TU WIEN Informatics

Über die Bestimmung Kognitiver Belastung bei der Nutzung von UbiComp-Systemen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Media and Human-Centered Computing

eingereicht von

Aaron Wedral, BSc
Matrikelnummer 01633070

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Assistant Prof. Dr.in phil. Mag. a phil. Astrid Weiss

Wien, 25. August 2023

Aaron Wedral	Astrid Weiss

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Cognitive Load of Pervasive Technology: How can it be determined?

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Media and Human-Centered Computing

by

Aaron Wedral, BSc
Registration Number 01633070

to the Faculty of Informatics

at the TU Wien

Advisor: Assistant Prof. Dr.in phil. Mag. a phil. Astrid Weiss

Vienna, 25th August, 2023

Aaron Wedral	Astrid Weiss

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Aaron Wedral, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 25. August 2023

Aaron Wedral

Danksagung

Ich möchte meiner Betreuerin für die rasche und ständige Unterstützung und Anleitung danken, die sie mir während des Schreibens dieser Arbeit bei allen Erfolgen und Rückschlägen gewährt hat. Und dafür, dass sie diese lange Zeit mit mir durchgestanden hat. Dankeschön. Ich möchte auch den lieben und hilfsbereiten Projektpartnern bei Profactor für die Entwicklung der getesteten Technologie und die Ermöglichung der Nutzer*innenstudie in dieser Arbeit danken. Es hat mich gefreut mit euch zu arbeiten. Danke an Rafael für die Hilfe und Begleitung bei der Planung und Durchführung der Nutzer*innenstudie. Ohne dich hätte ich es nicht geschafft. Nochmals vielen Dank an alle Versuchsteilnehmer*innen für ihre Zeit, Zusammenarbeit und Ehrlichkeit. Weiters danke ich den Expert*innen für ihre Teilnahme am Cognitive Walkthrough. Danke an die herzliche Gemeinschaft in der Human Computer Interaction Group, die meinem Gefasel gelauscht, mich unterstützt haben und mir einen Platz zum Schreiben gaben, wenn ich einen Tapetenwechsel brauchte. Und schließlich möchte ich Iris und meiner Familie danken. Ohne euch wäre dieses Unterfangen kaum zu bewältigen gewesen.

Acknowledgements

I would like to thank my supervisor for the immediate and constant support and guidance she has shown me during the process of my writing this thesis through every success and setback. And bearing with me through all this time. Thank you. I would also like to thank the nice and helpful project partners at Profactor for developing the tested technology and enabling the user study found in this thesis. It was a pleasure working with you. Thank you Rafael for the aid and companionship in planning and conducting the user study. I could not have done it without you. Thank you again, to all of the experiment participants for your time, cooperation and honesty. Furthermore, I want to thank the experts for their time and experience provided in the Cognitive Walkthrough. Thank you, to the lovely people at the Human Computer Interaction Group for listening to my ramblings about this thesis, giving support and guidance, as well as providing me a place to write when I needed a change of scene. Lastly, I want to thank Iris and my family. Such an endeavour would have been hardly achievable without you.

Kurzfassung

In den letzten Jahren hat *Cognitive Load* im Bereich der Mensch-Computer-Interaktion, insbesondere im Bereich *Ubiquitous-Computing (UbiComp)*, als ein wichtiger Faktor für die menschliche Leistung und das Lernen an Bedeutung gewonnen. Vor allem, da im UbiComp-Bereich viele die genaue Bestimmung von *Cognitive Load* zu einem wichtigen Meilenstein für adaptive Automatisierung erklärt haben. Bis zur Erstellung dieser Arbeit wurden zahlreiche Methoden entwickelt, um die Klassifizierungsgenauigkeit zu erhöhen, wobei sich die jüngste Forschung auf objektive Daten konzentriert, die in Echtzeit für die Verwendung in der adaptiven Automatisierung für UbiComp-Systeme analysiert werden können.

Die meisten Methoden schaffen es jedoch kaum, die Klassifizierungsgenauigkeit einer einfachen Likert-Skala zur Selbsteinschätzung zu übertreffen. Außerdem wiesen die verwendeten Methoden erhebliche methodische Mängel auf, die die Interpretierbarkeit ihrer Ergebnisse einschränken. *Cognitive Load* ist definiert als die Menge an *Working Memory*, die von bestimmten Lernenden mit ihrem spezifischen Vorwissen während einer bestimmten Lernaktivität verwendet wird. Die in dieser Arbeit untersuchten Methoden bezogen *Cognitive Load Theory* jedoch weder in ihre Interpretation der Ergebnisse ein noch gaben sie genau an, dass sie tatsächlich versuchten, die aktuelle Nutzung des *Working Memory* zu messen. Außerdem kontrollierten sie nicht das Vorwissen ihrer Proband*innen. Darüber hinaus ist nicht gewährleistet, dass die von ihnen verwendeten Methoden zur Induzierung mentaler Arbeitsbelastung bei jeder Person ein ähnliches Niveau der Nutzung des *Working Memory* hervorrufen, insbesondere da das Vorwissen nicht kontrolliert wurde. Nichtsdestotrotz wurden die Ergebnisse als *Cognitive Load* bezeichnet. Während die gegenwärtigen Mängel die Methoden für eine zuversichtliche Verwendung für die adaptive Automatisierung fragwürdig machen, bietet die Verwendung von Messungen von *Cognitive Load* zur Bewertung der *Technology Adoption* eine Alternative. Da es sich bei *Technology Adoption* im Wesentlichen um das Erlernen des Umgangs mit einer neuen Technologie handelt, könnten Messungen von *Cognitive Load* hier wertvolle Dienste leisten.

Mit diesem Ziel vor Augen teste ich einen gemischten Methodenansatz zur Schätzung von *Cognitive Load* und analysiere die Ergebnisse durch die Brille der *Cognitive Load Theory*. Die Methoden wurden in einer Nutzer*innen-Studie zur Evaluierung der Nutzbarkeit eines *Spatial-Augmented-Reality*-Systems mit fünfzehn repräsentativen Zielnutzer*innen getestet. Die getesteten Methoden waren der NASA-TLX, eine Sekundäraufgabe, die

Verwendung von Verhaltensmaßen, *Learnability*-Daten und ein angepasster *Cognitive Walkthrough*.

Beim Vergleich der möglichen Aussagen jeder einzelnen Methode wird deutlich, dass eine Methode allein keine ganzheitliche Sichtweise erlaubt, um das Geschehen in der Interaktion überzeugend zu erfassen, und zu viele Annahmen für eine detaillierte Interpretation erfordert. Erst durch die Einbeziehung mehrerer Methoden können Ursachen für Befunde durch Daten gestützt und relativiert werden, um ein schlüssiges und überzeugendes Bild des *Cognitive Load* der Nutzer*innen beim Erlernen der Interaktion zu zeichnen. Dennoch lieferten einige Methoden nicht die erwartete Qualität an Daten, während andere zu viel Aufwand erforderten, um angesichts ihrer Erklärungskraft eine künftige Verwendung zu rechtfertigen.

Auf der Grundlage meiner Ergebnisse vertrete ich die Auffassung, dass die Methodik der nahen Zukunft ein *Toolkit* mit gemischten Methoden zur Schätzung von *Cognitive Load* umfassen sollte. Dieses *Toolkit* kann ohne weiteres Methoden enthalten, die ich in dieser Arbeit aufgrund ihrer nicht belegten Behauptungen kritisiert habe, sofern ihre Grenzen in Aussagen miteinbezogen werden. *Cognitive Load Theory* ist keineswegs unumstritten und bedarf noch weiterer empirischer Validierung und Verfeinerung. Daher sollte die Methodik zur Bewertung des Phänomens so ganzheitlich wie möglich sein, um ihre Verfeinerung zu unterstützen und Schwächen oder Widersprüche aufzudecken. In dem Maße, wie Methodik und Theorie verfeinert werden, wird eine post-positivistische Sichtweise des Phänomens immer nützlicher. Solange es jedoch keine gefestigten Anhaltspunkte gibt, von der aus quantitative Ergebnisse interpretiert werden können, halte ich einen konstruktivistischen Ansatz mit gemischten Methoden, wie er in dieser Arbeit vorgeschlagen wird, für sinnvoller.

Abstract

In recent years, cognitive load has risen in importance in the field of human-computer interaction, particularly in the ubiquitous computing community, as an important factor for human performance and learning. Especially, since many in the field have proclaimed its accurate detection a necessity for adaptive automation. Until the time of writing, many method have been developed to estimate it chasing ever higher classification accuracy with recent research focusing on objective data which can be analysed in real-time for use in adaptive automation for ubiquitous and pervasive computer systems.

For most methods, however, the classification accuracy is barely comparable to a simple self-reporting Likert Scale. Additionally, the used methods had major methodological flaws limiting their interpretation of results. Cognitive load is defined as the amount of working memory used by a specific learner with their specific prior knowledge during a given learning activity. With little reference to the name giving Cognitive Load Theory, the methods analysed in this work neither included the theory in their interpretation of findings nor accurately attributed that they were actually trying to measure current working memory use. Additionally, they did not control for prior knowledge in their samples. Furthermore, their methods used to induce mental workload are not guaranteed to induce similar levels of working memory use for each individual, especially since prior knowledge was not controlled for. Nonetheless, the methods called their results cognitive load. While the current flaws of methodology make the methods questionable for confident use for adaptive automation, the use of cognitive load measures to evaluate technology adoption provides an alternative. As technology adoption is in essence learning how to use new technology, cognitive load measures could be of value here.

With this goal in mind, I explore a mixed method approach of estimating cognitive load and analyse the findings through a lens of Cognitive Load Theory. The methods are tested in a user study evaluating the usability of a Spatial Augmented Reality system with fifteen representative target users. The tested methods are the NASA-TLX, a secondary task, the use of behavioural measures, learnability measures and an adapted cognitive walkthrough.

When comparing the claims possible with each individual method, it quickly becomes clear that one method alone did not allow a holistic enough view to confidently capture what transpired in the interaction and requiring to many assumptions for accurate interpretation. Only when including multiple methods, causes for findings could be

supported by data and put into perspective to paint a conclusive and convincing picture of the cognitive load of users learning the interaction. Nonetheless, some methods did not provide the expected quality of data while others required too much effort to warrant future use given their explanatory power.

Based on my findings, I argue that near-future methodology should encompass a toolkit of mixed methods to estimate cognitive load. This toolkit can include methods earlier criticised for their unsupported claims if their limitations are acknowledged and counter-balanced by other methods. Cognitive Load Theory is by no means undisputed and still requires additional validation as well as refinement. Therefore, methodology to evaluate the phenomenon should aim to be as holistic as possible to aid in its refinement and identify weaknesses or contradictions. As methodology and theory get refined in turn, a post-positivist view on the phenomenon increasingly becomes useful. Until there is a proven baseline to interpret post-positivist findings from, however, I deem a constructivist mixed-method approach as suggested by this work more sensible.

Contents

CHAPTER 1

Introduction

In Human-Computer Interaction (HCI), there are a plethora of methods to research and analyse interactions between humans and computers, most of which were developed at a time where computers were mainly personal computers and interactions could mostly be done via mouse and keyboard. Of key interest was and still is usability: How efficiently, correctly and pleasantly a technology is to work with. Meanwhile, the technologies we use and how we interact with them changed leading to the issue that the tried and tested methods of evaluating usability no longer fully capture what happens in human-computer interactions [Rocha et al., 2017, Carvalho et al., 2018]. This problem is exacerbated for ubiquitous and pervasive computer (UbiComp) systems which leads researchers to expand on the traditional concept of what makes UbiComp systems usable [Rocha et al., 2017, Carvalho et al., 2018, Crabtree and Rodden, 2009]. UbiComp systems aim to make interactions less explicit and demanding by fading into the environment and embedding into the desired activities and workflows. This leads to an increasing importance of the environment and the established workflow on interaction outcomes as well as where the human attention and focus lie during the interaction; all of which are not adequately reflected by the traditional means to evaluate interactions requiring a rethinking and amending of methods [Rocha et al., 2017, Carvalho et al., 2018, Crabtree and Rodden, 2009].

Many researchers try to bridge this gap by developing and using novel methods to determine cognitive load to evaluate UbiComp system interactions [Haapalainen et al., 2010, Chen et al., 2011, Saha et al., 2018, Pillai et al., 2022, Fridman et al., 2018, Arshad et al., 2013, Yin et al., 2007, Gavas et al., 2017, Fujiwara and Suzuki, 2020, Li and De Cock, 2020]. This is most often accompanied by the goal of its automated evaluation for use in adaptive automation which favors quantitative analysis with objective measures. However, many different understandings and definitions of cognitive load serve as the conceptual underpinnings of these evaluation studies. Furthermore, there is little consensus on

methodological accuracy, which leads to hardly comparable results.[1].

Therefore, I explore the use of cognitive load for evaluation and assessment of UbiComp systems in this work. To this end, I framed the following research question[2]:

- How is cognitive load typically measured in HCI? How can it be determined?
- How is and can it be used for the evaluation (and development) of UbiComp systems?
- Which additional challenges arise for its estimation introducing the ubiquitous and pervasive characteristics of UbiComp systems?

In the related work chapter 2, I explore the current means of cognitive load estimation for use in human-computer interaction and why it is deemed important and useful. There, the current means of estimation are explored and its limitations discussed, answering most of the first research question. Additionally, challenges related to the evaluation of UbiComp systems are also explored answering the third research question.

Based on the findings of the literature review, I test alternative methods of cognitive load estimation in a case study with representative target users using a UbiComp system prototype developed by project partners at Profactor, covered in chapter 3. The prototype used is a Spatial Augmented Reality system providing users with task and policy related information via projections in a factory lab setting. The methods chosen for the user study are

- NASA-TLX [Hart and Staveland, 1988],
- secondary task completion,
- behavioural analysis
- thinking-aloud protocol [Rooden, 1998] and
- cognitive walkthrough.

