficulty should not affect the mental workload scores of either questionnaire. If this is the case, it is evidence that the questionnaires can capture MWL information specific to the task and not about unrelated previous tasks. To measure stability in **RQ3**, we used a Wilcoxon Signed-Rank Test between participants' responses to both questionnaires in two different conditions (1-back vs. 3-back). The mean and median of the MWL scores were $mean = 49.9, sd = 13.4$ (R-TLX) and $mean = 1.50, sd = 0.96$ (MRQ) in the 1-back condition and $mean = 50.6, sd = 14.$ (R-TLX) and $mean = 1.48, sd = 0.97$ (MRQ) in the 3-back condition. Both scales passed this test, and we found no statistically significant effect of N-back difficulty in the overall MWL score of either the TLX ($V = 1832.5, p = 0.49$) or MRQ ($V = 2124.5, p = 0.62$).

In summary, both the NASA-TLX and the MRQ showed acceptable levels of internal consistency and stability.

5.2.4. RQ4 - performance

We operationalised the SAK text entry performance as throughput. MacKenzie and Felzer (2010) define throughput as $T = (\frac{1}{SPC}) \times (\frac{1000}{SI}) \times (\frac{60}{5})$, where SPC is Scan per Character and SI is Scan Interval. Given that a major motivation for measuring MWL in previous works is to predict task performance, text entry throughput should be strongly negatively correlated with subjective MWL scores—lower MWL should lead to higher performance. So, to address **RQ4**, we computed Spearman's correlation coefficient between the overall MWL of each questionnaire and the throughput T of the corresponding task. The NASA-TLX ratings showed a small negative correlation with performance ($\rho(194) = -0.23, p < 0.01$). The MRQ ratings also showed a statistically significant correlation with performance, but this relationship was not as strong ($\rho(194) = -0.16, p = 0.028$).

These results suggest that the NASA-TLX was better than the MRQ at capturing differences in performance, but neither questionnaire showed a strong correlation with performance.

5.2.5. RQ5 - combining scales

Finally, to explore the best combination of subscales to measure MWL, we calculated the multicollinearity of all subscales of the two questionnaires using the variance inflation factor (VIF). VIF measures how much the behaviour (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with other independent variables and hints at the best subset of subscales. The ideal VIF score is equal to or less than 3.3 (Kock, 2015; Hair et al., 2019; Kock et al., 2012). The average VIF score for NASA-TLX subscales was 8.36 and 4.57 for MRQ, describing critical and moderate levels of multicollinearity, respectively.

We also examined the multicollinearity of subscales while combining the two questionnaires. To do so, first, we calculated VIF incorporating all the subscales of MRQ and NASA-TLX. Only seven subscales passed the threshold, all of which belong to MRQ. Then, we examined the effect of adding the rest of the subscales to these seven subscales,

Table 3
Multicollinearity of different subscales of MRQ and NASA-TLX.

Rank	Subscale	Questionnaire	VIF	Secondary VIF
1	Auditory emotional process	MRQ	1.97	1.85
2	Tactile figural process	MRQ	1.98	1.81
3	Vocal process	MRQ	2.06	1.90
4	Facial figural process	MRQ	2.17	1.94
5	Spatial quantitative process	MRQ	2.49	2.30
6	Auditory linguistic process	MRQ	2.68	2.12
7	Facial motive process	MRQ	2.87	2.62
8	Visual phonetic process	MRQ	3.38	2.57
9	Performance	NASA-TLX	3.54	2.03
10	Spatial concentrative process	MRQ	4.29	2.71
11	Physical Demand	NASA-TLX	4.38	2.58
12	Spatial categorical process	MRQ	5.07	>3.3
13	Spatial emergent process	MRQ	5.31	>3.3
14	Spatial positional process	MRQ	5.62	>3.3
15	Frustration	NASA-TLX	5.98	>3.3
16	Manual process	MRQ	7.01	>3.3
17	Short-term memory process	MRQ	7.83	>3.3
18	Visual temporal process	MRQ	8.25	>3.3
19	Temporal Demand	NASA-TLX	8.98	>3.3
20	Spatial attentive process	MRQ	9.49	>3.3
21	Visual lexical process	MRQ	12.35	>3.3
22	Mental Demand	NASA-TLX	17.96	>3.3
23	Effort	NASA-TLX	19.88	>3.3

one by one. Four subscales scored $VIF \leqslant 3.3$, two of which belong to NASA-TLX. Table 3 presents the results of the first and second rounds of multicollinearity tests and selected subscales.

These results suggest that the subscales from both questionnaires were highly interrelated, but the MRQ scales are less redundant than the NASA-TLX. In addition, if combining the scales, the NASA-TLX may be more useful in a reduced format.

5.3. Discussion

In this section, we discuss *convergent validity*, *sensitivity*, *reliability*, *the relationship between MWL and performance* and the *best combination of subscales* for measuring MWL in SAK.

5.3.1. RQ1: Convergence of scales

Convergent validity is demonstrated when a test strongly correlates with other variables or tests with which it shares the overlap of a construct. To consider two scales to measure the same construct, their *"correlations should be strong and positive"* and any test that does not correlate at least moderately would be highly suspect of not having convergent validity (Gregory, 2014). For both MRQ and NASA-TLX to be valid measures of MWL, to pass this test, they must exhibit a high correlation in their scores; however, as we discuss in Section 2.3, previous works have found that MWL scales exhibited divergence in several instances. In this experiment, we found only a small correlation between these scales, which is insufficient evidence of convergent validity.

We further investigated this by looking at the underlying factors affecting the rates in each questionnaire. As illustrated in Fig. 5(a), the underlying factor behind the NASA-TLX subscales seems to differ from those behind MRQ. This means the factors contributing to MRQ's subscales differ from NASA-TLX's, suggesting that they measure different constructs.

These findings caution against considering MWL as a hypothetical construct. Instead, MWL seems to be an intervening variable describing what different scales measure. This makes comparing scores from different scales difficult, as they probably refer to different constructs. Another possible explanation for the lack of convergence between these two scales relates to our previous discussion of the multivariate nature of MWL. These scales may measure different dimensions of MWL, so although both capture facets of MWL, these facets are different. It warns researchers to be cautious when generalising findings based on one MWL scale or adopting findings from studies using distinct MWL scales. This also strengthens the need for better conceptualising and defining MWL for each given study.

5.3.2. RQ2: Sensitivity to difficulty alternations

Sensitivity denotes the capacity to detect statistically significant differences in the magnitudes of questionnaire scores for different MWL levels (Lewis, 2002). As previously mentioned, we developed two versions of the SAK varying in difficulty, each completed by one group of participants in our experiment. We expected a statistically significant difference between the MWL scores of these questionnaires induced by the differences in difficulty. However, our results reveal no statistically significant difference between captured MWL scores for these two versions. This implies that the scales may not be sensitive enough to capture changes in MWL caused by changes in the difficulty of SAK, a standard HCI task, in our experimental setting. Though this could be attributed to a lack of statistical power, it still creates a challenge for HCI research given that our sample ($N = 98$) was much larger than most quantitative studies published at CHI ($N = 20 \pm 12$) (Caine, 2016), many of which involve the use of the NASA-TLX.

The lack of sensitivity of these MWL scales may also be attributed to the difference in task difficulty in our experimental design. However, we find this unlikely as variations of SI have shown to impact error rate (MacKenzie and Felzer, 2010). Further, various controls (described in Section 5.1.1) we adopted confirmed changes in performance in the two versions of SAK. In between-group experimental designs like ours, where each participant was only exposed to one difficulty level, the lack of a reference task means these scales are prone to insensitivity. In addition, in more complex tasks where participants cannot identify the source of MWL or in cases where the source causing changes in MWL is not a subscale of these questionnaires — both of which are common situations in HCI tasks — these questionnaires are prone to insensitivity.

Another potential explanation is that because the task difficulty manipulation only tapped into one resource — visual perception — it is possible that this was not sufficient to yield a substantial difference in MWL. However, given that our task is representative of the kinds of HCI tasks in which MWL questionnaires are applied, if large changes

in difficulty do not yield differences in MWL, perhaps MWL is not a relevant construct to measure in this kind of experiment.

Though a single study is not enough to invalidate a scale, it emphasises the need for a collective effort in validating the scales we adopt from other disciplines in HCI tasks. Until then, HCI researchers should be cautious that not observing a statistically significant difference in questionnaire ratings between conditions does not necessarily mean that there is no difference in MWL. It may be due to the lack of sensitivity of the scales in capturing changes in MWL.

5.3.3. RQ3: Reliability in biased conditions

Internal consistency is a measure of how well all the items in a scale measure the same construct (Tavakol and Dennick, 2011). It is typically measured using Cronbach's coefficient alpha and is considered adequate if alpha is equal to or greater than 0.7 (Ley et al., 2009). In terms of internal consistency, both questionnaires showed promising results ($\alpha > 0.7$ (Gliem and Gliem, 2003)). This is evidence that the scales might still be reliable enough for application in other domains, even if they are inappropriate for HCI. We found no evidence for issues in test-retest reliability, finding no effect of the N-back task on MWL scores in either scale. This could be because the scales are indeed stable, but it could also be because of the lack of sensitivity to the task's difficulty, as evidenced by the results pertaining to RQ2. It is also possible that the N-back tasks did not lead to differences in MWL. We did not collect data to confirm the manipulation, but we find it unlikely that this is the case, as previous work has shown through physiological tests that N-back tasks do induce higher MWL (Arana-De las Casas et al., 2023).

5.3.4. RQ4: Performance

We found statistically significant small negative correlations between NASA-TLX overall MWL and performance in our data, representing the decrease in words per minute as MWL increases. However, this correlation is too small to make meaningful predictions about performance based on NASA-TLX ratings. An even smaller correlation was found for the MRQ.

These results can be interpreted in three ways. First, we can conclude from these results that MWL is not a good predictor of performance. This reiterates our discussion of the need to consider more complex models for the relationship between MWL and performance, as the traditional uni-dimensional view cannot present this relationship entirely, and there is no guarantee that reducing MWL improves performance (Young et al., 2015). A second interpretation of these results suggests that the low correlation between performance and MWL may be due to the inability of subjective MWL scales to accurately measure MWL, consequently resulting in a low correlation between MWL subjective scores and performance. A third interpretation is that the scales did measure MWL accurately, but as we discussed in Sections 2.5 and 2.7, MWL and performance have a dynamic relationship, with users regulating MWL to maintain a given level of performance. Subjective scales cannot measure this two-way relationship, and we may fail to observe a correlation between subjective MWL scores and performance, even if there is a causal effect between MWL and performance. This also suggests that more objective MWL measurements might be more appropriate for predicting performance, such as physiological signals. However, to confirm which of these views holds, more research is required on the relationship between MWL and performance.

In any case, if the goal of administering a NASA-TLX questionnaire as part of the evaluation of an interactive system is to predict user performance — as is often the case in HCI — it is likely that the research will not yield meaningful results. As such, we do not recommend using either the NASA-TLX or the MRQ for this purpose.

5.3.5. RQ5: Combining questionnaires

Multicollinearity tests demonstrate that the subscales of NASA-TLX are highly correlated ($VIF = 8.36$), which means changes in one subscale are associated with shifts in another one. This implies we can obtain the same results using only a subset of these subscales. MRQ showed better results ($VIF = 4.57$); which is evidence that MRQ is better designed with fewer correlated subscales. Still, some of its subscales are redundant due to the high correlation (see Table 3). This could be because we only used MRQ in a single task, lacking enough variation to show differences in other subscales. Our factor analysis shows that merging these two questionnaires could be valuable as different factors influence them. The results of the multicollinearity tests corroborated this as well. *Performance* and *physical demand* were two subscales of NASA-TLX, which could be used alongside nine subscales of MRQ. This is in line with the findings of Finomore et al. (2013), who suggested NASA-TLX and MRQ can complement each other.

5.4. Summary

We presented a validation study on NASA-TLX and a more modern alternative, MRQ, assessing their convergent validity, sensitivity, and reliability on a standard HCI task. Validating a scale in a field of research requires a collective effort and running multiple validation studies on various applications of that area of research. This validation experiment aims to encourage further refinement of the concept of MWL and the validation of scales in various contexts of HCI. Our results suggest that NASA-TLX and MRQ are stable and consistent on the SAK task, while their convergent validity and sensitivity are questionable. Further, researchers must be cautious in making predictions about performance based on subjective MWL ratings, as the correlations we found were too small to be meaningful. We also reiterate de Waard and Lewis-Evans (2014)'s suggestion of using various scales to capture different aspects of MWL, as the underlying factors behind MRQ and NASA-TLX seem to be different. Combining these two scales could paint a more comprehensive picture of users' MWL.

6. Limitations

We selected 75 papers from the ACM CHI Proceedings as a sample of HCI research involving MWL. However, HCI covers a very broad range of proceedings and journals. Therefore, one limitation of this review is the fact that our selection of papers may not be representative of all types of MWL research in HCI. In addition, we found our sample to be prone to *Jingle-Jangle Fallacy* issues in which researchers use constructs with similar names interchangeably while referring to different constructs (Aeschbach et al., 2021). We believe this may also exist in the context of MWL, and terms such as "cognitive load" and "cognitive workload" may be used to describe MWL. In our review, we excluded these terms from our search as the claims of this paper are limited to MWL constructs. Additionally, it was not possible to detect when researchers actually intended to measure MWL instead of cognitive load when they used these terms. In their review, Kosch et al. (2023) used a broader search string that captured these terms on a broader subset of the literature beyond CHI and identified further instruments. Some of them were appropriate MWL measures, such as the Instantaneous Self-Assessment (a one-question measure designed to be administered multiple times during a task) (Tattersall and Foord, 1996) and the Bedford Workload Scale (a unidimensional scale answered in a two-step process) (Roscoe and Ellis, 1990). However, they also identified the use of a number of inappropriate scales designed to measure other constructs, such as the Dundee Stress State Questionnaire (stress responses) (Matthews et al., 1999), System Usability Scale (usability) (Brooke, 1996), Rating Scale Mental Effort (Zijlstra, 1993) (perceived effort). This reinforces that the issues we identified in our sample may also be prevalent in the broader HCI literature.

Another limitation of the scope of our review is that we focused on subjective MWL ratings. Other approaches that have been explored in the literature include biosignals (cerebral, ocular, cardiovascular, dermal), task performance measures, input device signals, speech, body movements and saliva (Kosch et al., 2023). We make no claims about

the validity or appropriateness of these approaches and refer the reader to Kosch et al. (2023) for a discussion.

We reiterate that validating an instrument or construct is a collective scientific effort, and a single experiment cannot validate or invalidate an instrument. Our experiment invites the community to validate and refine MWL instruments and concepts within an HCI context. Different HCI tasks and setups may result in different results, and the outcomes of this experiment should not be used to invalidate previous studies as long as there are not enough consequent studies that confirm these results. Our validation experiment focused on one canonical HCI task — text entry — instantiated as a specific interaction technique—SAK. For example, the task we chose focused mostly on visual demands without involving other resources, such as spatial or auditory demands. In real-world settings, MWL is a multidimensional concept spanning cognitive, perceptual, and motor resources, so a more ecologically valid evaluation could yield different findings. Though we believe that a standard tool in the HCI belt should apply to any standard HCI task, it could be that idiosyncrasies of the technique or general properties of text entry affect the validity of these instruments. In addition, this validation experiment was conducted online, and its outcomes may be limited to online studies. Lab and in-the-wild studies may have different outcomes.