The specific methods were chosen since similar methods have yielded promising results with non-pervasive technology [Chen et al., 2011] and they are not covered by works in the corpus of the related work chapter 2. Additionally, interviews were conducted to elicit if participants recognised and correctly identified the used in-situ projections or if they fulfilled the given task without using them (which is possible since the pervasive

[1] Both topics are covered in detail in chapter 2

[2] Before starting my research, I expected the first question to be the most trivial with the third question yielding the most interesting results, as I had trouble limiting potential causes for findings in previous work with UbiComp systems. As it turned out in my literature review (covered in the related work chapter 2), it was exactly the opposite.

technology only provides aid relevant to the task and is not required for completing it). The cognitive walkthrough was based on videos derived from the user study and conducted by HCI experts. Chapter 3 finally explores how the different alternative individual methods compare in terms of their performance.

In the discussion 4, I combine the findings from the individually tested methods combined with the literature review to answer and discuss all research questions with a focus on the first and second question, since these turned out the be the more pressing matter considering the current state of the use of cognitive load.

CHAPTER 2

Related Work

2.1 Why measure cognitive load?

In many studies, cognitive load has been identified and verified to be an important factor for human performance, especially during learning. Since then, it has become a widely used measure for ease of interaction of computer interfaces and has become an integral part for adaptive automation [Chen et al., 2011]. Considering these aspects, cognitive load could be of great interest in HCI research. The goal of measuring cognitive load for adaptive automation is to determine adequate times to interrupt users or to provide additional information without increasing the mental workload or breaking concentration [Chen et al., 2011, Fridman et al., 2018]. With this knowledge, interactions could be designed to be increasingly seamless providing functionality and information only when they are needed. This would keep interactions simple while maintaining a high level of functionality which is otherwise a common trade-off in user experience design [Chen et al., 2011, Fridman et al., 2018]. To achieve this however, the load would need to be determined real-time during the interaction without inhibiting it. How well current methods achieve this goal will be looked at later in this chapter (2.2).

Additionally, cognitive load is already used to help informing on the effectiveness of learning material [Duran et al., 2022, Kelleher and Hnin, 2019] and the ease of interaction [Hart and Staveland, 1988, Hart, 2006] as a usability measure. Beyond that, cognitive load could be used to evaluate technology adoption as I would argue that adoption is little else than learning how to operate new technology. The benefits of good usability and easy adoption are very intertwined and allow for earlier effective and satisfactory use of the technology in question. The dream of adaptive automation, its evaluation potential for learning and its use as a usability measure make clear why cognitive load has become of increasing interest to human-computer interaction research. Therefore, I would argue that incorporating accurate estimations of cognitive load into user studies when evaluating technology is of great benefit to user research.

To this end, I reviewed a corpus of works developing or testing methods to estimate cognitive load. The corpus was built using the following criteria:

- All works in the corpus had to be research publications and be written in English.
- The papers were collected solely from ACM and IEEE digital library to especially focus on the current strands of research within HCI due to my perceived focus on cognitive load by the UbiComp research community after some papers using cognitive load.
- The search terms were *cognitive load, cognitive load measure, cognitive load estimate* and *cognitive load estimation*.
- Then, I selected the works that did not want to use cognitive load within a study but instead developed, compared or tested methods to estimate cognitive load.
- In the end, I only selected works that introduced new sensors or entirely different ways of interpreting already covered sensors. Not much selection using this method was necessary, as not many works used already similar methods.

The resulting corpus will be analysed item by item in the following section.

2.2 Cognitive Load Estimation Right Now

In this section, the we answer the question of: how can cognitive load be measured or determined? There is a lot of work currently being done in this field with greatly varying measures that are used to try and gleam into the inner workings of working memory. In general, they can be summarized into two groups: self reporting and objective measures.

2.2.1 Self reporting

Self reporting measures rely on participants evaluating their condition after being confronted with the technology in question. Likert-Scales [Chen et al., 2011] or standardised questionnaires like the NASA-TLX [Hart and Staveland, 1988, Hart, 2006] are the most commonly used self-reporting methods to estimate cognitive load [Duran et al., 2022]. Self-reporting methods are often under scrutiny for being inherently subjective measures since they rely heavily on self-reflection and have to work with the fallibility of human self-perception. Since human perceptual and, more importantly, cognitive biases are well documented and researched, this critique is understandable. However, when trying to detect and replicate induced cognitive load, self-reporting methods are among the most accurate measures present in this corpus of methods [Chen et al., 2011]. Even in the early work of [Chen et al., 2011] in 2011 which compared the accuracy of three different objective measures prevalent at the time with self-reporting, self-reporting clearly came out on top in all classification tasks. The classification tasks varied in the number of

load levels that were induced and attempted to be replicated by classification of collected data. For all levels of load measured self-reporting performed best of all methods tested in the experiment (see 'SR' in figure 2.1).

2.2.2 Objective Measures

Even in this corpus alone, the use of objective measures used for the estimation of cognitive load has become vast and diverse furthered by the critique of self-reporting measures. With the availability of an increasing variety of sensors measuring bodily functions, they are used to try and estimate induced cognitive load by classifying their sensor values in different conditions using varying computational models. In this chapter, the different objective measures of the corpus will be reported and compared.

Starting again with [Chen et al., 2011], the authors used pupil diameter, blink number ('PD + BN', both measured with an eye-tracker), response time ('RT') and performance accuracy ('PA') in addition to a self-reporting scale ('SR') for the same tasks and trained computational models using each. The authors then compared their accuracy in classifying the induced levels of cognitive load. The load was induced using increasingly difficult addition tasks. The tasks were classified into five difficulty categories with increasing difficulties having more digits and carries. These five difficulties of tasks were then also grouped to form two-, three- and four-class classifications based on performance accuracy to test the classifiers for different amounts of classes. The authors used Gaussian Mixture Models (GMM) as classifiers to build their classification model. While pupil diameter combined with blink number as well as response time had comparably high accuracy across all levels, they came nowhere near to a simple single Likert scale (see figure 2.1).

Method	SR	RT	PA	PD + BN
5 classes (1-2-3-4-5)	66%	39%	29%	39%
4 classes (1,2-3-4-5)	75%	50%	35%	54%
3 classes (1,2,3-4-5)	85%	69%	51%	69%
2 classes (1,2,3-4,5)	90%	82%	68%	84%

Table 2. Classification accuracy using a GMMs approach. Levels in brackets with ',' between are grouped into one class and levels with '-' between are distinct classes.

Figure 2.1: Table of Methods Accuracy as Reported by [Chen et al., 2011].

[Saha et al., 2018] measured electroencephalography (EEG) signals taken from participants while performing second language English reading comprehension tasks. The levels of task difficulty (and the resulting expected cognitive load) were defined using Kincaid readability tests. They used a deep learning pipeline of stacked denoising autoencoder followed by a multilayer perceptron followed by a long short term memory

and another multilayer perceptron to classify their EEG data. This results in a good accuracy of 86.33% for their three classes of cognitive load which would be on-par with the self-reporting of [Chen et al., 2011]. As they only used EEG-data from a total of four participants, their results can be considered as debatable.

[Fridman et al., 2018] use a camera to film participants driving cars. The camera is used to extract eye-tracking features unobtrusively and in a real-world setting. The features (similar to features extracted by eye-trackers) were then classified by a hidden Markov model approach and a 3D-CNN separately. The HMM achieved an average accuracy of 77.7% and the 3D-CNN an average accuracy of 86.1%. The cognitive load was induced using the standardised n-back task. The n-back task requires participants to recite numbers given to them via audio. In 0-back, participants are asked to recite the latest number given to them. In 1-back, the number before that and in 2-back one further back et cetera. To induce three levels of load, the 0-back, 1-back and 2-back variants were used.

[Arshad et al., 2013] tried to use mouse activity during the interaction as an indication to classify for cognitive load. The authors attributed different lengths of pauses to different behavioural patterns. The two levels of load were discerned between simply completing a primary task compared to completing the same task while being interrupted by a similar sub-task. Both tasks could be achieved using a mouse and a screen. T-tests showed that these behavioural patterns differed significantly in frequency between their two load classes low and high load. But since their distributions have a lot of overlap they did not try to classify load using their features themselves. The large overlap would make mouse behaviour a bad sole classifier since classes would not be consistently correctly decidable without a distinguishing feature in the overlap region.

[Murata and Suzuki, 2015] try to predict cognitive load levels by measuring cerebral blood flow. There was no classification in levels as with previous studies. Instead, the authors measured the flow of participants while resting and during calculation tasks. Afterwards, they combined spectrum analysis and correlation of the blood flow and their load specification resulting in a correlation coefficient over time which was computed using piece-wise fast Fourier transforms. The authors report significant differences between rest and calculation blood flows. However, they only included subtractions in the study, since they were the only tasks showing any difference in blood flow in previous tests. Additionally, the required equipment was very intrusive needing to be strapped to the head tightly with skin contact.

[Haapalainen et al., 2010] compared the classification accuracy of many different measures. Heat flux, electrocardiograms median absolute deviation, galvanic skin response, heart rate, pupil diameter and electroencephalogram data captured during cognitive load inducing tasks were used for classification. The load was induced with six different types of tasks which each had high and low difficulty variants. The authors expected that cognitive load might physically manifest differently depending on task difficulty and the cognitive capabilities used in the task. This was their reason for including various ways to induce load in addition to a plethora of measures. According to their findings,

participants had different features that performed best for classification of their cognitive load levels. Heat flux and electrocardiograms median absolute deviation, however, were the best classifiers for a clear majority of participants. Their average classification accuracy among all participants were 76.1% (heat flux) and 71.4% (electrocardiograms) respectively. Combined, they reached an accuracy of 81.1%. This is especially impressive considering that their data on cognitive load levels are more noisy and transferable than other works of this corpus since the authors used many different tasks to establish them.

[Gavas et al., 2017] wanted to classify load using frequency domain analysis of pupil size variation. They used a mental addition task to induce load. Low load tasks were additions with numbers from 0-5 and high load tasks used numbers from 6-19, but not 10, 11 and 15. In total, ten numbers had to be added in sequence with three seconds time between each number being displayed. To ensure consistent data, participants were forced to keep their eyes open during calculation. To estimate cognitive load, they used multiple metrics gathered from the eye-tracker. Their proposed measure of cognitive load is a function of pupil size pulses of varying frequencies. A frequency analysis of pulses with varying magnitude is used to form bins for a given measurement period. The proposed measure of cognitive load is then calculated by multiplying the mean frequency by its corresponding power for each trial.

In their analysis, they compared previously used measures derived from eye-trackers to their new definition of load. The other measures are *percentage change in pupil diameter* and two measures from a visual field analysis *perimeter-area ratio* and *form factor*. For their trials, their new definition of cognitive load performed best and was able to identify between their induced low and high loads for all trials. However, their data shows very minuscule differences between low and high load induction for many trials. This either means that the loads did not differ much for most participants or that their measure is very prone to noise and relies on great measurement accuracy. The cognitive loads measure proposed by the authors also varied greatly for individual participants between trials leading to inaccuracies when deferring a load classification from their measure. While load for their high load tasks was on average higher than for their low load tasks, the difference per participants was often very minuscule making classification inaccurate. While outperforming the other measures used to compare their proposed cognitive load, the resulting accuracy of 71.3% is not great compared to the best performing methods of this corpus [Fridman et al., 2018, Chen et al., 2011, Yin et al., 2007].

[Li and De Cock, 2020] use measures derived from a wristband monitor to infer cognitive load levels of participants using different machine learning methods. The data used from the wristband monitor are *galvanic skin response*, *heart rate*, *rr intervals* and *skin temperature*. The data was provided as part of the challenge in the UbiTtention2020 dataset and was collected from 23 users (18 for training and 5 for testing). The authors had no control over how the data was collected but also did not report the means used to induce cognitive load on participants. Without going too much into detail, their approaches while sophisticated yielded sobering results with their highest accuracy resulting in 63% accuracy for classifying between no load and load. This was done using

a logistic regression model of all four measures and performed not much better than random guessing. The results are especially meager when compared to best performing methods of this corpus [Fridman et al., 2018, Chen et al., 2011, Yin et al., 2007], with [Chen et al., 2011] and [Yin et al., 2007] using less sophisticated and energy intensive methods.

[Yin et al., 2007] use different speech features to discern cognitive load levels. Self reporting techniques were used to verify their three levels of induced load. There is one major asterisk for their data: they expected adults over 18 to have similar levels of reading comprehension without confirming this in any way. However, this does not seem to have impacted their results much, but maybe they also sampled from a homogeneous group. It is not clear from the paper. Their used measures are *utterances in the frequency domain*, the *Mel-Frequency Cepstral Coefficients*, the prosodic features of *pitch* representing tone and *intensity* to indicate emphasis. These were modelled using a *Gaussian Mixture Model* and *Cepstral Mean Subtraction* as well as *Feature Warping* were used to reduce channel mismatch. Their resulting best model was able to reach 71% accuracy in classifying between three different levels of load which is one of the best results reported in this section and is only beaten by the camera information while driving of [Fridman et al., 2018] and the self-reporting of [Chen et al., 2011].

[Kelleher and Hnin, 2019] tried to adaptively predict the cognitive load of code puzzles based on puzzles previously completed by the same participant. To achieve this, they collect biometric, behavioural and self-reported features to build Random Forest Classifiers. However, since one single classifier for all types of puzzles had unacceptable performance, they instead created a classifier for each type of code puzzle. Afterwards, they reported the weight of each measure for each classifier and separated them into germane, intrinsic and extraneous load to explain their findings based on Cognitive Load Theory. They aim to use this predicted load to provide appropriate learning material to induce the optimal cognitive load to facilitate learning. This study was mainly done with middle school aged girls since most of the learning camps in which the study was conducted were hosted for girls only. The established ground truth for load was self-reported from participants by a 9-point likert scale which was complemented by an additional level of load for unfinished puzzles. From this base, other features were evaluated. Their resulting predictive pairing comparison accuracy (so which of two compared puzzles will induce higher load) ranged between 71% and 79% for the different types of code puzzles.