We operationalised MWL with two instruments. Our findings are limited to the validity of MRQ and NASA-TLX, and they should not be used to invalidate previous research findings using other scales on other types of HCI tasks. However, our findings demonstrate the urgency for additional validation studies of other MWL scales on text entry and other HCI tasks. For example, an MWL scale validated in a GUI task might fail to measure MWL in voice-based interfaces, as MWL instruments might behave differently across modalities.

Typically, the psychometric evaluation of questionnaires includes several tests, all of which cannot be addressed in a single study. In our validation experiment, we investigated convergent validity, sensitivity, stability, and internal consistency; however, other psychometric tests, such as inter-rater reliability, face validity and content validity, should also be addressed in future studies.

7. Implications for HCI

Though our review and experiment highlighted several critical issues in the conceptualisation, measurement, and application of MWL in HCI, there are several directions through which HCI researchers can mitigate these issues and exploit MWL potentials in guiding designs.

7.1. Consensus definition

Building our work upon solid theoretical grounds is critical to ensuring the reproducibility of MWL practices in HCI. We still cannot conclude whether MWL is an intervening variable or a hypothetical construct, as evidence for both positions exists. On the one hand, if we consider it a unitary construct to assess users' perceived MWL, as in the NASA-TLX, we lean towards seeing it as an intervening variable that generates an abstract of users' perceived workload. On the other hand, when considering it as a multivariate construct as described in multiple resource theory, we lean more towards seeing it as a hypothetical construct as it displays other visible cues, such as performance deterioration. However, considering MWL as a multivariate construct adds complexity to its analysis since not only are the demands for each resource important, but the interference of demands is also of interest. For simplistic tasks and settings, NASA-TLX can be an insightful tool as the overlap between resources is not significant; however, these are not typical in realistic interactions with digital systems. To us, a multivariate definition of MWL seems more appropriate for HCI, as real-world tasks of interest to our community are complex, demanding various attentional and cognitive resources. In addition, the multivariate view is supported well by Multiple Resource Theory and gives HCI researchers a better theoretical basis for interpreting results and informing design decisions.

7.2. New scales and validation studies

The reviews and validation experiment reveal a potential for developing new subjective scales for measuring MWL that are more applicable in HCI. Longo et al. (2022), in a recent review of literature, identified 22 subjective scales that have been used for measuring MWL, most of which were outdated, did not reflect recent MWL research, or had been developed for measuring constructs other than MWL. NASA-TLX, as the de facto MWL scale used in CHI papers, did not show enough sensitivity in a standard HCI task and its subscales seem too far removed from our application domain to inform HCI applications. Besides *physical demand*, the other five subscales of NASA-TLX are not instructive enough to inform HCI design requirements. For instance, if NASA-TLX identifies *frustration level* as the main source for high MWL in a given design alternative, what does that mean for a re-design?

MRQ minimises this issue by using concrete subscales (e.g. *Spatial attentive process*), which explicitly specify what features in the design must be modified (e.g. task demands that require a focused sense of vision). However, to date, the number of validation studies on MRQ is still limited, and we have failed to confirm its validity for the task we studied. Like NASA-TLX, it did not present promising results in terms of sensitivity to differences in difficulty level on a standard HCI task in our experiment, which emphasises the need for further validation studies of MWL scales on different HCI applications. Moreover, although MRQ is inspired by Multiple Resource Theory, due to a large number of attentional resources, Boles et al. (2007) selected only 14 resources through multiple factor analysis studies, which were specifically designed for dual-task performance. Later, they added three more resources to the list and released MRQ as a 17-subscale questionnaire. Boles et al. (2007) explicitly state that this set of resources should not be considered complete. Therefore, MRQ may also suffer from an arbitrary choice of subscales, similar to the NASA-TLX. In addition, using the MRQ in a given study implies that it builds upon Multiple Resource Theory. Multiple Resource Theory explicitly explains what resources may intervene with each other and what resources can be used in parallel; however, the complete mapping between multiple resource theory frameworks of resources and MRQ subscales is challenging. In sum, our community must rethink *why* we measure MWL and what *actionable insights* we seek from these measurements. Only then can we decide what data to collect and design instruments for that purpose.

7.3. Alternative constructs

MWL has rich potential for use by HCI researchers to quantify and, thus, manage cognitive demands. However, the conceptualisation of this construct is still evolving, even in the literature from which HCI research appropriated it. In a considerable body of HCI work, MWL is still used to evaluate other constructs, such as usability or performance. In this kind of work, measuring MWL itself is not the goal. As such, rather than just haphazardly administering a NASA-TLX questionnaire at the end of a study, it is worth considering what construct the research team is really interested in and which instrument to use to measure it. This may lead to having a more solid theoretical background, specifically in cases where relying on the instrument's validity is critical. Although MWL still has the potential to be insightful for exploratory research and iterative design in HCI, researchers must practice caution measuring MWL with subjective questionnaires in confirmatory research as our findings highlight disparities in its concept and deficiencies in its scales.

7.4. Advice for measuring MWL in HCI

Based on our theoretical, methodological, and empirical analysis of MWL in HCI, we conclude with some advice for future work.

Table 4
Minimal reporting standards for mental workload measurement.

Category	Reporting requirements
Conceptual and Theoretical Justification	- The definition of MWL being used. - The theoretical framework supporting this definition. - The theoretical justification for measuring MWL in the context of the study.
Instrument Details	- The instrument used (e.g., NASA-TLX, MRQ, or see Kosch et al. (2023) for alternatives). - Justification for this choice of instrument. - Whether the instrument was modified. If so, provide details and validation status. - Whether this version of the instrument has been validated or whether additional validation was conducted. - Whether responses will be analysed at the aggregate or subscale level and the implications for the theory being tested (this decision should be made a priori).
Questionnaire Administration	- How the questionnaire was administered (e.g., paper, web, mobile, VR). - The scale format (e.g., 5-pt Likert, 7-pt Likert, Visual Analogue Scale (VAS)). - If using NASA-TLX, specify whether the weighting step was conducted.
Materials and Data	- The instrument materials (if using a custom or modified questionnaire). - Files with raw, anonymised response data and analysis code, preferably accessible through an open repository (e.g., OSF, GitHub) or in the supplementary materials associated with the paper.

7.4.1. Define what you mean by mental workload:

Our review revealed a wide variety of definitions in the literature, with different implications for its measurement. Before administering a questionnaire, it is critical to clarify which definition is being built upon and with which theoretical framework (e.g. Multiple Resource Theory, Cognitive Load Theory, etc.) the work is aligned. We suggest Longo et al. (2022) as a useful catalogue of definitions.

7.4.2. Reflect about why you are measuring mental workload:

MWL involves a multimodal, dynamic, feedback-driven process—it is a complex construct to measure. In many of the papers in our sample, it was not a key construct in the research question. As such, we encourage researchers to reflect on whether it is necessary to measure it in the first place.

7.4.3. Reconsider the administration of the NASA-TLX in HCI:

We invite researchers to reconsider the NASA-TLX as a default choice of questionnaire to administer in our experiments. It was not originally designed to be administered in the contexts in which we do so. Further, its structure is not designed to yield actionable re-design insights. Alternative instruments might be more informative in this sense, but a combination of instruments might work even better. We refer the reader to Kosch et al. (2023) for a list of possibilities.

7.4.4. Do not assume that ratings are linearly related:

Though many analyses in the literature employ tests that involve linear assumptions (t-tests, ANOVAs, etc.) to analyse MWL ratings, these assumptions are rarely met. For statistical significance tests, a better approach for analysing this data is the use of non-parametric tests (e.g. Spearman's correlation, Wilcoxon signed-rank tests). If taking an estimation-driven approach, a suitable method is to use a generalised linear model with a cumulative probit link, which allows you to model participant- and item-level random effects. These models allow you to estimate effects in the scale of the latent normally distributed MWL distribution as opposed to the ratings themselves. For guidance on this procedure, we recommend Liddell and Kruschke (2018).

7.4.5. Be transparent in the reporting of your measurements:

A common theme in our review was the lack of detail about how MWL was measured. Many studies provided insufficient information about the theoretical foundation of MWL, the choice and implementation of measurement instruments, or the data processing methods used. Given the conceptual ambiguities surrounding MWL and the methodological limitations identified in previous work, transparent reporting is essential for improving reproducibility and ensuring meaningful comparisons across studies. To address these gaps, we present a checklist (Table 4) outlining key details that should be reported when measuring MWL. This checklist highlights critical aspects of MWL measurement, including theoretical justifications, instrument selection and modifications, administration procedures, and data availability—elements that are frequently omitted in prior studies.

8. Conclusion

In this paper, we presented a comprehensive review of debates around the MWL concept, a methodological review of CHI papers that involve the use of MWL, and an experiment investigating the validity, reliability, and sensitivity of subjective MWL scales on a standard HCI task. We found that MWL is still an amorphous and poorly understood construct that, nevertheless, is widely used in CHI papers. Importantly, our findings suggest that commonly used scales that operationalise MWL lack sensitivity and convergent validity on a standard HCI task. Consequently, we identify an urgent need for improved practices and increased awareness of methodological standards. In particular, we advise caution when using NASA-TLX for user experience evaluations, particularly when predicting user performance or seeking actionable insights.

CRediT authorship contribution statement

Ebrahim Babaei: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tilman Dingler:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Benjamin Tag:** Writing – review & editing, Validation. **Eduardo Velloso:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Aeschbach, Lena Fanya, Perrig, Sebastian A.C., Weder, Lorena, Opwis, Klaus, Brühlmann, Florian, 2021. Transparency in measurement reporting: A systematic literature review of CHI PLAY. Proc. ACM Hum.-Comput. Interact. 5, http://dx.doi.org/10.1145/3474660, CHI PLAY.

Afridi, Ahmad Hassan, Mengash, Hanan Abdullah, 2020. NASA-TLX–based workload assessment for academic resource recommender system. Pers. Ubiquitous Comput. 26 (3), 881–899. http://dx.doi.org/10.1007/s00779-020-01409-z.

Albers, Michael J., 2011. Tapping as a measure of cognitive load and website usability. SIGDOC '11, Association for Computing Machinery, New York, NY, USA, pp. 25–32. http://dx.doi.org/10.1145/2038476.2038481.

Anon, 2021a. https://humansystems.arc.nasa.gov/groups/tlx/downloads/TLX_pappen_manual.pdf. (Accessed 30 August 2021).

Anon, 2021b. Subjective workload assessment technique (SWAT): A user's guide. https://apps.dtic.mil/sti/citations/ADA215405. (Accessed 30 August 2021).

Anon, 2021c. TLX @ NASA Ames - NASA TLX paper/pencil version. https://humansystems.arc.nasa.gov/groups/tlx/tlxpaperpencil.php. (Accessed 06 September 2021).

Anthony Deutsch, J., Deutsch, Diana, 1963. Attention: Some theoretical considerations. Psychol Rev 70 (1), 80–90. http://dx.doi.org/10.1037/h0039515.

Arana-De las Casas, Nancy Ivette, De la Riva-Rodríguez, Jorge, Maldonado-Macías, Aide Aracely, Sáenz-Zamarrón, David, 2023. Cognitive analyses for interface design using dual N-back tasks for Mental Workload (MWL) evaluation. Int. J. Environ. Res. Public Heal. 20, 1184. http://dx.doi.org/10.3390/IJERPH20021184.

Babaei, Ebrahim, Tag, Benjamin, Dingler, Tilman, Velloso, Eduardo, 2021. A critique of electrodermal activity practices at CHI. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3411764.3445370.

Beevis, David, 1999. Analysis techniques for human-machine systems design: A report produced under the auspices of NATO defence research group panel 8. Crew Systems Ergonomics/Human Systems Technology Information Analysis Center.

Blacker, Kara J., Negoita, Serban, Ewen, Joshua B., Courtney, Susan M., 2017. N-back versus complex span working memory training. J. Cogn. Enhanc. 1 (4), 434–454. http://dx.doi.org/10.1007/s41465-017-0044-1.

Boles, David B., Adair, Lindsey P., 2001. The Multiple Resources Questionnaire (MRQ). Proc. Hum. Factors Ergon. Soc. Annu. Meet. 45 (25), 1790–1794. http://dx.doi.org/10.1177/154193120104502507.

Boles, David B., Bursk, Jonathan H., Phillips, Jeffrey B., Perdelwitz, Jason R., 2007. Predicting dual-task performance with the Multiple Resources Questionnaire (MRQ). Hum. Factors 49 (1), 32–45. http://dx.doi.org/10.1518/001872007779598073, PMID: 17315841.

Bowman, Robert, Nadal, Camille, Morrissey, Kellie, Thieme, Anja, Doherty, Gavin, 2023. Using thematic analysis in healthcare HCI at CHI: A scoping review. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3544548.3581203.

Braun, Michael, Mainz, Anja, Chadowitz, Ronee, Pfleging, Bastian, Alt, Florian, 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–11. http://dx.doi.org/10.1145/3290605.3300270.

Brehmer, Matthew, McGrenere, Joanna, Tang, Charlotte, Jacova, Claudia, 2012. Investigating interruptions in the context of computerised cognitive testing for older adults. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 2649–2658. http://dx.doi.org/10.1145/2207676.2208656.

Broadbent, Donald E., 1956. Listening between and during practiced auditory distractions. Br. J. Psychol. 47 (1), 51–60. http://dx.doi.org/10.1111/j.2044-8295.1956.tb00561.x.

Broadbent, Donald E., 1957. A mechanical model for human attention and immediate memory. Psychol Rev 64 (3), 205–215. http://dx.doi.org/10.1037/h0047313.

Brooke, John, 1996. SUS: A quick and dirty usability scale. In: Jordan, Patrick W., Thomas, Bruce, McClelland, Bernard, Weerdmeester, Ian L. (Eds.), Usability Evaluation in Industry. Taylor & Francis, pp. 189–194.

Brown, Scott W., 1997. Attentional resources in timing: Interference effects in concurrent temporal and nontemporal working memory tasks. Percept. Psychophys. 59 (7), 1118–1140. http://dx.doi.org/10.3758/bf03205526.

Bustamante, Ernesto A., Spain, Randall D., 2008. Measurement invariance of the Nasa TLX. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 52 (19), 1522–1526. http://dx.doi.org/10.1177/154193120805201946.

Butmee, Totsapon, Lansdown, Terry C., Walker, Guy H., 2019. Mental workload and performance measurements in driving task: A review literature. In: Bagnara, Sebastiano, Tartaglia, Riccardo, Albolino, Sara, Alexander, Thomas, Fujita, Yushi (Eds.), Proceedings of the 20th Congress of the International Ergonomics Association. IEA 2018, Springer International Publishing, Cham, pp. 286–294.

Byrne, Aidan, Tweed, Nathan, Halligan, Claire, 2014. A pilot study of the mental workload of objective structured clinical examination examiners. Med. Educ. 48 (3), 262–267. http://dx.doi.org/10.1111/medu.12387.

Cain, Brad, 2007. A Review of the Mental Workload Literature. Defence research and development Toronto (Canada).