In this section, we have seen many different methods of using objective measures to estimate a previously induced cognitive load. And then, in most cases, a classification accuracy of the induced load based on these measures is reported. Among all methods, the self-reporting of [Chen et al., 2011] (2011), the CNN of videos of pupils of [Fridman et al., 2018] (2018) and the manual speech feature extraction of [Yin et al., 2007] (2007) performed best. [Chen et al., 2011] is the only work in the corpus to test multiple numbers of load level classes within one work.

However, the authors in this corpus use widely different means to induce cognitive load with widely differing participant populations and means to estimate load. So, what is it

that was measured and compared here? What are these increasing heights of accuracy the authors are chasing?

2.2.3 Accuracy and what is measured

To measure the success of their methods, the researchers had participants complete tasks aimed to induce a specific amount of cognitive load and recorded them using different measures. Afterwards, they tried to reconstruct these induced levels of load using their collected data. Finally, they compare how accurate their reconstructions of load levels were by reporting a percentage accuracy and sometimes a confusion matrix. This approach, however, assumes that the load that is *aimed* to be induced by the employed tasks was actually induced in participants. The aimed levels of load of the tasks are taken as an absolute ground truth for further analysis.

Apart from [Fridman et al., 2018, Pillai et al., 2022], no study used standardised tasks which are confirmed to induce comparable levels of load per participant. And even the standardised tasks acknowledge that the total level of load varies highly between individuals and only relative load for each individual can be standardised (so, lower load inducing tasks for one individual can induce higher total load than high load tasks on a different individual but do very rarely for the same individual). The difference in levels are also highly individualistic. One cannot assume that the high load tasks induce a significantly higher load to the low load tasks simply because a standardised task is used. Both could be considered high or low total levels of load when considering the capabilities and state of the individual tested. So, even using standardised methods to induce load, we can only assume that load will be higher if we use a task inducing higher load but we cannot make confident claims on total load or when comparing load between individuals. Therefore, errors in classification might simply occur due to noisy data or minuscule differences in total load.

Considering that in the best case a standardised method of induction is used as a ground truth for quantitative analysis and model training, there can never be a single measure indicating load levels but only individual calibration. When not using standardised tests, it cannot even be ensured that the used tasks induce the desired comparability of load levels. Nonetheless, they are used as ground truths for most model training of the works in this corpus. For classification, one therefore does not know if the error happened in the estimation of the induced load or while inducing the load, especially, when not using standardised means to induce load. Therefore, comparing reported accuracies of classification is not as meaningful as one might assume.

As indicated earlier and also discussed in detail in the following section, cognitive load is highly individualistic. This means that the choice of participant sample heavily influences findings. However, this is done with greatly varying levels of detail throughout this corpus. [Saha et al., 2018] had only four participants proclaimed as “healthy males and females". [Yin et al., 2007] simply assumed that reading comprehension levels of adults

over 18 in a native language are the same or at least similar. This is a very strict claim to make and depends for instance heavily on levels of education.

Even though, prior knowledge and skill levels in the tested task are hard to control for, it makes no sense to simply hope levels are similar enough and be done with it. The result is that errors might arise due to high differences in the prior knowledge of participants and not having a way to account for these differences in the models and instead simply hoping that they are not impactful enough in skewing results. Since there is currently no exact way to determine how cognitive load manifests physically [Duran et al., 2022], it makes no sense to assume that it does so in the same way for every sample. Therefore, without carefully managing the participant sample, research measuring cognitive load might never yield the results necessary to understand the phenomenon.

To further put the findings into perspective, only one paper did their study outside of a lab setting. Cognitive load depends heavily on the current state of the individual in question as well as the state of their environment [Duran et al., 2022]. Additionally, much of the measurement equipment used in the above studies [Gavas et al., 2017, Fujiwara and Suzuki, 2020, Saha et al., 2018] cannot be used to evaluate outside a tightly controlled setting. This means that lab findings might highly differ from real world environments with many additional mentally taxing factors.

Still, not every work in the corpus using objective measures introduced the above mentioned sources of bias. [Fridman et al., 2018] is a positive example succeeding where many other papers mentioned fail. They use a standardised definition of both their tasks used to induce load and reference Cognitive Load Theory before applying their models. Their tests are done on driving participants which is one of the best studied fields for mental workload in general according to their sources and has the most standardised methods tested in similar environments making their data the most comparable within the corpus. Additionally, they use a controlled sample and practice the task ahead of time to reduce additional load it might otherwise induce while still learning it. They do not rely on tightly controlled environments and clean data for their analysis since it was already conducted with noisy data from a real world driving environment. Despite (or maybe because of) this lack of control and while using standardised comparable methods of load induction, they reach similar accuracy levels of the aforementioned self-reporting metrics accomplished in [Chen et al., 2011] which is impressive comparing them to all other objective measures from the corpus.

However, it is still unclear what we are comparing when we compare the reported accuracy. Is it how well models classify using their tested methods? Or is it how well they managed to induce the desired levels of load? Or is it instead how well they selected their sample? The works within this corpus seldom used the same tasks with n-back being the only one to be used by several. The authors worked with widely differing populations and used a plethora of measures and analysis methods, some more realistic for real world use than others. Even standardized tasks cannot guarantee the induction of similar load levels. So what are the comparable results? And why can present researchers using state-of-the-art

computational models barely compete with self-reporting methodology's accuracy which has been around forever?

These findings made me skeptical about self-proclaimed "objective" measures for cognitive load. I would not argue, that this means that our skepticism of self reporting methods should cease but instead that we should question the infallibility of "objective" measures. Even though their use of mathematics in their computational models might be flawless and the source of their data objective, many factors influence the findings before the models try to capture reality, leading us back to the cognitive biases mentioned above. Most results in the corpus are only published as accuracy and not put into perspective of what is actually tested using the methods at hand and not put into relation of what might additionally influence findings. For the mathematical methods to yield accurate answers, the underlying models need to be correct and represent reality as accurately as possible with current knowledge. Based on the above discussion, this is a hard claim to make. And after talking about all these questionable findings, let us discuss the elephant in the room: what are we measuring? Why are the methods used to establish the ground truth so different between works? What even is this *cognitive load*?

2.3 What is Cognitive Load?

Looking more closely at what the authors of the works in this corpus are doing when working with cognitive load, I realised that there is a glaring issue: there is no single accepted definition of cognitive load. Even worse, most papers referred to in the previous section do not acknowledge this fact. They simply cite an early work from the field before *Cognitive Load Theory* [Duran et al., 2022] was formulated and researched and define what cognitive load could be from their interpretation and continue working from there. Sometimes, their definition of load is simply an alias for mental workload or cognitive activity [Saha et al., 2018, Pillai et al., 2022, Haapalainen et al., 2010, Li and De Cock, 2020]. To avoid this mistake, I will reiterate on the current standings of Cognitive Load Theory. I will discuss what cognitive load is and what it is not and afterwards highlight its inappropriate use in the aforementioned works.

2.3.1 Cognitive Load Theory

Originating in the educational sciences, Cognitive Load Theory is a theoretical framework that describes the influence of the human cognitive architecture on the learning process [Duran et al., 2022]. *Cognitive load* is described as the amount of working memory used by an activity. Based on cognitive load research, a plethora of effects on learning have been found resulting in pedagogical recommendations. In Cognitive Load Theory, learning is seen as the forming of knowledge in long-term memory which is seen as an unlimited capacity. For the knowledge to be formed, however, it needs to be worked through and constructed in working memory. Working memory is known to be very limited in capacity and duration (even though it can be improved by training) [Duran et al., 2022].

How to overcome this limitation of the working memory effectively by using adequate instructional design is at the core of Cognitive Load Theory.

To describe the interaction of learning material and working memory, two concepts are used in Cognitive Load Theory: *element interactivity* and *schemata*. Element interactivity refers to how many elements and connections between elements need to be considered at the same time in working memory to process a situation. Interactions between elements arise when they need to be "compared, contrasted, integrated or otherwise consciously processed together" [Duran et al., 2022](p.40:3). So when the amount of interactions required increases, the used working memory increases resulting in a higher load for the time of learning. When the load becomes too much, the learning is not successful. When the learning is successful, the knowledge is packed into a schema for long-term memory storage. Schemata are domain-specific bundles of knowledge which can be retrieved from long-term memory and processed as a single bigger element in working memory. This reduces interactivity which decreases the required cognitive load to work with the knowledge while increasing the total cognitive capacity. Therefore, previous knowledge of a learner not only heavily influences the effectiveness of learning material for that learner but also the cognitive load for a given activity [Duran et al., 2022].

In Cognitive Load Theory, there are two types of cognitive load which are undisputed namely *intrinsic load* and *extraneous load*. They are both cognitive load which are only differentiated for the purpose of analysis. Intrinsic load is the load which is minimally required to allow a specific learning to happen based on specific prior knowledge. Extrinsic load is load which is not necessary for learning and stems from e.g., instructional activities and learning materials. It can therefore be reduced safely and when reduced can lead to easier learning, since more working memory is available to tackle the intrinsic load [Duran et al., 2022].

These two different terms can be used to explain various effects identified by cognitive load research. For example, the worked-example effect compares learners solving a problem to learners studying an already solved example. Learners solving the problem have a harder time because additionally to studying the example, they need to consider how they can maneuver within the problem space to get to the desired solution. Therefore, the extrinsic load while solving the problem is higher than while studying the example [Duran et al., 2022]. Even when considering that a learner might have increased intrinsic load when learning an example and maneuvering the problem space at the same time, the compound effect of element interactivity would add extraneous load. Either way, the extraneous load is higher. Extraneous load is not always bad. In cases, where cognitive overload is not a concern extraneous load does not need to be minimised and can be increased to further some other goal (e.g., increasing motivation). When the intrinsic load is high due to a combination of low prior knowledge and high complexity of the content, reducing extraneous load is good [Duran et al., 2022].

These are the concepts consistent through Cognitive Load Theory. However, given its history, big influence and lack of definitive evidence, there is no single or unchanging Cognitive Load Theory. Currently, there are two main strands of theory existing in

parallel. One is a reworked version backed by the original authors of Cognitive Load Theory which reduced the theory to its simplest form yet and is proclaimed as the current version. The other, older theory, promoted in the time between 1998 and 2010, is still currently widely used, yet no longer supported by its original authors [Duran et al., 2022].

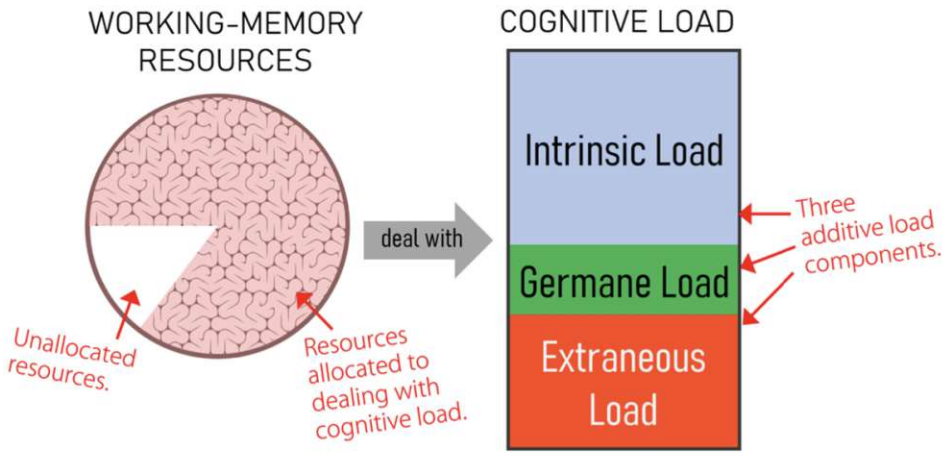


Fig. 1. Old CLT, as interpreted by the present authors. In the abstract scenario depicted, most of a learner's working-memory capacity is devoted to dealing with the cognitive load imposed by a learning activity. (In this scenario, there is no cognitive overload and some resources are left unallocated).

Figure 2.2: Cognitive Load Theory Model of the Working Memory While Learning including Germane Load interpreted by [Duran et al., 2022].

In the older strands of the theory a third type of cognitive load was differentiated called *germane load*. Germane load was understood as good cognitive load that should be maximised to facilitate learning. It goes beyond managing intrinsic load and is explicitly needed to make learning happen (see figure 2.2). The theory claimed in the case that intrinsic and/or extraneous load were too high, no capacity was left for germane load and therefore no learning could happen. Additionally, it was also used to explain the effect of motivation on learning: if a learner showed no motivation, the germane load induced by the learning activity was not high enough for motivation to arise. It was also claimed by some scholars that germane load was a result of additional learning aspects like self-explanation and reflection of learning material [Duran et al., 2022].

These claims stand in contrast to empirical findings. Load increases were never measured to yield increased learning success. On the contrary, many studies showed an increase in learning results when load was reduced [Duran et al., 2022]. This directly contradicts a typical claim of the original Cognitive Load Theory that decreases in extraneous load should lead to equal increases in germane load which would facilitate learning. Furthermore, unfalsifyable post-hoc claims would be made using germane load. If load would be reduced and it lead to an increase in learning success, one could claim that germane load was decreased. However, if the learning success increased extraneous load was decreased. The opposite argument can be made vice versa for load increases. This

can in no way be proven or contradicted making these arguments nearly meaningless. However the effect on learning and however the measured cognitive load, germane and extraneous load could be used to explain the results with no way of detecting which type of load was present. Additionally, many authors found germane load unnecessary to explain results and it could not be used to explain many detected effects. Therefore, the theory was revised in 2010 [Duran et al., 2022].