Caine, Kelly, 2016. Local standards for sample size at CHI. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. CHI '16, Association for Computing Machinery, New York, NY, USA, pp. 981–992. http://dx.doi.org/10.1145/2858036.2858498.

Cockburn, Andy, Dragicevic, Pierre, Besançon, Lonni, Gutwin, Carl, 2020. Threats of a replication crisis in empirical computer science. Commun. ACM 63 (8), 70–79. http://dx.doi.org/10.1145/3360311.

Cockburn, Andy, Gutwin, Carl, Dix, Alan, 2018. HARK no more: On the preregistration of CHI experiments. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, Association for Computing Machinery, New York, NY, USA, pp. 1–12. http://dx.doi.org/10.1145/3173574.3173715.

Cooper, George E., Harper, Robert P., 1969. The use of pilot rating in the evaluation of aircraft handling qualities. In: NASA technical note, National Aeronautics and Space Administration.

Curry, Renwick, Jex, Henry, Levison, William, Stassen, Henk, 1979. Final report of control engineering group. In: Mental Workload. Springer, pp. 235–252.

De Waard, Dick, Brookhuis, K.A., 1996. The Measurement of Drivers' Mental Workload. Groningen University, Traffic Research Center Netherlands.

de Waard, Dick, Lewis-Evans, Ben, 2014. Self-report scales alone cannot capture mental workload. Cogn. Technol. Work. 16 (3), 303–305. http://dx.doi.org/10.1007/s10111-014-0277-z.

de Winter, Joost C.F., 2014. Controversy in human factors constructs and the explosive use of the NASA-TLX: a measurement perspective. Cogn. Technol. Work. 16 (3), 289–297. http://dx.doi.org/10.1007/s10111-014-0275-1.

Di Campli San Vito, Patrizia, Shakeri, Gözel, Brewster, Stephen, Pollick, Frank, Brown, Edward, Skrypchuk, Lee, Mouzakitis, Alexandros, 2019. Haptic navigation cues on the steering wheel. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–11. http://dx.doi.org/10.1145/3290605.3300440.

Dismukes, Robert Key (Ed.), 2017. Human Error in Aviation. Routledge, http://dx.doi.org/10.4324/9781315092898.

Emerson, Terry J., Reising, John M., Britten-Austin, Harold G., 1987. Workload and situation awareness in future aircraft. SAE Trans. 96, 1130–1137, http://www.jstor.org/stable/44473023.

Estes, Steven, 2015. The workload curve: Subjective mental workload. Hum. Factors 57 (7), 1174–1187. http://dx.doi.org/10.1177/0018720815592752.

Finomore, Victor S., Shaw, Tyler H., Warm, Joel S., Matthews, Gerald, Weldon, Dave, Boles, David B., 2009. On the workload of vigilance: Comparison of the NASA-TLX and the MRQ. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 53 (17), 1057–1061. http://dx.doi.org/10.1177/154193120905301703.

Finomore, Jr., Victor S., Shaw, Tyler H., Warm, Joel S., Matthews, Gerald, Boles, David B., 2013. Viewing the workload of vigilance through the lenses of the NASA-TLX and the MRQ. Hum. Factors 55 (6), 1044–1063. http://dx.doi.org/10.1177/0018720813484498, PMID: 24745198.

Fréard, Dominique, Jamet, Eric, Le Bohec, Olivier, Poulain, Gérard, Botherel, Valérie, 2007. Subjective measurement of workload related to a multimodal interaction task: NASA-TLX vs. Workload profile. In: Jacko, Julie A. (Ed.), Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 60–69.

Galy, Edith, Cariou, Magali, Mélan, Claudine, 2012. What is the relationship between mental workload factors and cognitive load types? Int. J. Psychophysiol. 83 (3), 269–275. http://dx.doi.org/10.1016/j.ijpsycho.2011.09.023.

Galy, Edith, Mélan, Claudine, 2015. Effects of cognitive appraisal and mental workload factors on performance in an arithmetic task. Appl. Psychophys. Biof. 40 (4), 313–325. http://dx.doi.org/10.1007/s10484-015-9302-0.

Galy, Edith, Paxion, Julie, Berthelon, Catherine, 2018. Measuring mental workload with the NASA-TLX needs to examine each dimension rather than relying on the global score: an example with driving. Ergonomics 61 (4), 517–527. http://dx.doi.org/10.1080/00140139.2017.1369583, PMID: 28817353.

Gavas, Rahul, Chatterjee, Debatri, Sinha, Aniruddha, 2017. Estimation of cognitive load based on the pupil size dilation. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics. SMC, pp. 1499–1504. http://dx.doi.org/10.1109/SMC.2017.8122826.

Gerjets, Peter, Scheiter, Katharina, Catrambone, Richard, 2004. Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures. Instr. Sci. 32 (1/2), 33–58, http://www.jstor.org/stable/41953636.

Gerjets, Peter, Scheiter, Katharina, Catrambone, Richard, 2006. Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? Learn. Instr. 16 (2), 104–121. http://dx.doi.org/10.1016/j.learninstruc.2006.02.007, Recent Worked Examples Research: Managing Cognitive Load to Foster Learning and Transfer.

Gliem, Joe A., Gliem, Rosemary R., 2003. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for likert-type scales.

Gopher, Daniel, Donchin, Emanuel, 1986. Workload: An examination of the concept. In: Boff, K.R., Kaufman, L., Thomas, J.P. (Eds.), Handbook of Perception and Human Performance, Vol. 2. Cognitive Processes and Performance. John Wiley & Sons, pp. 1–49.

Gregory, Robert J., 2014. Psychological Testing: History, Principles, and Applications, Global Edition, Seventh ed. Pearson Education, London, England.

Grier, Rebecca A., 2015. How high is high? A meta-analysis of NASA-TLX global workload scores. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 59 (1), 1727–1731. http://dx.doi.org/10.1177/1541931215591373.

Hair, Joseph F., Anderson, Rolph E., Tatham, Ronald L., Black, William C., 1998. Factorial analysis. In: Multivariate Data Analysis, fifth ed. Prentice Hall, New Jersey.

Hair, Joseph F., Risher, Jeffrey J., Sarstedt, Marko, Ringle, Christian M, 2019. When to use and how to report the results of PLS-SEM. Eur. Bus. Rev. 31 (1), 2–24.

Hancock, Peter A., 1989a. A dynamic model of stress and sustained attention. Hum. Factors 31 (5), 519–537. http://dx.doi.org/10.1177/001872088903100503, PMID: 2625347.

Hancock, Peter A., 1989b. The effect of performance failure and task demand on the perception of mental workload. Appl. Ergon. 20 (3), 197–205. http://dx.doi.org/10.1016/0003-6870(89)90077-X.

Hancock, Peter A., Caird, Jeff K., 1993. Experimental evaluation of a model of mental workload. Hum. Factors 35 (3), 413–429.

Hancock, Gabriella M., Longo, Luca, Young, Mark S., Hancock, Peter A., 2021. Mental workload. In: Handbook of Human Factors and Ergonomics. John Wiley & Sons, Ltd, pp. 203–226. http://dx.doi.org/10.1002/9781119636113.ch7.

Hancock, Peter A., Matthews, Gerald, 2019. Workload and performance: Associations, insensitivities, and dissociations. Hum. Factors 61 (3), 374–392. http://dx.doi.org/10.1177/0018720818809590, PMID: 30521400.

Hanoch, Yaniv, Vitouch, Oliver, 2004. When less is more: Information, emotional arousal and the ecological reframing of the Yerkes-Dodson law. Theory Psychol. 14 (4), 427–452. http://dx.doi.org/10.1177/0959354304044918.