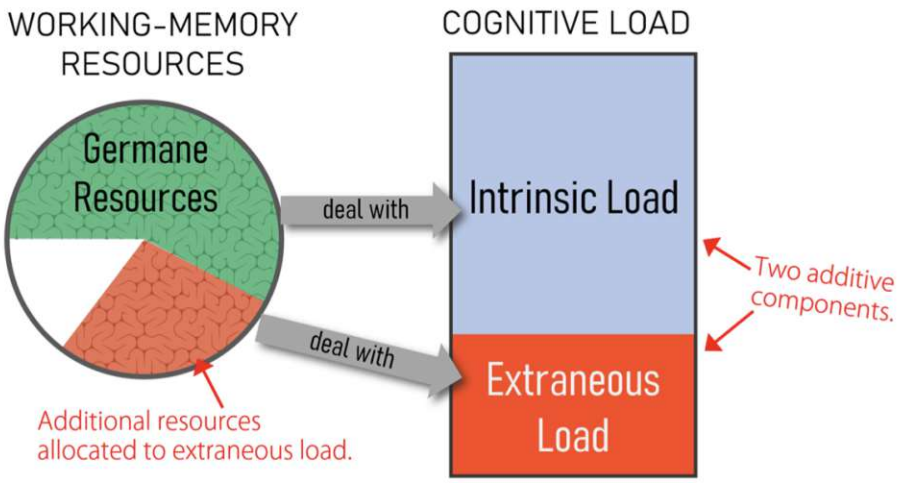


Fig. 2. New CLT, as interpreted by the present authors; cf. Old CLT in Figure 1. In this scenario, extraneous load requires the learner to devote some working-memory resources, but the learner is also able to devote germane resources to intrinsic load, i.e., learning. (This figure, too, depicts a scenario without cognitive overload.)

Figure 2.3: Cognitive Load Theory Model of the Working Memory While Learning including Germane Resources interpreted by [Duran et al., 2022].

The revised theory cuts down on the differentiation of different forms of cognitive load. Intrinsic and extraneous load are defined strictly by using element interactivity and all cognitive load is composed of the two. Any load that is not extraneous for a learning activity is intrinsic. Self-explanation and reflection are simply defined as additional learning goals and therefore increase intrinsic load. Working memory is separated into resources dealing with intrinsic load (germane resources) and other resources either unused or used to handle extraneous load (see figure 2.3). Motivation is also extracted from the model. A motivated learner is simply assumed, since its effects could not be explained by any version of the model [Duran et al., 2022].

So, there are two main types of Cognitive Load Theory currently alive and in discussion with the newer trying to solve or bypass many critiques of the old which could not be empirically confirmed. The newer strand is also backed by the original author of Cognitive Load Theory, John Sweller. But "as things stand, it is contentious which constructs compose cognitive load, how best to interpret cognitive load measurements, and how to phrase hypotheses that involve relationships between load components, overall load, and

learning outcomes" [Duran et al., 2022](p. 40:6).

To summarize, Cognitive Load Theory tries to model the inner workings of working memory of a learner during a learning activity. The proclaimed current version assumes motivated learners. Cognitive load is used as a term to describe the working memory resources a learning activity requires. These depend heavily on previous knowledge of the learner for a given activity. If a learning activity requires more resources than are currently available, cognitive overload occurs and no learning can happen. It also separates the cognitive load into two categories: load that is necessary for the learning process (intrinsic) and everything else (extraneous).

2.3.2 Cognitive Load Estimation without Cognitive Load Theory?

From my reviewing perspective, it is unclear if the works in this corpus imagine cognitive load according to theory. [Li and De Cock, 2020] cite no work of Cognitive Load Theory. They instead cite a source which claims that cognitive load is a confirmed physiological phenomenon and should therefore be aimed to be measured without further elaboration on what it is they are measuring. [Saha et al., 2018, Pillai et al., 2022, Haapalainen et al., 2010] cite papers that predate any Cognitive Load Theory. They use works discussing mental workload to define their own concept of cognitive load which in turn they proceed to “measure". And whatever findings might arise from the research just have to be cognitive load which they previously defined without any foundation in Cognitive Load Theory.

[Arshad et al., 2013, Kelleher and Hnin, 2019, Chen et al., 2011, Fridman et al., 2018, Gavas et al., 2017, Yin et al., 2007, Fujiwara and Suzuki, 2020] all cite a source of the original version of cognitive load. They refer to the three way split of load into intrinsic, extraneous and germane load. However, the only work in the corpus to analyse their findings using the before cited Cognitive Load Theory is [Kelleher and Hnin, 2019]. They categorise their measures into the three categories of load and also report the individual impact of measures on their cognitive load predictions. The rest simply reports an accuracy of their reproduction of their methods to induce load and simply claims that this must be cognitive load since their methods to induce cognitive load were mentally taxing on varying levels (some to greater success than others). For context, the works of [Chen et al., 2011, Yin et al., 2007] are too old to have used the new strand of Cognitive Load Theory.

Referring back to Cognitive Load Theory, however, it is very unclear what they are trying to measure. Most of the tasks used to “induce load" are not really learning experiences. The only learning experience measured is in [Kelleher and Hnin, 2019] of mostly teenage girls working with programming puzzles. All other measure studies in the corpus use a method to tax the resources of working memory which they are in turn trying to measure. They identify one method which best reproduces their presupposed definition of a hard(er) task without acknowledging possible influences on their measures and then report an accuracy without interpreting where the errors might have occurred. And then,

they call it cognitive load because cognitive load seems like a self-explanatory term (it is the load on cognitive resources, after all) and sometimes even find a theory with the same name to cite which is discarded as soon as the experiment starts. And even if the theory is cited, it is not the less disputed reworked version but instead the older theory which could not explain many effects found in research of learning material.

As previously stated, only few works in the corpus use standardised methods to tax the working memory which is used to define the ground truth to define the cognitive resources required for the task of their participants. Some methods to tax the working memory used are confirmed to be taxing at appropriate levels, but they do not indicate anything regarding cognitive load. Cognitive load is a term used to describe the resources of the working memory necessary for successful learning given a specific individual and learning activity. In turn, this means without proper assessment or control of the previous knowledge and motivation of the individuals tested and the investigation of a learning activity, it makes little sense to call the resulting metric cognitive load and to refer to Cognitive Load Theory.

To summarize, the ground truth for analysis is defined without a means to control if it is actually accurate, then a model to reproduce this ground truth is created by various methods. Afterwards, how accurately the ground truth could be reproduced is reported without acknowledging that it is not clear if the errors arose in reproduction or at the definition of ground truth. In the end, a theory not fit to describe the phenomenon is used to name the results without an effort to interpret them through the lens of the name giving theory.

However well the use of computational models, I do not think that the results of the work in this corpus (apart from [Kelleher and Hnin, 2019]) can be used to confirm or refine Cognitive Load Theory even though they claim to be measuring the cognitive load the theory describes. In their work, they try to infer some state of the working memory without having any proven point of reference to interpret the findings from. There is only the knowledge that changes of the working memory's state should be detectable by observing bodily functions. Therefore, the works in the corpus intensively observe and analyse various measures of bodily functions. And changes in bodily functions they detect, when subjecting participants to mentally taxing tasks. Then, they try to form a computational model that is able to reproduce their predefined truth about the state of working memory of their participants. But they use methods that are not accurate enough to evoke the state of working memory that is used as ground truth for building the model.

In no work in the corpus, an effort was made to establish a level of minimal working memory resource use and cognitive overload (over-attribution of working memory) as points of reference to interpret their findings. In the best case, when standardised methods (e.g., n-back) are used, the only points of reference available are multiple levels of resource use of the working memory including the knowledge that in nearly all cases they can be accurately ordered by how taxing they are. But with no reference to a minimum or maximum, there is only ever an order that can be used to infer further. That

means, there is currently no way to confidently gauge a level of working memory use by using objective measures. Therefore, we cannot use objective measures to meaningfully analyse the states of working memory. Since cognitive load describes the use of working memory required by learning material for a motivated learner with a specific level of previous knowledge, cognitive load can therefore not be meaningfully analysed as well using objective measures.

Cognitive Load Theory itself also claims additional limitations which make the use of objective measures for analysis even more questionable if they are used in isolation. A fundamental assumption of the revised Cognitive Load Theory is that it only applies on a motivated learner, since the model including motivation could not explain many results. Controlling for motivation to correctly incorporate it into quantitative analysis seems near impossible. To correctly apply Cognitive Load Theory to isolated quantitative analysis, would therefore require being able to control for motivation in addition to the currently impossible accurate induction of working memory resource use.

However, this does not mean that researchers wanting to study cognitive load are out of options. Instead of relying solely on unproven post-positivist measures with little point of reference, one could instead turn to a mixed-methods approach by incorporating constructivist methods not relying on a measurable ground truth to complement the analysis. By interpreting potential findings through the lens of Cognitive Load Theory, it might be possible to still gather valuable insights. However, a live-assessment of the current state of working memory is not possible through constructivist methods. Therefore, the dream of interfaces adapting to the current state of working memory proclaimed by most works of this corpus is currently not achievable without previously being able to assess a ground truth of individual working memory limits, which are required to establish ground truths for objective measure analysis.

2.3.3 How to Instead Use Cognitive Load Theory in HCI

Since Cognitive Load Theory is a theory about learning, specifically how taxing the learning is on working memory, it can be used for one of the core interests of HCI: technology adoption. A big part of technology adoption is people learning how to interact with a given technology. The analysis of interactions with new technology in user studies is common practice and learnability is already of great interest there [Lewis and Sauro, 2009]. Currently, learnability is mostly evaluated based on performance of early use over time. The most common measures of performance are efficiency (how fast tasks are completed with the technology) and effectiveness (if the tasks are completed successfully with the technology). The development of efficiency and effectiveness from start of use to their plateauing is then often called learnability. An interaction with good learnability reaches fast plateauing. Good learnability alone, however, does not suffice to indicate its usability e.g., an interaction with fast plateauing but a remaining error rate of 30% is not desirable.

The current methods to evaluate learnability are easy to implement requiring only

behavioural data in time on task and successful completions, often complemented by the System Usability Scale [Brooke, 1995] using its factor structure [Lewis and Sauro, 2009] which allows to not only gather data on satisfaction (how well users liked to interact with the system and felt confident in their ability) but additionally on learnability itself. However, they do not yield much insight into why issues for learnability and interaction complexity arise. Analysing interactions through the lens of Cognitive Load Theory could therefore be of great interest to complement the findings of current measures of usability with an interpretation based on element interactivity and working memory usage.

For long-term adoption and usage impact estimation, the perspective of Cognitive Load Theory can be additionally helpful. Because the completion of a task with and without using the technology often differs widely, a clear definition of the required learning for both situations is helpful. With this exact specification of learning, the changes of required knowledge for the differently learned skill can be analysed in contrast to cognitive load while learning, especially intrinsic load. This different learning required for technology adoption, while possibly easier due to lower required intrinsic load, might not be desirable. An example for such an undesirable learning: learners might be unable to complete the task in a case of technology malfunction due to no longer understanding the requirements of the task without the technological assistance. For successful long-term adoption, users might additionally have to learn how to repair the technology increasing the overall learning required to complete the task [Bainbridge, 1983]. Even though a reduction of cognitive load might occur while learning the new, easier interaction, it might not be a prudent goal to blindly chase cognitive load reduction as the resulting learning might not be the intended learning and too much intrinsic load was reduced. Vice versa, a technology could be deemed unsuitable to a task in its current form, if the newly learned interaction, while providing great long-term benefits, requires too much cognitive load in its learning for the population who currently completes the task without the technological assistance. Keeping Cognitive Load Theory in mind, possible interventions could be to simplify the interaction by reducing element interactivity or by supplying the possible users with more schemata relevant to the interaction which both reduce cognitive load.

To further complement typical user research analysis with Cognitive Load Theory, one can try to separate extraneous load by identifying sources of distraction and unnecessary elements or redundant information. This can be done e.g., by analysing behavioural data, thinking-aloud [Rooden, 1998] data and interview data. These can in turn be removed and the interaction tested again, to confirm if in fact there was extraneous load inhibiting the learning of the interaction. Using e.g., A/B testing [Tullis and Albert, 2008], confirming Cognitive Load Theory based findings on interaction problems can be easily and cheaply included into modern technology development cycles.

In my opinion, speaking about cognitive load beyond technology adoption in HCI makes little sense. For continued use, it becomes increasingly harder to separate learning from doing and cognitive load is proven hardly enough to transfer its claims about learning to acquiring mastery with ever harder to define previous knowledge. When talking about

adaptive interfaces I would therefore not use the term cognitive load to describe their used data, especially after adoption. They would want to adapt their interactions based on working memory usage and other state of mind indications like states of deep work or flow. However, as argued in this section, this does not mean that there is no place for Cognitive Load Theory in HCI. Just not in the way many of the works in Ubiquitous Computing of the corpus claim.

2.4 How and why we measure “cognitive load"

As explored in the previous chapter, cognitive load is currently being used as an umbrella term for both the use of working memory resources and the cognitive load term of Cognitive Load Theory, which describes the working memory resources required to have successful learning of a specific learning activity and learner accordingly. This makes publications using the term at least confusing, if not misusing the name giving theory. In this work, there is a separation of the two to talk about their differences and why many of the works in the corpus were written with a questionable understanding of cognitive load and working memory resource use while claiming to measure the prior. I would encourage other authors to differentiate as well in accordance with Cognitive Load Theory, but there is little prospect of this changing in the near future. Therefore, one needs to be careful when reading about cognitive load since the theoretical background of the work has a high chance to be dubious.

Currently, there is no confident way of measuring cognitive load. As discussed in the section on Cognitive Load Theory (2.3.1), there are very hard to control factors required for cognitive load to be measurable. Additionally, in the work of this corpus, researchers are currently not really measuring cognitive load but instead are trying to measure the current state of working memory resource use. Due to their limited incorporation of Cognitive Load Theory into the study design and working with its limitations, the necessary precautions for accurate, reproducible findings to further the development of Cognitive Load Theory are not taken. Instead, claims of a (more or less) successful reproduction of a questionable ground truth are chased.

There are many valuable reasons for wanting real-time data on the state of the working memory in many fields. As claimed by most works of the corpus, adaptable interfaces are their main driver, which should adequately respond to the state of working memory and issue interruptions when appropriate in addition to providing the right information at the right time. These measures are currently questionable at best, due to the aforementioned classification based on an uncontrollable ground truth which is backed by a misunderstood Cognitive Load Theory (as discussed in 2.3.1). Instead of acknowledging this, the works in this corpus aim for ever greater accuracy of reproduction using ever more complex computational models without addressing the underlying issue of their arguments being baseless.

As the analysis in this work deems the use of real-time objective measures with the current means of establishing specific working memory resource use futile, I suggest an

alternative for the use of Cognitive Load Theory in HCI (see 2.3.3). As a theory of learning it is fit for the analysis of phenomena arising in technology adoption, since a major part of technology adoption is the learning of possible interactions. It should therefore be possible to use Cognitive Load Theory to complement current methods of evaluating the learnability (and in turn the usability) of technology to provide the theory to identify possible hindrances in the adoption process.

Learning is a complicated and broad subject and Cognitive Load Theory only a small part of it. Using it to hide limitations and incertitude while claiming to have found an accurate, objective measure of the current state of working memory, and therefore, the human mind can be described as hubris or, depending on the motivation, something else entirely [Frankfurt, 2005].

2.5 Measuring Usability of Ubiquitous or Pervasive Technology

As already pointed out by many authors before me, there are fundamental differences in how people interact with UbiComp systems compared to traditional computer technology. The main one being that in addition to interactions between users and the technology, interactions between users and the environment and the environment and the technology take place during the time of interaction [Carvalho et al., 2018]. Furthermore, the goal of UbiComp systems usually entails seamless integration into everyday tasks and activities and preferably becoming imperceptible to the user [Bezerra et al., 2014].

Based on this, [Carvalho et al., 2018, Rocha et al., 2017, Bezerra et al., 2014] formulated additional characteristics of UbiComp systems which should be incorporated into a usability analysis to gain a more holistic evaluation of interactions taking place during the use of the UbiComp system. While the proposed heuristics are likely not the be-all-and-end-all of evaluation of UbiComp systems but rather the beginning of further research of its refinement, it still provides a good starting point for UbiComp system usability evaluation. [de Souza Filho et al., 2020] demonstrated this by incorporating the heuristics proposed by [Rocha et al., 2017](see figure 2.4) into two different usability evaluations for a UbiComp system.

With this greater importance of interactions with the environment (both from the user and the system) and the goal of seamlessly embedding the technology into the environment, traditional user studies in lab setups cannot capture everything that occurs in an interaction [Crabtree and Rodden, 2009, Rocha et al., 2017]. In my literature search, I have found no work which diagnoses what exactly is lost and which was never captured in the first place when applying traditional user studies to UbiComp systems. From the works I have found that aim to improve usability evaluation for UbiComp systems [de Souza Filho et al., 2020, Rocha et al., 2017, Carvalho et al., 2018], there is no consensus on what is necessary or highly helpful to include as additional factors

ID	Name	Definition
UH1	Visibility of system status	The system should always provide feedback to the user in response to an interaction performed. This feedback should neither disrupt the user in his current activity nor overwhelm the user with information, but must exist in the form of a noticeable change in some of the interaction modalities of the interface.
UH2	Correspondence between the system and the real world	The system must speak the language of the user, with words, symbols, concepts and interactions familiar to the user, instead of being system-oriented. One must follow the conventions of the real world, making the information appear logical and natural and easily reaching the intended goal.
UH3	User control and freedom	The user must feel free to interact with the application or not. When the user wishes to interact with the system, must be in control, and at any time can abort a task or undo an operation and return to the previous state. When the application interacts with the user in a given context, the user should not feel obligated to respond to the interaction and should have the option to ignore it in order to keep the focus on their current activity. All of these actions must be clearly marked on the system and their visualization, if any, should maintain the standard throughout the application.
UH4	Consistency and standards	Application interfaces, data inputs, ways of interacting or adapting to the context, must be consistent and follow conventions and standards familiar to the user, so that the software can be understood, learned and used.
UH5	Error prevention	It is important to know the situations that cause most errors and modify the interfaces and interactions so that users do not make these mistakes when interacting with the application. In addition, the application must be able to keep its services and performance always available when used by one or more users, under specific or adverse conditions.
UH6	Recognition rather than recall	When there is a dialog or interaction available, the system should minimize the user's memory load, leaving objects, actions, and options available to at least one of the user's senses.
UH7	Flexibility and efficiency of use	The application should provide shortcuts to accelerate the interaction, in order to reduce the effort required to achieve the intended goal, especially for the advanced user. Completeness of functionality must be maintained when using shortcuts or not. In addition, the system must be flexible, giving the user the ability to customize settings according to their needs and experiences.
UH8	Aesthetics and minimalist design	Dialogues should contain only relevant and necessary information, neither more nor less. Each extra unit of information in a dialogue competes with relevant units of information. The sequence of interaction and access to components and functionalities should be available depending on the context, in a simple and natural way.
UH9	Help users recognize, diagnose and recover from errors	Error messages should be simple and expressed in clear language (without codes), accurately indicate the problem and constructively suggest a solution. In addition to texts, messages can be displayed in other formats available on mobile devices, such as image, audio, vibration. Error messages should guide the user with caution, without stress, so that the user does not stop using the system.
UH10	Help and documentation	Ideally, the application should be so easy to use (intuitive) that it does not need help or documentation. If necessary, the help should be easily accessible, centered on the user's current activity. Help guidelines should be simple and objective.
UH11	Mobility and devices	Ubiquitous applications should maintain their functionality with the physical displacement of the user and on devices with different capacities. Aspects such as wireless networking, device connection, screen size, limited hardware capacity and power capacity are factors what the application needs to take into account to adjust during use without causing inconvenience to the user.
UH12	Privacy and safety	The application must be able to keep the information saved and protected, so that there is no risk of damage in a context of specific use. Information must be transported and stored securely, as well as the application's access controls.
UH13	Invisibility and transparency	The system must be able to hide computational components so users do not worry about them. Interactions must take place through natural interfaces.
UH14	Context awareness and adaptive interfaces	The ubiquitous application should react according to the context information that the user encounters. Interfaces must adapt to these contexts and bring only relevant information in a way that facilitates the use of the system.
UH15	Sensors and data input	It must be checked whether data input, either by the user or captured from the sensor, is being effective and happening naturally to the user. The application should operate correctly in the presence of invalid inputs or stressful environmental conditions.

Figure 2.4: A table of usability heuristics for UbiComp systems proposed by [Rocha et al., 2017] as interpreted and used by [de Souza Filho et al., 2020].

and how they are to be considered. The only consensus is that the previous ways of determining usability are not suited to capture the increasing complexity of interactions.

[Zilz, 2011] aim to capture some of this complexity in an easily applicable way using a virtual environment. By simulating a virtual UbiComp environment including its users, the authors want to avoid the cost and difficulty of real-life tests with real participants in real environments. Instead, they try to identify usability problems using expert evaluations of their virtual UbiComp system. While acknowledging that it is not a substitute for a real interaction in the environment of use with target users, the authors are confident that in using their method many usability issues can still be identified, especially issues that would not be as easily detected in typical usability evaluations in lab settings, while keeping costs at low.[1]

[1]The user study in this work was also not conducted in the environment the technology aims to be embedded in, but instead in the factory lab of Profactor. Since the tested UbiComp system was built as a prove of concept in its development stage, it seemed a sensible choice to keep costs low for our partners at Profactor. While still not a real environment and an artificial setup, a factory lab has similar looks and policies to factories even though no live production is taking place. More details are covered in chapter 3.

Thinking about cognitive load in UbiComp systems similar issues arise. A naive measure simply aiming to extract the level of mental resources used during the interaction cannot determine which part of the interaction is responsible for the increase measured. Therefore, an interpretation through the lens of Cognitive Load Theory is essentially required to identify its sources. Since there are multiple interactions taking place at the same time during a simple UbiComp interaction, element interactivity is of even greater concern. With the goals of seamless integration and high levels of interaction, it makes sense to aim for low cognitive load for the overall interaction. Otherwise, we run the risk of cognitive load being too high for the interaction to be easily learned and in turn not adopted quickly into the workflows the UbiComp system aims to be embedded in.

In summary, currently, there is no means to confidently interpret cognitive load measures which requires the development and/or testing of methods for its interpretation. Therefore it is complicated enough to interpret cognitive load for heavily controlled study setups. However, cognitive load is of especially great interest in setups which cannot easily be controlled e.g., real-world settings with simultaneous interactions or UbiComp system environments. This leads to the additional problems outlined above. Nonetheless, this work aims to provide some form of progress in this difficult environment with a case study by testing different methods to estimate cognitive load interpreted through the lens of Cognitive Load Theory in a semi-controlled, semi-realistic setting with representative target users of a UbiComp system.

CHAPTER 3

Case Study: Cognitive Load of UbiComp System in an Industrial Setting

The case study was conducted as part of a usability evaluation of a spatial augmented reality (SAR) system developed by our project partners at Profactor which also provided the location and setup for the user study. The tested technology was solely developed by Profactor. The usability study was planned by Rafael Vrecar, Astrid Weiss and myself and conducted by Rafael Vrecar and myself. The data evaluation was mainly done by myself. The technology combined with its evaluation are a separately published work from which figures and data were reused [Wedral et al., 2023][1].

Instead of relying on a sole objective measure during task completion, I opted to use a plethora of methods, some of which only tangentially have to do with cognitive load (and aid with other usability metrics evaluated in the study) to have more starting points for analysis for the cognitive load. How well they aided in determining cognitive load for the study and how I interpret their usage now with a deeper knowledge on Cognitive Load Theory is covered in this chapter.

3.1 Design and Description of Evaluated Artefacts

To start this chapter, however, a description of the evaluated artefacts is required.

[1] Even though at the time I did not yet engage myself as deeply into Cognitive Load Theory and did not grasp the implications of the current state of cognitive load estimation practice on my used methods, the methods used within the case study comply surprisingly well with Cognitive Load Theory compared to many methods discussed in earlier chapters (see 2.3.2).

(a) Ergonomic Notification

(b) Collaborative Robot Safety Zone Awareness

(c) Manual Assembly Assistance

Figure 3.1: 3D sketches of the study use cases for the SAR system developed by Profactor [Wedral et al., 2023].

The aim of the SAR system was to provide users with live feedback on their actions in their working environment without requiring a media breach to screens to access the feedback information. Our partners at Profactor developed three different SAR systems supporting three different use cases all prevalent in an industrial production setting. The *Manual Assembly Assistance* (figure 3.1c) avoids the media breach by directly projecting the location of parts' placement onto the work piece. For this lab setup, a simpler proof of concept setup was chosen, where workers had to peg in clamps on a board. The SAR system highlights where the part is to be placed and if the parts are placed correctly. Locations where clamps should be placed were indicated by blue circles. Locations of correctly placed clamps were indicated with green squares and locations of incorrectly placed clamps (or other occlusions of holes) were marked with red Xs.

The second system, called *Collaborative Robot Safety Zone Awareness*, was developed as a means to dynamically and responsively communicate the safety zone around a collaborative robot (Cobot). Cobots are robots especially designed to collaborate in a workspace together with human workers. To pose less of a danger, they have to operate on lower speeds during this interactions. ISO/TS 150662 defines spaces around robots with according different robot speeds and resulting dangers for humans. To communicate this safety zone, the SAR system projects an area around the cobot as depicted in figure 3.1b. This area turns red when a human enters its space. In a real factory application it could for instance also stop or slow the robots movement. The technology was developed with the idea of alerting workers of overhead dangers such as cranes and autonomously moving robots entering their spaces (or vice-versa). Compared to standard safety measures, SAR safety measures would only need to encompass currently active dangers and not all possible dangers for safety zone depiction.

The third system, Ergonomic Notification, aims to mitigate the risk factors to occupational safety and health of ergonomic issues in the workplace. The SAR system was developed to communicate detected risks to workers the moment the ergonomic issues arise. However, since the detection of ergonomic issues was not fully implemented by Profactor, the SAR system was controlled wizard-of-oz style based on skeleton extraction of the worker. As depicted in figure 3.1a, the projection turns green when a user lifts ergonomically in its space and pulsates red when a user has ergonomic issues during lifting. Otherwise, it

simply communicates its presence as a white circle. The triangle and rectangle were applied using tape and are only present to mark different locations for the test setup.

3.2 Study Design and Structure

In addition to evaluating classic usability metrics such as effectiveness, efficiency and satisfaction, the plan for the study was to use a mixed methods approach to pin-point metrics that are harder to meaningfully interpret using single measures like learnability and the aforementioned cognitive load.

The research questions for this study were:

1. How does the use of the defined projector assistance systems affect efficiency, effectiveness, and satisfaction?
2. How is the usage affecting cognitive load for task completion?
3. How effective is the learnability of the projector technology?

The lab study setup was designed in a way to incorporate the usage of all three SAR systems into a single task scenario. To gather comparable quantitative data, we comprised the task of multiple iterations of the same interaction. To make the interaction meaningful to the participants, we presented it as one overarching assignment. The most important reason for the repeated trials of single participants was to gather a naive measure of learnability using the changes in task completion time over the trials. As cognitive load stems from a theory on learning, learnability should be usable as a pillar to gauge cognitive load, since the load of interaction should decrease after the initial learning phase. During conceptualisation of the study, the focus was put on cognitive load and how the technology usage affected it for accomplishing the task which was especially important to the project partners at Profactor. [2]

Additionally, the SAR systems where not active in every step of the procedure. The tasks were setup in a way that the participants could complete them without interacting with the technology, as would be possible in their real work environment. This was done with the aim to gauge how useful the representative target participants perceive the SAR systems, even in early interaction. Instead of having one phase of the experiment with the SAR system active and one with the system deactivated, the projections switched between active and inactive at fixed repetitions to allow the observation of behavioural changes. The goal was to detect changes in policy compliance and task performance not only between on and off states but especially in off states after having interacted with the technology. If seen it could indicate that the positive behaviour might persist even after deactivation or disuse of the SAR system. The time with the technology active,

[2] This was before my deep dive into the topic and therefore my learnings were not incorporated into the study design.