Harris, David, Wilson, Mark, Vine, Samuel, 2019. Development and validation of a simulation workload measure: the simulation task load index (SIM-TLX). Virtual Real. 24 (4), 557–566. http://dx.doi.org/10.1007/s10055-019-00422-9.

Hart, Sandra G., 2006. Nasa-Task Load Index (NASA-TLX); 20 years later. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 50 (9), 904–908. http://dx.doi.org/10.1177/154193120605000909.

Hart, Sandra G., Staveland, Lowell E., 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock, Peter A., Meshkati, Najmedin (Eds.), Human Mental Workload. In: Advances in Psychology, vol. 52, North-Holland, pp. 139–183. http://dx.doi.org/10.1016/S0166-4115(08)62386-9.

Hayashi, Takahiro, Kishi, Reo, 2014. Utilization of NASA-TLX for workload evaluation of Gaze-writing systems. In: 2014 IEEE International Symposium on Multimedia. pp. 271–272. http://dx.doi.org/10.1109/ISM.2014.18.

Hertzum, Morten, 2021. Reference values and subscale patterns for the task load index (TLX): a meta-analytic review. Ergonomics 64 (7), 869–878. http://dx.doi.org/10.1080/00140139.2021.1876927, PMID: 33463402.

Hollender, Nina, Hofmann, Cristian, Deneke, Michael, Schmitz, Bernhard, 2010. Integrating cognitive load theory and concepts of human–computer interaction. Comput. Hum. Behav. 26 (6), 1278–1288. http://dx.doi.org/10.1016/j.chb.2010.05.031, Online Interactivity: Role of Technology in Behavior Change.

Howard, Zachary L., Evans, Nathan J., Innes, Reilly J., Brown, Scott D., Eidels, Ami, 2020. How is multi-tasking different from increased difficulty? Psychon. Bull. Rev. 27 (5), 937–951. http://dx.doi.org/10.3758/s13423-020-01741-8.

Hyland, Michael, 1981. Hypothetical constructs and intervening variables. In: Introduction To Theoretical Psychology. Macmillan Education UK, London, pp. 32–41. http://dx.doi.org/10.1007/978-1-349-16464-6_3.

Israel, Glenn D., 1992. Determining Sample Size (Fact Sheet PEOD-6). University of Florida, Gainesville, FL.

Isreal, Jack B., Chesney, Gregory L., Wickens, Christopher D., Donchin, Emanuel, 1980. P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. Psychophysiology 17 (3), 259–273. http://dx.doi.org/10.1111/j.1469-8986.1980.tb00146.x.

Jabeen, Farzana, Tao, Linmi, 2017. An efficient text entry model for scanning ambiguous keyboard. In: 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics, Vol. 1. IHMSC, pp. 71–76. http://dx.doi.org/10.1109/IHMSC.2017.23.

Jabeen, Farzana, Tao, Linmi, Wang, Xinyue, Mei, Shanshan, 2018. C-SAK: Chinese scanning ambiguous keyboard for Parkinson's disease patients. In: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress. DASC/PiCom/DataCom/CyberSciTech, pp. 792–799. http://dx.doi.org/10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00137.

Jaeggi, Susanne M., Seewer, Ria, Nirkko, Arto C., Eckstein, Doris, Schroth, Gerhard, Groner, Rudolf, Gutbrod, Klemens, 2003. Does excessive memory load attenuate activation in the prefrontal cortex? Load-dependent processing in single and dual tasks: functional magnetic resonance imaging study. NeuroImage 19 (2), 210–225. http://dx.doi.org/10.1016/S1053-8119(03)00098-3.

Jahns, Dieter W., 1973. Concept of Operator Workload in Manual Vehicle Operations. Technical Report No. 14, Forschungsingstitut Anthropotechnik.

Johansson, Gunnar, von Hofsten, Claes, Jansson, Gunnar, 1980. Event perception. Annu. Rev. Psychol. 31, 27.

Kahneman, D., 1973. Attention and effort. Prentice Hall,

Kantowitz, Barry H., 2000. Attention and mental workload. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 44 (21), 3–456–3–459. http://dx.doi.org/10.1177/154193120004402121.

Keller, Johannes, Bless, Herbert, Blomann, Frederik, Kleinböhl, Dieter, 2011. Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. J. Exp. Soc. Psychol. 47 (4), 849–852. http://dx.doi.org/10.1016/j.jesp.2011.02.004.

Kjærup, Maria, Skov, Mikael B., Nielsen, Peter Axel, Kjeldskov, Jesper, Gerken, Jens, Reiterer, Harald, 2021. Longitudinal studies in HCI research: A review of CHI publications from 1982–2019. In: Karapanos, Evangelos, Gerken, Jens, Kjeldskov, Jesper, Skov, Mikael B. (Eds.), Advances in Longitudinal HCI Research. Springer International Publishing, Cham, pp. 11–39. http://dx.doi.org/10.1007/978-3-030-67322-2_2.

Klein, Stanley A., 2001. Measuring, estimating, and understanding the psychometric function: A commentary. Perception; Psychophys. 63 (8), 1421–1455. http://dx.doi.org/10.3758/bf03194552.

Klepsch, Melina, Schmitz, Florian, Seufert, Tina, 2017. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Front. Psychol. 8 (NOV), 1–18. http://dx.doi.org/10.3389/fpsyg.2017.01997.

Kock, Ned, 2015. Common method bias in PLS-SEM: A full collinearity assessment approach. Int. J. E-Collab. 11, 1–10. http://dx.doi.org/10.4018/ijec.2015100101.

Kock, Ned, Texas A&M International University, Lynn, Gary, Stevens Institute of Technology, 2012. Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. J. Assoc. Inf. Syst. 13 (7), 546–580.

Kokini, Christina M., Lee, Sangwon, Koubek, Richard J., Moon, Seung Ki, 2012. Considering context: The role of mental workload and operator control in users' perceptions of usability. Int. J. Human–Comput. Interact. 28 (9), 543–559. http://dx.doi.org/10.1080/10447318.2011.622973.

Kosch, Thomas, Hassib, Mariam, Woźniak, Paweł W., Buschek, Daniel, Alt, Florian, 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, Association for Computing Machinery, New York, NY, USA, pp. 1–13. http://dx.doi.org/10.1145/3173574.3174010.

Kosch, Thomas, Karolus, Jakob, Zagermann, Johannes, Reiterer, Harald, Schmidt, Albrecht, Woźniak, Paweł W., 2023. A survey on measuring cognitive workload in human-computer interaction. ACM Comput. Surv. 55 (13s), http://dx.doi.org/10.1145/3582272.

Kramer, Arthur F., 2020. Physiological metrics of mental workload: A review of recent progress. Multiple-Task Perform. 279–328.

Leplat, Jacques, 1978. Factors determining work-load. Ergonomics 21 (3), 143–149. http://dx.doi.org/10.1080/00140137808931709, PMID: 27349.

Leppink, Jimmie, Paas, Fred, Van der Vleuten, Cees P.M., Van Gog, Tamara, Van Merriënboer, Jeroen J.G., 2013. Development of an instrument for measuring different types of cognitive load. Behav. Res. Methods 45 (4), 1058–1072. http://dx.doi.org/10.3758/s13428-013-0334-1.

Lewis, James R., 2002. Psychometric evaluation of the PSSUQ using data from five years of usability studies. Int. J. Human–Comput. Interact. 14 (3–4), 463–488. http://dx.doi.org/10.1080/10447318.2002.9669130.

Ley, Jacqui M., McGreevy, Paul, Bennett, Pauleen C., 2009. Inter-rater and test–retest reliability of the monash canine personality questionnaire-revised (MCPQ-R). Appl. Anim. Behav. Sci. 119 (1), 85–90. http://dx.doi.org/10.1016/j.applanim.2009.02.027.

Liddell, Torrin M., Kruschke, John K., 2018. Analyzing ordinal data with metric models: What could possibly go wrong? J. Exp. Soc. Psychol. 79, 328–348.

Longo, Luca, 2017. Subjective usability, mental workload assessments and their impact on objective human performance. In: Bernhaupt, Regina, Dalvi, Girish, Joshi, Anirudha, K. Balkrishan, Devanuj, O'Neill, Jacki, Winckler, Marco (Eds.), Human-Computer Interaction - INTERACT 2017. Springer International Publishing, Cham, pp. 202–223.

Longo, Luca, 2018. Experienced mental workload, perception of usability, their interaction and impact on task performance. PLoS One 13 (8), 1–36. http://dx.doi.org/10.1371/journal.pone.0199661.

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

Longo, Luca, Orru, Giuliano, 2019. An evaluation of the reliability, validity and sensitivity of three human mental workload measures under different instructional conditions in third-level education. In: McLaren, Bruce M., Reilly, Rob, Zvacek, Susan, Uhomoibhi, James (Eds.), Computer Supported Education. Springer International Publishing, Cham, pp. 384–413.

Longo, Luca, Wickens, Christopher D., Hancock, Gabriella M., Hancock, Peter A., 2022. Human mental workload: A survey and a novel inclusive definition. Front. Psychol. 13, http://dx.doi.org/10.3389/fpsyg.2022.883321.

Mack, Kelly, McDonnell, Emma, Jain, Dhruv, Lu Wang, Lucy, E. Froehlich, Jon, Findlater, Leah, 2021. What do we mean by "Accessibility Research"? A literature survey of accessibility papers in CHI and ASSETS from 1994 to 2019. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3411764.3445412.

MacKenzie, I. Scott, 2009. The one-key challenge: Searching for a fast one-key text entry method. In: Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility. Assets '09, Association for Computing Machinery, New York, NY, USA, pp. 91–98. http://dx.doi.org/10.1145/1639642.1639660.

MacKenzie, I. Scott, Felzer, Torsten, 2010. SAK: Scanning ambiguous keyboard for efficient one-key text entry. ACM Trans. Comput. Hum. Interact. 17 (3), 11:1–11:39. http://dx.doi.org/10.1145/1806923.1806925.

MacKenzie, I. Scott, Felzer, Torsten, 2022. ScanningKeyboardExperiment. http://www.yorku.ca/mack/ExperimentSoftware/doc/ScanningKeyboardExperiment.html. (Accessed 08 August 2022).

Malacria, Sylvain, Bailly, Gilles, Harrison, Joel, Cockburn, Andy, Gutwin, Carl, 2013. Promoting hotkey use through rehearsal with exposehk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 573–582. http://dx.doi.org/10.1145/2470654.2470735.

Matthews, Gerald, De Winter, Joost, Hancock, Peter A., 2020. What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. Theor. Issues Ergon. Sci. 21 (4), 369–396. http://dx.doi.org/10.1080/1463922X.2018.1547459.

Matthews, Gerald, Joyner, Lucy, Gilliland, Kirby, Campbell, Sian, Falconer, Shona, Huggins, Jane, 1999. Validation of a comprehensive stress state questionnaire: Towards a state 'Big Three'? In: Personality Psychology in Europe, Vol. 7. Tilburg University Press, pp. 335–350.

Matthews, Gerald, Reinerman-Jones, Lauren E., Barber, Daniel J., Abich, Julian, 2014. The psychometrics of mental workload: Multiple measures are sensitive but divergent. Hum. Factors: J. Hum. Factors Ergon. Soc. 57 (1), 125–143. http://dx.doi.org/10.1177/0018720814539505.

Matthews, Gerald, Reinerman-Jones, Lauren E., Barber, Daniel J., Abich, IV, Julian, 2015a. The psychometrics of mental workload: Multiple measures are sensitive but divergent. Hum. Factors 57 (1), 125–143. http://dx.doi.org/10.1177/0018720814539505, PMID: 25790574.

Matthews, Gerald, Reinerman-Jones, Lauren, Wohleber, Ryan, Lin, Jinchao, Mercado, Joe, Abich, Julian, 2015b. Workload is multidimensional, not unitary: What now? In: Schmorrow, Dylan D., Fidopiastis, Cali M. (Eds.), Foundations of Augmented Cognition. Springer International Publishing, Cham, pp. 44–55.

McKendrick, Ryan D., Cherry, Erin, 2018. A deeper look at the NASA TLX and where it falls short. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 62 (1), 44–48. http://dx.doi.org/10.1177/1541931218621010.

Merton, Robert K., 1968. The matthew effect in science. Science 159 (3810), 56–63. http://dx.doi.org/10.1126/science.159.3810.56.

Moray, Neville, 1979. Models and measures of mental workload. In: Mental Workload. Springer US, pp. 13–21. http://dx.doi.org/10.1007/978-1-4757-0884-4_2.

Moray, Neville, 2013. Mental Workload: Its Theory and Measurement, Vol. 8. Springer Science & Business Media.

Moroney, William F., Biers, David W., Thomas Eggemeier, F., 1995. Some measurement and methodological considerations in the application of subjective workload measurement techniques. Int. J. Aviat. Psychol. 5 (1), 87–106. http://dx.doi.org/10.1207/s15327108ijap0501_6.

Moustafa, Karim, Luz, Saturnino, Longo, Luca, 2017. Assessment of mental workload: A comparison of machine learning methods and subjective assessment techniques. In: Longo, Luca, Leva, M. Chiara (Eds.), Human Mental Workload: Models and Applications. Springer International Publishing, Cham, pp. 30–50.

Noyes, Jan M., Bruneau, Daniel P.J., 2007. A self-analysis of the NASA-TLX workload measure. Ergonomics 50 (4), 514–519. http://dx.doi.org/10.1080/00140130701235232.

Nygren, Thomas E., 1991. Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. Hum. Factors 33 (1), 17–33. http://dx.doi.org/10.1177/001872089103300102.

Paas, Fred G., 1992. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. J. Educ. Psychol. 84 (4), 429–434.

Paas, Fred, van Gog, Tamara, 2006. Optimising worked example instruction: Different ways to increase germane cognitive load. Learn. Instr. 16 (2 SPEC. ISS.), 87–91. http://dx.doi.org/10.1016/j.learninstruc.2006.02.004.

Paas, Fred G.W.C., Van Merriënboer, Jeroen J.G., 1993. The efficiency of instructional conditions: An approach to combine mental effort and performance measures. Hum. Factors 35 (4), 737–743. http://dx.doi.org/10.1177/001872089303500412.

Pauzie', Annie, 2008. A method to assess the driver mental workload: The driving activity load index (DALI). IET Intell. Transp. Syst. 2 (4), 315. http://dx.doi.org/10.1049/iet-its:20080023.

Perry-Houts, Jonathan, 2022. jperryhouts/Dual-N-Back: Memory game that's been shown to improve working memory.. https://github.com/jperryhouts/Dual-N-Back. (Accessed 09 August 2022).

Ramkumar, Anjana, Stappers, Pieter Jan, Niessen, Wiro J., Adebahr, Sonja, Schimek-Jasch, Tanja, Nestle, Ursula, Song, Yu, 2017. Using GOMS and NASA-TLX to evaluate human–computer interaction process in interactive segmentation. Int. J. Human–Comput. Interact. 33 (2), 123–134. http://dx.doi.org/10.1080/10447318.2016.1220729.

Reid, Gary B., Nygren, Thomas E., 1988. The subjective workload assessment technique: A scaling procedure for measuring mental workload. In: Hancock, Peter A., Meshkati, Najmedin (Eds.), Human Mental Workload. In: Advances in Psychology, vol. 52, North-Holland, pp. 185–218. http://dx.doi.org/10.1016/S0166-4115(08)62387-0.