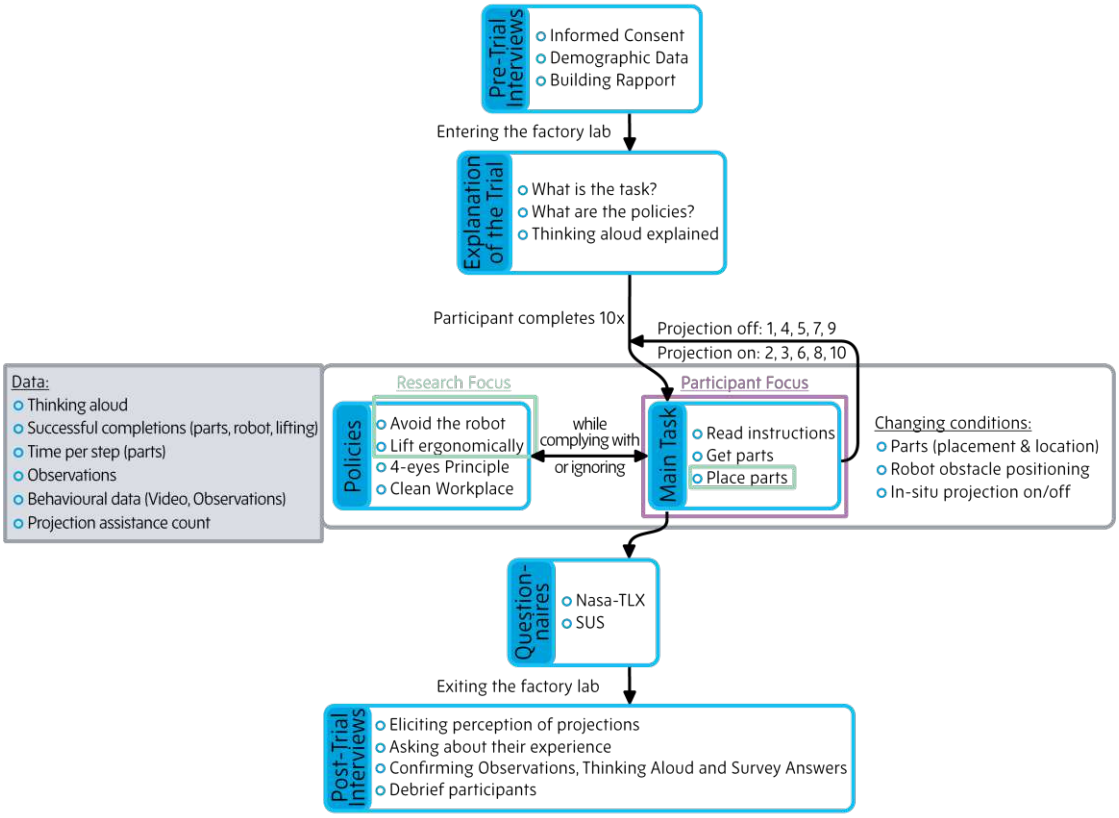


Figure 3.2: Shows the flow of the experiment procedure. It details which methods were used and data was gathered in which step. The green region marks the research focus during the repeated trials with the technology. The lavender region marks where the participant focus was expected during the repeated trials.

however, was not long enough for participants to become accustomed to its aid and it was therefore only perceived as a nuisance that the system was not always active.

Figure 3.2 shows how the resulting experiment was structured and in which step which data was captured. The study was conducted with fifteen representative target users mainly from the automotive industry with varying degrees of previous experience in working with automation. They were recruited by Profactor from the local area surrounding Steyr, Upper Austria.

Data and Measures

In the next paragraphs, I will cover the decisions on how the experiment was setup and a brief overview which data was chosen why. A more detailed analysis and explanation of method choice is covered in their individual sections.

As mentioned before (and backed by my findings in the related work chapter 2), the goal

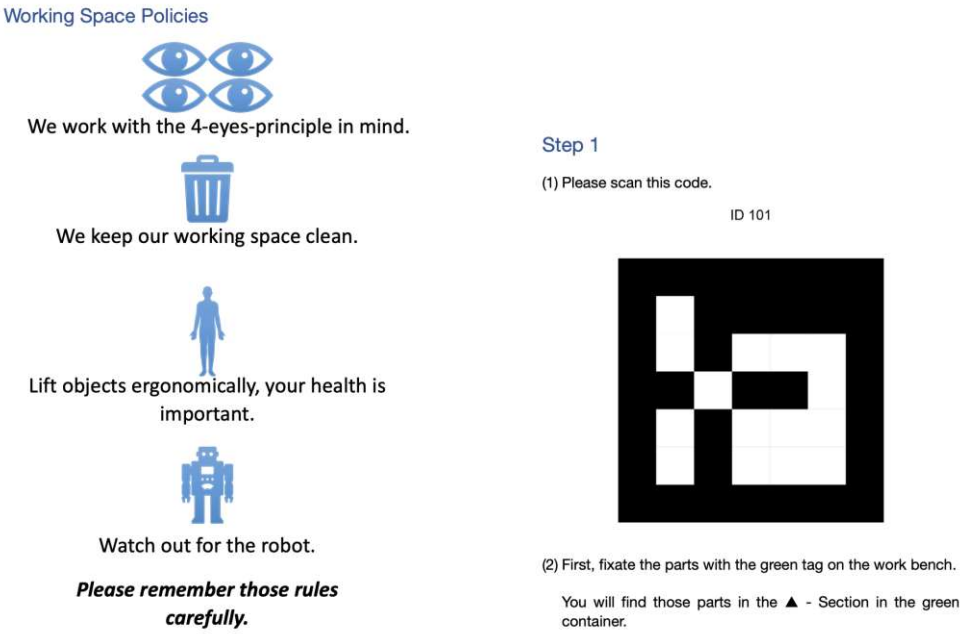


Figure 3.3: Policy and exemplary step description. The policy is translated from German

was not to rely on a single measure to determine cognitive load. Therefore, we used a mixed methods approach which incorporated various quantitative and qualitative data with the goal of gaining a comprehensive understanding of the matter. Most usability measures had one single source of quantitative data to evaluate it and the complex measures were complemented by insights from further qualitative techniques. The main measure of interest was cognitive load which has been extensively covered in the related work chapters. To evaluate cognitive load but also as an additional metric for usability, learnability was evaluated. The usability metrics covered by the ISO standard were also included [ISO 9241-11:2018, 2018]. The metrics of the standard entail:

- *effectiveness* - accuracy and completeness with which users achieve specified goals
- *efficiency* - resources used in relation to the results achieved
- *satisfaction* - extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations

The following data and collection methods were therefore selected for the study. In [square brackets] the type of data is depicted, followed by a description on point of collection. For which criteria it was used is displayed in (parentheses).

- [Behavioural] Time required for each step at the work bench (Efficiency)

- [Behavioural] Successful policy compliance and completion of tasks (Effectiveness)
- [Self-reporting] NASA-TLX Survey [Hart and Staveland, 1988] (Cognitive Load)
- [Self-reporting] SUS-Survey [Lewis and Sauro, 2009] (Usability and Learnability)
- [Self-reporting] Semi-structured interviews (Satisfaction, Learnability, and control whether the system is actually considered during the experiment, since it is not required for completion or compliance)
- [Self-reporting] Think-aloud-protocol [Rooden, 1998] [Tomitsch et al., 2018, p. 158] (Cognitive Load, a control for Learnability and control whether the system is actually considered during the experiment)
- [Behavioural] A count of how many times the SAR system was active including the current run (Cognitive Load)

Notes and observations made during the study and during the analysis of the video footage were used to complement or contradict findings of other measures when appropriate. However, they were not coded or analysed in-depth since it would break the scope of the work.

To have quantifiable and comparable data for analysis, the behavioural data extracted from the videos needed to be defined and coded. As *completion time* requires a frame of time with task focus and little distractions to be a meaningful between subject comparison, it could only be compared for the manual assembly assistance. The beginning of the measure was defined the moment the first part of three placed by the participant was no longer touched by the participant. The end of the measure was defined as the last part placed or repositioned by the participant before proclaiming the current run as finished was released. With this definition, participants had implicit control over the recorded starting and end time with little room for distractions and noise to arise from other interactions with the system or the environment. There was no walking, no searching for parts and little else than the interaction with the SAR system. Thinking-aloud explanations, however, are noise that could not be excluded from the data.

Success also had to be defined individually for each SAR system. For the *manual assembly assistance* it was defined as the number of correctly placed parts (out of three) at the end time defined for completion time. Even though parts may be moved or corrected later, participants were instructed that they take responsibility for the correctness of their placements when they proclaim a current run as completed. Additionally, it would complicate the measure when re-attributing correct placement to a specific run if earlier mistakes are corrected since the measure would then have to incorporate when the mistake was made and if the SAR system was activated for this run. To simplify the measure and its analysis, the above definition was chosen. For the *collaborative safety zone awareness* success was defined as "participant did not collide with the robot" and errors would therefore be counted when collisions occurred. Even though trespassing of the marked

area was also extracted and analysed, it was not chosen as the success criteria. Since the policy participants were instructed in did not specify the restricted area as such but only stated that the robot should be avoided, avoidance is the only sensible measure. For the *ergonomic monitoring* a lifting was successful when participants did not bend their lower back and instead lifted from their knees. Errors were therefore counted each time participants did not lift correctly for each run. The minimum required lifting for a given run was one time, but sometimes participants lifted more often depending on remembering which parts to pick up or other instructional misconceptions.

The last behavioural measure extracted from the recordings was the count of how many runs the SAR system was active including the current one. This secondary task of counting was chosen in favor of counting the runs themselves because the instructions numbered the runs as continuous steps of a single assignment and the current run could therefore simply be read in the instructions. Participants were asked to recount from the forth run onward having seen two runs with and without SAR system assistance, so they were able to separate between these conditions.

The other measures did not have to be extracted from the video recordings. For the non-cognitive-load measures, there is a well defined method for their use for usability evaluation which was used in the study. As previously discussed in the section on current use of cognitive load theory 2.3.2, it is unclear what exactly is measured by e.g., the NASA-TLX and other cognitive load measures. Therefore, in my opinion it makes little sense to name one of these measures cognitive load. However, trying to understand cognitive load using multiple methods makes a lot of sense. Especially, since they are used in the context of learning interactions with new technology. This is what I speak of when referring to cognitive load from here on out in the further analysis if not specified otherwise. Nonetheless, there is still no proven methodology to interpret their findings especially in mixed method approaches. In this work, they were used to complement or contradict each other as appropriate and discussed in the further specific analysis of the single methods and cross-references will be made. This will also clarify, how the methods were used in conjunction within the user study.

To derive *satisfaction*, the SUS survey score was combined with experience reports from the post trial interviews (and thinking-aloud data when appropriate). *Learnability* was comprised of a naive measure combined with self-reporting results. The naive measure was the improvement of completion time over multiple runs. For the self-reporting data, participants were asked about their learning progress in the post-trial interviews and the learnability score was extracted from the SUS survey according to the methods of [Lewis and Sauro, 2009].

For the user study, cognitive load was mainly constructed from the NASA-TLX answers combined with the thinking-aloud participation and the secondary task in recounting the number of runs the SAR systems were active.

In this work, the methods and how they were used will be presented in detail. In the next sections, I will cover why which method was deployed and which findings were used

for the evluation of cognitive load in the scope of the user study. Furthermore, I will expand on the possibilities of the methods including possible reinterpretations and reflect on their effectiveness, both using the theoretical background established in the related work chapter on cognitive load 2.

3.3 NASA-TLX

The NASA Task Load Index is a potent tool to gain self-reporting data on mental workload and possible overload for a given activity. The workload is split into six factors comprising the total workload for a given task. The identified factors are

- mental demand - (how mentally demanding the task was)
- physical demand - (how physically demanding the task was)
- temporal demand - (how urgent or rushed the pace of the task was)
- performance - (how successful the rater was in accomplishing the task)
- effort - (how hard the rater had to work to reach their level of performance)
- frustration - (how unsure, discouraged, annoyed, stressed or angry the rater was during task completion)

All six factors are rated with a 21-point scale presented with vertical lines, the first bar being equated with the numeric 0 and the last one with 20[3].

To reduce between-rater variability, the raters themselves give weights to each of the six factors in its original version [Hart and Staveland, 1988]. In turn, the weights determine to which extent the scores of individual items of the survey attribute to the total score. Later research showed, however, that it is unclear if the weights substantially influence the accuracy of the resulting rating [Hart, 2006]. While not recommended by the original author, it is therefore common practice to simply assume equal weight of all six factors and remove the weighting from the process. This was also done in the user study to reduce the already long experiment procedure. The scores of the NASA-TLX are transformed to be between 0 and 100. The total score with equal weights is therefore computed by averaging the six individual factors which are multiplied by 5. For our specific case, participants often did not mark the vertical lines but rather made an X between two vertical lines. Since individual values were not included in analysis, we opted to interpret the Xs as halfway between vertical lines and used them for calculation accordingly as .5 numerical values. Even though, the minimum and maximum of the scale move .5 toward the middle and the value exactly in the middle is lost, we still deemed it accurate enough

[3]For a visual description, the german version of the NASA-TLX survey used in the study is included in the appendix at the end of the document

to incorporate it into the user study, since we did not compare their results to other NASA-TLX results and did not need exact values for the interpretation of the data.

The arguments for choosing the NASA-TLX survey over alternatives like a simple 9 point Likert scale were the following: first, familiarity with the survey and secondly, it allows for more detailed analysis since it encapsulates more factors which can be analysed individually when appropriate. For instance, frustration and temporal demand can be explicitly evaluated and do not have to be assumed in the subsumed Likert scale. While it was not done so for the study, the more detailed separation of factors also allows for the easier identification of sources for the cognitive load when combined with Cognitive Load Theory.

In general, the choice of quantitative self-reporting data on cognitive load should be determined by usage factors. If the extraction and nuanced identification of possible sources for cognitive load is a priority, then the NASA-TLX is most likely the right choice. In contrast, if the study requires the testing of multiple interactions or conditions and therefore the data collection needs to be carried out multiple times during the trial, then the Likert scale is most likely the right choice. When other data is collected, the Likert scale also loses the disadvantage of having little context and classification of the collected data.

During our study setup, the NASA-TLX survey was to be filled out immediately after the trial. This was done to gather immediate and unreflecting data on the interaction with the SAR system. The more time passes between the stimulus and the emotional state that we want to capture, the more conscious reflection of how one is expected to feel or wants oneself to feel can impact findings and the more cognitive biases can distort the data. Especially with data that is already hard to interpret or quantify like cognitive load, introducing more noise makes meaningful analysis harder. This immediacy of feedback is one of the main drivers of researchers aiming to find objective real-time data on cognitive load (see chapter 2.2). Therefore, trying to mitigate this inherent shortcoming of retrospective data collections for the state of mind is well advised when using the NASA-TLX or the Likert Scale mentioned above.

As can be seen in table 3.1, mental workload is surprisingly high for the simple interaction. This was however most likely due to the complex study setup with the constantly changing conditions, participants issues with reading comprehension while having complicated instructions and the multiple secondary tasks participants were put up to.

The high variability is hard to interpret using only the results from the survey and require further points of reference. The most probable answer is differences in our population even though recruited locally and all having work experience in automotive manufacturing for which the full version of the Manual Assembly Assistance was developed.

Considering that participants had multiple secondary tasks and policies to keep in mind during the whole interaction, it is very likely that much of the cognitive load from the interaction arose from element interactivity of these factors. The main goal of the presented technology for policy compliance (Cobot Safety Zone Awareness and Ergonomic

Part.	Mental D.	Physical D.	Temporal D.	Performance	Effort	Frustration	Total Score
T01	32,5	7,5	7,5	-	7,5	87,5	-
T02	35	15	30	40	25	35	30
T03	77,5	37,5	32,5	32,5	52,5	32,5	44,167
T04	17,5	2,5	12,5	17,5	7,5	67,5	20,833
T05	50	0	25	10	0	0	14,167
T06	7,5	2,5	2,5	17,5	7,5	7,5	7,5
T07	37,5	2,5	2,5	50	50	57,5	33,333
T08	22,5	17,5	22,5	50	17,5	2,5	22,083
T09	22,5	7,5	7,5	17,5	32,5	17,5	17,5
T10	47,5	2,5	7,5	2,5	32,5	7,5	16,667
T11	17,5	12,5	17,5	50	27,5	42,5	27,917
T12	20	0	0	20	50	0	15
T13	10	10	10	75	15	5	20,833
T14	22,5	22,5	50	5	15	15	21,667
T15	2,5	2,5	2,5	27,5	17,5	2,5	9,167

Table 3.1: Table of Individual NASA-TLX Survey Answers. While overall load levels are not high, results for frustration, effort and especially mental demand vary highly between participants.

	Task Load
N	14
Mean	21.488
Standard Deviation	9.814
Minimum	7.5
1st Quartile	15.417
Median	20.833
3rd Quartile	26.458
Maximum	44.167

Table 3.2: NASA-TLX Descriptive Statistics. Participant T01 did not fill in the performance score and could therefore not be used for the total score. Results indicate low to medium overall load.

Notification) was to alleviate workers from these exact load creating elements of constantly having to keep policies in mind and abiding by them. Therefore, looking at the high mental demand results, it is sensible to claim that the proposed SAR systems failed to deliver the desired load reducing effects for the whole interaction/task. (This claim is strengthened when combined with data reported later in this work.)

As can be seen in the descriptive statistics (table 3.2) and the individual scores, the overall task workload is very low. However, further interpretation of possible sources for the outliers in the data is hard to do without further insight and the NASA-TLX

is therefore a sub-optimal sole measure for cognitive load. The only real interpretation without a point of reference and no knowledge about the distribution of NASA-TLX scores, is a t-test with the median of possible scores (50) as the expected value. For the sample in the study ($M = 21.488$, $SE = 2.623$), the scores are significantly lower $t(13) = -10.871$, $p < 3.4 * 10^{-8}$, $r = 0.949$. But without any further point of reference, this can only be seen as a weak indicator that the overall task workload is low.

Another issue that arose within the user study, was the participants' view on what was considered when filling out the survey answers. Participants reported filling out the survey with mainly the Manual Assembly Assistance in mind, since they perceived it as the main system and therefore wanted to give the most accurate and impactful feedback for this system. Without this information, the data from the TLX survey would have been interpreted completely inaccurately as we would have had to assume that all systems would be included equally in their perception.

Task Load Index through CLT lens

From a perspective of Cognitive Load Theory, the NASA-TLX immediately makes a positive impression since it does not claim to directly capture cognitive load. Since it is not clear if cognitive load even exists in the way asserted by the theory, claiming to be a direct and infallible cognitive load measure in the worst case without acknowledging its limitations to learning activities does not promise a great understanding of the current standing of Cognitive Load Theory. The NASA-TLX does not assert to be such a measure, although it was not developed recently enough to do so in the first place. The NASA-TLX claims to envelop the factors contributing to the total workload of a rater's task [Hart and Staveland, 1988]. This workload encompasses factors which are not covered in Cognitive Load Theory and go beyond it, but can be used for e.g., a usability evaluation.

While 'mental demand' at surface seems to be the sole cognitive load indicator within the NASA-TLX survey, most other measures can be used to analyse how efficient the learning was. As introduced in the chapter on Cognitive Load Theory 2.3.1, cognitive load is the use of working memory resources required for a given learning activity with given prior knowledge. While the mental demand adequately captures the use for working memory resource (and is therefore very similar to the often used nine point Likert scale), it does not paint a complete picture of the learning. Temporal demands, frustration, effort and performance are all great indicators for possible sources of cognitive load. If time is of the essence, it needs to be kept in mind constantly during the task and is therefore an inherent additional element required for the task which due to element interactivity leads to increases in cognitive load. Frustration is a good indication of usability errors in the interaction or task which most likely draws focus and attention away from the task to the frustrating interaction in turn creating extraneous cognitive load. Effort is a great measure in addition to mental demand to indicate whether the resulting cognitive load stays in adequate levels for the testing population. Finally, performance tells us how confident participants were in their task indicating if the learning was successful or if the

correct learning took place. Especially, when other behavioural measures to analyse task success are used to verify raters' doubts or confidence about their performance.

As can be seen, the NASA-TLX goes beyond identifying mental load levels for raters and additionally can be used to provide starting points to identify potential sources of cognitive load. However, without any additional data from the context of collection, the results of the task load survey can only be used to identify the raters' perception on the task's difficulty and their success. Score levels without the raters' expectations on difficulty or levels of skill in the task domain are hard to interpret as both might vary highly between raters. Therefore, at the very least, observations during the interaction/task completion or a clear description of the rating population and task procedure need to accompany the data to interpret the results. Otherwise, there is no point of reference to interpret score levels. As a consequence, I argue that the NASA-TLX survey is best used in conjunction with other qualitative methods which allow participants to further elaborate on their perception of the task and tested technology e.g., Interviews or the Thinking Aloud Protocol [Rooden, 1998]. That being the case, it is not sensible to use the NASA-TLX as a sole measure for cognitive load estimation. [4]

Nonetheless, the NASA-TLX has its merits for cognitive load estimation. The first major benefit is its flexibility of use, requiring no extra equipment and only a condition that wants to be evaluated. While it is hard to pin-point cognitive load levels of specific intervals of moments of the interaction, its use to cheaply and easily get a quantifyable overview of multiple factors for task load and learning of an interaction or task makes it still a valuable tool and gives a starting point for further analysis. Its difficulty in identifying moments or sources of load, however, make it a less optimal choice for evaluating UbiComp systems as argued for in the chapter on usability for UbiComp systems 2.5. Even though it implicitly suggests otherwise being a quantitative measure, it is therefore ill suited to identify or confirm small effects on cognitive load. Its best use based on this analysis is a comparison between two different conditions within participants, since its total values are hard to interpret and the individual perceptional biases should behave similarly for both interactions. However, if not used as a sole measure, it can still provide the aforementioned overview or starting point and give insight into possible factors otherwise overlooked; like the surprisingly high frustration score for some participants of the conducted user study.

3.4 Secondary Task

In this study, a secondary task was deployed to account for how much cognitive load was present. The argument is simple: if participants are able to complete the main task in addition to a secondary task which requires working memory resources, the main task's working memory resource use cannot be high enough to lead to mental overload when

[4] As opposed to the imagined objective measure sought for by researchers as reflected in chapter 2.2 which can be deployed without context to accurately capture real-time working memory resource use levels for adaptable interfaces.

Participant	#4	#5	#6	#7	#8	#9	#10	Total Runs Correct
T01	1	1	1	1	1	1	1	7
T02	1	1	1	1	1	1	1	7
T03	0	0	0	0	0	0	0	0
T04	1	1	1	0	0	0	0	3
T05	1	1	1	1	1	1	1	7
T06	1	0	1	0	0	0	0	2
T07	1	1	1	0	0	0	0	3
T08	1	1	1	1	1	1	1	7
T09	1	1	1	0	0	0	0	3
T10	1	1	1	1	1	1	1	7
T11	1	1	-	1	0	0	0	-
T12	1	1	1	1	1	1	1	7
T13	1	1	-	1	1	1	1	-
T14	1	1	1	1	1	1	1	7
T15	1	1	1	1	1	1	1	7

Table 3.3: Secondary Task: Recounting how many runs the SAR system was active after each run starting with the fourth. Most errors arose between step 6 and 7 where the experiment conditions changed most drastically. Participant T03 did not understand what the question was supposed to elicit and therefore recounted how many parts were placed in the manual assembly.

combined. For the user study, participants were asked to recount how many runs the SAR system was active. The questions started at run 4 to spare participants the first few trivial answers and to shorten the experiment time.

In Table 3.3, it quickly becomes clear, that most participants did not have an issue with the secondary task. Nine participants had no wrong answers. Most participants who failed, started doing so after the sixth step which included a more drastic condition change and interruption of the second experimenter turning on an additional obstacle in the form of a projection which was accompanied by a Collaborative Safety Zone Awareness. The goal of this obstacle was to identify if participants would also act in accordance with the safety zone, even if there was no physical obstacle present. But since participants ignored the safety zone despite the robot obstacle, they did so as well for only a projection being in their way.

While yielding quantifiable results, how to interpret them is less clear. For the participants who made no mistakes in recounting, the premise that no mental overload was reached still holds. But that on its own is not a lot of information. It is still impossible to gauge actual levels of cognitive load for learning the interaction due to having no point of reference and the additional introduction of working memory resource use skewing the levels of working memory available.

Its even worse for participants who made mistakes. Because one cannot even assume

Participant	Number of Policies Recounted Correctly
T01	-
T02	3
T03	3
T04	3
T05	0
T06	3
T07	1
T08	3
T09	3
T10	2
T11	1
T12	3
T13	3
T14	2
T15	3

Table 3.4: Table of the number of policies recounted correctly by the participants. Participant T01 was not asked due to an error. Three correctly recounted policies was the achieved maximum and each policy was forgotten more than once including all participants.

mental overload, since there are a lot of reasons participants might have forgotten the current count e.g., simply not caring to keep track. And even if mental overload was observed, it is unclear if the mental overload would have occurred if the secondary task did not have to be completed.

Reviewing the metric post-hoc, I probably would not have included it in the study, as even in the best case it yields surprisingly little new information but in any case introduces a lot of bias that all other measures are influenced by.

After completing the trial, participants were asked if they remember the policies that they were told in the beginning without us telling them beforehand that we would be asking for them. Table 3.4 shows the number of policies that could correctly be recounted by each participant. No one recounted all four policies correctly while most participants recounted three. However, every policy was forgotten at least once with no clear tendency in occurrence.

Recounting all policies is a hard task considering the usual limits of working memory when untrained [Klingberg, 2010] and the amount of other tasks required of the participants in the meantime. Especially, since participants were not told beforehand that they would be asked to recount them after completion of the other tasks. Additionally, they did not put much emphasis on the policies in the first place.

This measure while hardly being classifiable as a secondary task yielded even less information than the previous recount. There are a lot of possible influences how and

why participants might have remembered or forgotten and most of them are incredibly hard to control for. Additionally, since participants were not told beforehand that they should remember them, it is hard to argue for any relation to their working memory states during the experiment.

When not using n-back as the secondary task, the data becomes increasingly hard to interpret. With the large amount of studies using n-back to put consistent strain on the working memory, a success rate can be determined and compared. And even using n-back does not specifically aid in determining the cognitive load of learning the interaction. It can be used to use control an increase in working memory usage but it does not help in determining how successful the learning was or when the learning happened. To determine how much cognitive load the learning required, you would either need many participants with negligibly similar previous knowledge which is hard to argue for or have single individuals learn the same interaction multiple times which is impossible. In any case, it complicates the study procedure and distracts participants from the main interaction you are interested in in the first place.

While n-back is the most consistent way I have found in my research for this work to induce reproducible levels of working memory use, I would still not include it as a cognitive load measure in a user study. And since it is the most consistent and easy to interpret secondary task, I therefore would also not include secondary tasks in future studies trying to determine cognitive load for learning interactions. The increase in working memory usage due to the artificially introduced element interactivity yields too little information in return.