Romero, Joaquim, 2017. An Investigation of the Correlation between Mental Workload and Web User's Interaction (Ph.D. thesis).

Roscoe, Alan H., Ellis, Georges A., 1990. A Subjective Rating Scale for Assessing Pilot Workload in Flight: A Decade of Practical Use. Technical Report, Royal Aerospace Establishment.

Rubio, Susana, Díaz, Eva, Martín, Jesús, Puente, José M., 2004. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. Appl. Psychol. 53 (1), 61–86. http://dx.doi.org/10.1111/j.1464-0597.2004.00161.x.

Schnotz, Wolfgang, Kürschner, Christian, 2007. A reconsideration of cognitive load theory. Educ. Psychol. Rev. 19 (4), 469–508. http://dx.doi.org/10.1007/s10648-007-9053-4.

Sperandio, Jean-Claude, 1978. The regulation of working methods as a function of work-load among air traffic controllers. Ergonomics 21 (3), 195–202. http://dx.doi.org/10.1080/00140137808931713.

Staal, Mark A., 2004. Stress, cognition, and human performance: A literature review and conceptual framework.

Stager, Paul, 1991. Error models for operating irregularities: Implications for automation. In: Wise, John A., Hopkin, V. David, Smith, Marvin L. (Eds.), Automation and Systems Issues in Air Traffic Control. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 321–338.

Stanton, Neville Anthony, Hedge, Alan, Brookhuis, Karel, Salas, Eduardo, Hendrick, Hal W. (Eds.), 2004. Handbook of Human Factors and Ergonomics Methods. CRC Press, http://dx.doi.org/10.1201/9780203489925.

Sweller, John, 2010. Element interactivity and intrinsic, extraneous, and germane cognitive load. Educ. Psychol. Rev. 22 (2), 123–138. http://dx.doi.org/10.1007/s10648-010-9128-5.

Sweller, John, 2011. Cognitive load theory. In: Mestre, Jose P., Ross, Brian H. (Eds.), In: Psychology of Learning and Motivation, vol. 55, Academic Press, pp. 37–76. http://dx.doi.org/10.1016/B978-0-12-387691-1.00002-8.

Sweller, John, Ayres, Paul, Kalyuga, Slava, 2011. Measuring cognitive load. In: Cognitive Load Theory. Springer New York, New York, NY, pp. 71–85. http://dx.doi.org/10.1007/978-1-4419-8126-4_6.

Sweller, John, van Merrienboer, Jeroen J.G., Paas, Fred G.W.C., 1998. Cognitive architecture and instructional design. Educ. Psychol. Rev. 10 (3), 251–296. http://dx.doi.org/10.1023/A:1022193728205.

Tattersall, A.J., Foord, P.S., 1996. An experimental evaluation of instantaneous self-assessment as a measure of workload. Ergonomics 39 (5), 740–748. http://dx.doi.org/10.1080/00140139608964495.

Tavakol, Mohsen, Dennick, Reg, 2011. Making sense of Cronbach's alpha. Int. J. Med. Educ. 2, 53–55. http://dx.doi.org/10.5116/ijme.4dfb.8dfd.

Thorpe, Alexander, Nesbitt, Keith, Eidels, Ami, 2020. A systematic review of empirical measures of workload capacity. ACM Trans. Appl. Percept. 17 (3), http://dx.doi.org/10.1145/3422869.

Tsang, Pamela S., Vidulich, Michael A., 2006. Mental workload and situation awareness. In: Handbook of Human Factors and Ergonomics. John Wiley & Sons, Ltd, pp. 243–268. http://dx.doi.org/10.1002/0470048204.ch9, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470048204.ch9.

United States. National Aeronautics and Space Administration, 1988. NASA conference publication. In: NASA Conference Publication, no. 2504, Scientific and Technical Information Office, National Aeronautics and Space Administration.

Van Acker, Bram B., Parmentier, Davy D., Vlerick, Peter, Saldien, Jelle, 2018. Understanding mental workload: from a clarifying concept analysis toward an implementable framework. Cogn. Technol. Work. 20 (3), 351–365. http://dx.doi.org/10.1007/s10111-018-0481-3.

Vashistha, Aditya, Sethi, Pooja, Anderson, Richard, 2017. Respeak: A voice-based, crowd-powered speech transcription system. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1855–1866. http://dx.doi.org/10.1145/3025453.3025640.

Vygotsky, Lev Semenovich, Cole, Michael, 1978. Mind in Society: Development of Higher Psychological Processes. Harvard University Press.

Wacharamanotham, Chat, Eisenring, Lukas, Haroz, Steve, Echtler, Florian, 2020. Transparency of CHI research artifacts: Results of a self-reported survey. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–14. http://dx.doi.org/10.1145/3313831.3376448.

Waddington, Chris T., MacKenzie, I. Scott, Read, Janet C., Horton, Matthew, 2017. Comparing a scanning ambiguous keyboard to the on-screen QWERTY keyboard. In: Electronic Visualisation and the Arts. EVA 2017, pp. 1–6.

Welford, Alan T., 1978. Mental work-load as a function of demand, capacity, strategy and skill. Ergonomics 21 (3), 151–167. http://dx.doi.org/10.1080/00140137808931710.

Wichmann, Felix A., Hill, N. Jeremy, 2001. The psychometric function: I. Fitting, sampling, and goodness of fit. Perception; Psychophys. 63 (8), 1293–1313. http://dx.doi.org/10.3758/bf03194544.

Wickens, Christopher D., 2002. Multiple resources and performance prediction. Theor. Issues Ergon. Sci. 3 (2), 159–177.

Wickens, Christopher D., 2008. Multiple resources and mental workload. Hum. Factors 50 (3), 449–455. http://dx.doi.org/10.1518/001872008X288394.

Wickens, Christopher D., Boles, David, Tsang, Pamela, Carswell, Melody, 1984. The Limits of Multiple Resource Theory in Display Formatting: Effects of Task Integration. Technical Report, Illinois University at Urbana-Champaign.

Wiebe, Eric N., Roberts, Edward, Behrend, Tara S., 2010. An examination of two mental workload measurement approaches to understanding multimedia learning. Comput. Hum. Behav. 26 (3), 474–481. http://dx.doi.org/10.1016/j.chb.2009.12.006.

Wilson, Max L., 2023. Mental workload vs cognitive load vs everything else in HCI. https://medium.com/@cogpi/mental-workload-vs-cognitive-load-vs-everything-else-in-hci-575722d14572. (Accessed 26 February 2025).

Wilson, Mark R., Poolton, Jamie M., Malhotra, Neha, Ngo, Karen, Bright, Elizabeth, Masters, Rich S.W., 2011. Development and validation of a surgical workload measure: The surgery task load index (SURG-TLX). World J. Surg. 35 (9), 1961–1969. http://dx.doi.org/10.1007/s00268-011-1141-4.

Xie, Bin, Salvendy, Gavriel, 2000. Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. Work. Stress. 14 (1), 74–99. http://dx.doi.org/10.1080/026783700417249.

Young, Mark S., Brookhuis, Karel A., Wickens, Christopher D., Hancock, Peter A., 2015. State of science: mental workload in ergonomics. Ergonomics 58 (1), 1–17. http://dx.doi.org/10.1080/00140139.2014.956151, PMID: 25442818.

Zijlstra, Ferdinand R.H., 1993. The Construction of a Scale to Measure Perceived Effort. TU Delft, The Netherlands.

Züger, Manuela, Fritz, Thomas, 2015. Interruptibility of software developers and its prediction using psycho-physiological sensors. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 2981–2990. http://dx.doi.org/10.1145/2702123.2702593.