3.5 Behavioural Analysis

One of the lesser used methods to estimate cognitive load, is inference using behavioural data e.g., effectiveness and efficiency measures. Studies like [Chen et al., 2011] show why it probably is not sufficient to use it as a sole classifier for cognitive load, since while it correlates with the cognitive load levels "measured"[5] in the study, when used as a classifier of cognitive load, it does not perform very well. Taking a closer look at what both measures try to quantify, it becomes clear why effectiveness and efficiency measures cannot tell the whole story of cognitive load. Behavioural effectiveness and efficiency measures not in the slightest try to include the difficulty of the task and how demanding it was for its solver. They only include how fast, resource efficient and correctly the task was completed. However, it is also clear, that there is a relation between the subjective difficulty and the mental resources required of a task and how successfully and efficiently it was completed by that individual. Therefore, when evaluating the learning of a new task (which is done during user studies evaluating new technologies) it is sensible to use behavioural effectiveness and efficiency measures to aid in cognitive load estimation. How well this method worked for our user study is covered in this section. First, I will report

[5]For my critique on cognitive load measures and why it is very likely that cognitive load was not actually measured by most of them, refer to chapter 2.3.2

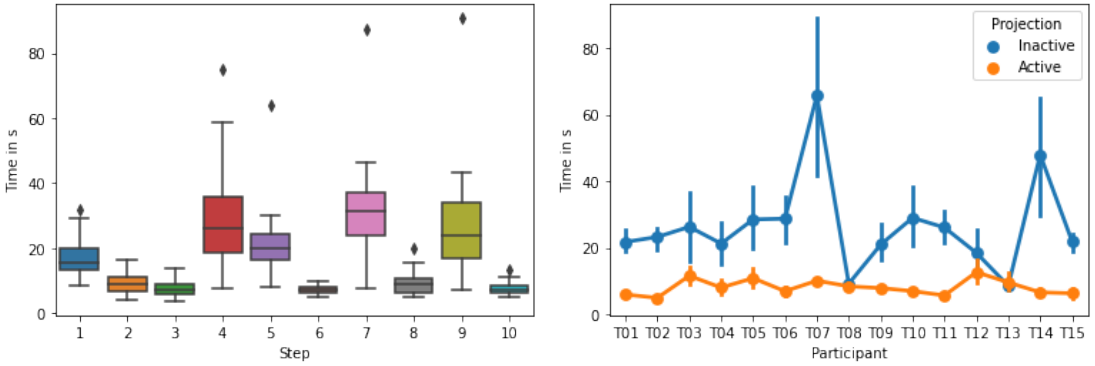


Figure 3.4: Boxplot of Completion Time per Step (left), Avg. Completion Time per Participant with Manual Assembly Assistance Active/Inactive with Error Rates (right). Steps with Manual Assembly Assistance (runs 2, 3, 6, 8, 10) were Considerably Faster.

the findings of the user study and afterwards analyse how meaningfully they can be used to estimate cognitive load.

The behavioural effectiveness and efficiency measures collected in the user study were already defined in the chapter on data and measures 3.2. As a reminder, the only efficiency data gathered in the study was the time required at the Manual Assembly Assistance. The effectiveness data for the Manual Assembly Assistance was how many parts were correctly placed. For the Collaborative Safety Zone Awareness, the number of collisions with the robot was defined as success while transgressions of the zone were also collected. Success for the Ergonomic Monitoring was defined as healthy lifting from the knees.

The outlined expteriment procedure as instructed by myself and the second experimenter was followed most of the time by participants. However, there are some outlier runs which are not comparable to the others due to heavy changes in behaviour. This makes the runs no less interesting but skews quantitative analysis of their behavioural data massively. The first outlier is participant 7 in run 5 who checked the position of every placed part manually by counting out because they realised that the Manual Assembly Assistance indicated previous placement errors. Therefore, the run at the work bench took 00:07:01. This run was consequently not included in further calculations. The last outlier was participant 8, since they ignored any placement information available at the Manual Assembly Assistance, be it on screen or via SAR system. Since all parts were placed incorrectly as a result, it would heavily skew error rates for both conditions: SAR system active and not active. Therefore, their success data for the Manual Assembly Assistance were not included in the analysis.

Manual Assembly Assistance The boxplot in figure 3.4 shows all completion times per run. As can be seen, the completion times were not normally distributed. As confirmed by a Wilcoxon test, the compppletion time for runs with an inactive projection

In-situ \ Location	Correct	Incorrect
Inactive	46	24
Active	68	2

Table 3.5: Contingency Table of Part Placement on the Work Bench. Participants made significantly fewer mistakes in placement with an active manual assembly assistance.

SAR \ Lifting Ergonomically	Yes	No
Inactive	43	45
Active	27	33

Table 3.6: Contingency Table of Participants Lifting Ergonomically. We observed no significant difference with an active Ergonomic Notification.

($Mdn = 23.259s$) was significantly higher than for runs with an active projection ($Mdn = 7.912s$), $T = 2.0$, $p < 0.0002$, $r = -0.601$. This suggests that our participants could use the SAR system to great advantage having considerably smaller time on task. From all participants, only two had a low completion time at the workbench task without the Manual Assembly Assistance (participant 8 and 13). They could achieve this by not trying to correctly place the parts which is reflected in their error rate, as we will see later in this chapter.

Looking at the effectiveness measure, as can be seen in the data from table 3.5, in runs with an active SAR system, less mistakes were made by participants at the workbench. A Barnard's exact statistic of the contingency table comparing active vs inactive Manual Assembly Assistance is -4.781 with a p-value of $2.074 * 10^{-5}$. Based on the resulting odds ratio, a participant was 17.739 times more likely to place a part correctly with an active Manual Assembly Assistance than without. Therefore, the usage of the Manual Assembly Assistance not only increased the speed with which the task at the workbench was completed but also how correctly participants were able to place the parts. This in turn suggests that the Manual Assembly Assistance made the task significantly easier, since the comparison was made within participants and can therefore hardly be attributed to differences in population. Additionally, due to the choice of condition change timings it is also very unlikely that a higher proficiency in the task was the main source of influence on these findings.

Ergonomic Notification How successful the Ergonomic Notification was in aiding participants was similarly analysed to the effectiveness of the Manual Assembly Assistance. The Barnard's exact statistic is 0.462 with a p-value of 0.719. The corresponding decrease in odds ratio of 0.856 is therefore not significant. Considering this data alone, the Ergonomic Notification apparently has little impact in how successfully participants were in lifting ergonomically as suggested by the policy during the experiment. The only condition consistently varied between trials was the activation state of the SAR system, this is the only condition that can be statistically tested for. Therefore, with no other

data source, it cannot be determined why the Ergonomic Notification had little impact.

Collaborative Robot Safety Zone Awareness Evaluating effectiveness of the Collaborative Robot Safety Zone Awareness is as hard as evaluating successful robot avoidance is easy: there was no collision of participant and robot within the 150 runs of the study. Therefore, without a larger sample (preferable from live accident data with and without the system) and without insurance-statistical methods its impact is hard to determine. What can however be evaluated, is not a measure of success, but how often participants transgressed the safety zone. But as it was only active during runs with active SAR systems, it is hard to evaluate its usefulness. The participant with the least trespasses was T08 with 2 total and 3 participants (T05, T11 and T15) had the most trespasses with 8. Without further data, it is hard to determine the reason for the trespasses, be it deeming the robot as unthreatening or not attributing any meaning to the SAR system. What can however be said, is that trespassing is common. About the cognitive load of the SAR system, however, not a lot can be argued for by this data.

Behavioural Analysis with Cognitive Load Theory

From the behavioural data alone, it is very hard to make definitive conclusions on the cognitive load of learning the new technology. This is mainly due to the fact, that the behavioral data is only influenced by cognitive load but it is unclear to what degree. While less cognitive load should yield an improvement in all chosen performance measures [Duran et al., 2022], how strong its effect is depends heavily on other factors as well. Assuming a similar load due to the similar means of communication of all three SAR systems, it is very unclear how to explain the drastic differences in success of the three technologies. The only assumption one can make, is that the differences in cognitive load have to be drastic or issues for their effective usage arise elsewhere.

Only when including other data like observations from the experiment, interviews or thinking aloud data, probable sources become apparent. Then, it becomes clear that most participants did not notice or attribute much meaning to the Ergonomic Notification and mostly ignored the safety zone after deeming the robot completely harmless. As with most quantitative measures for cognitive load estimation I analysed or used in this work, it is not easy to formulate a holistic view on how the interaction went using only a single one. Only when combining measures, an increasingly complete picture of how the interaction went and which changes can be suggested can be formulated. Due to its complexity and frequent interactions with environmental factors (especially in a UbiComp environment), cognitive load is hard to grasp otherwise having many other possible and hard to control sources of influence to attribute findings to. Since in this chapter, we are covering the value of the individual methods used a holistic interpretation of the combined cognitive load measures will be covered later in the discussion 4.

3.6 Learnability

Learnability is an interesting measure for evaluating usability. In HCI, it is most often simply defined and tested as a performance measure on how quickly new users become proficient in using the tested technology. In literature, [Grossman et al., 2009] identify two main understandings of learnability: initial learnability, which concerns itself with the performance of new systems and extended learnability, which considers performance over time.

Using the repetitive measures taken of the behavioural data, extended learnability (further referred to simply as learnability) was evaluated in this study. While only looking at the performance is a simplified view and evaluation of learning but it is nonetheless a valuable usability measure, since it allows the consideration of when usage will be effective. This is especially important for highly complex interactions.

High increases in performance alone, however, are not a guarantee for good usability, since the reached performance could still be drastically lower than the desired levels. A technology might therefore have a good learnability while not really being of much value, because the performance increase is negligible or it might even be a detriment compared to the performance before using it.

There are other measures of learnability but most of them require the acknowledgement that learning is a complex subject and cannot easily be simplified while retaining meaningful insights. Another, simple learnability measure was included in the user study, which is the factorisation of the SUS by [Lewis and Sauro, 2009], which separates the scale into two sub-scales with eight items for usability and two items for learnability.

Since cognitive load tries to estimate the level of working memory resources required for learning, it should probably be considered a learnability measure itself. Nonetheless, the chosen learnability measures give insight into how easy and successful the learning was by giving a plateau and the time required to reach it and how confident participants felt about their learning with the SUS sub-scale. Therefore, I suggest that both of these measures should be helpful in arguing the levels cognitive load of learning possible interactions with the technology. Easy and successful learning imply that no cognitive overload took place. The levels of confidence in their learning indicate perceived performance and how well participants could verify their performance. Under the older strands of Cognitive Load Theory, this would probably fall into the domain of the elusive germane load. Considering the reworked Cognitive Load Theory, it is probably only a post-hoc indicator of learning success and can therefore be used to rule out cognitive overload during the learning if the confidence was high enough.

As can be seen in figure 3.5, there is no real trend to be observed for the time required at work bench. It rather suggests, that while the individual run tasks were very similar (always having to place three parts at the work bench) their difficulty varied. Ten runs per individual is very likely not enough data to observe the first plateauing. Nonetheless, it is very unusual to see no impact of learning on the task performance time for repetitive

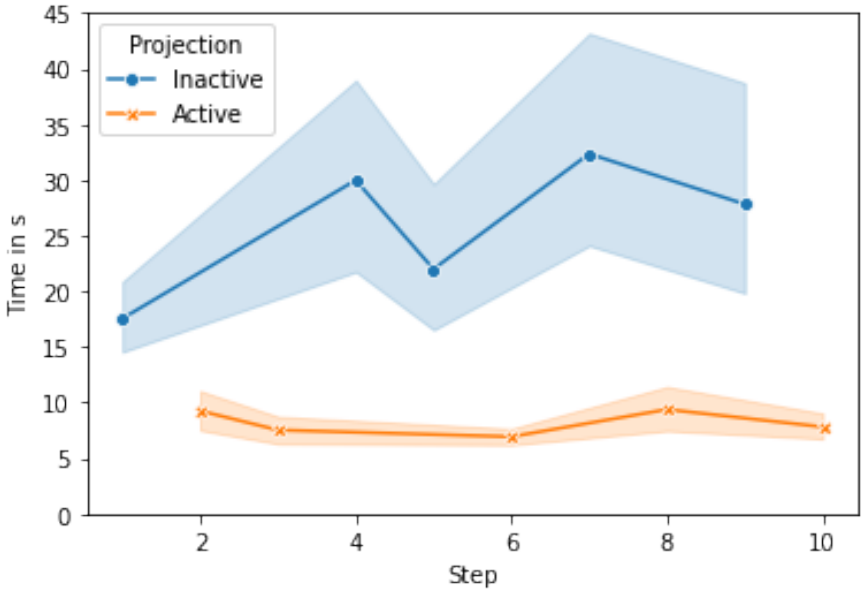


Figure 3.5: Avg. Completion Times per Step with a 95% Confidence Interval. Blue: Manual Assembly Assistance Inactive, Orange: Manual Assembly Assistance Active. No decrease in time required over trials was observed. This suggests, that not much learning was required or possible for the interaction.

measures. This suggests that, whatever the cause, little learning took place during the interaction.

Immediately before participants filled out both the NASA-TLX survey and the System Usability Scale, they reported perceiving the workbench task and the Manual Assembly Assistance as the main task an main system and therefore filled out both surveys with this in mind. Therefore, this needs to be considered for the evaluation of the SUS answers. The System Usability Scale[Brooke, 1995] (for the German version used in the user study refer to the appendix 5) consists of ten items including two factors identified by [Lewis and Sauro, 2009]. Eight items are used for the Usability factor and two items (numbers 4 and 10) for the Learnability factor. The even numbered questions are framed negatively while the odd numbers are framed positively. Each question is answered by marking on a 5-point scale ranging from 1-“Strongly disagree” to 5-“Strongly agree”. To calculate the resulting score, positively marked items contribute their position on the scale minus 1 and negatively marked items 5 minus their position on the scale. To achieve the same result with an easier to look at computation, simply add all the positive results and subtract all the negative results and add 20 again to compensate the adding and subtracting done in the step before. To have resulting scores range between 0 and 100, the resulting score is multiplied by 2.5.

Missing values for individual items should simply be replaced by the middle possible

	Score	Usability	Learnability
N	13	13	14
Mean	81.731	80.769	86.607
Standard Deviation	11.105	13.612	18.647
Minimum	60	50	37.5
1st Quartile	72.5	75	78.125
Median	85	84.375	93.75
3rd Quartile	87.5	90.625	100
Maximum	100	100	100

Table 3.7: Descriptive Statistics of SUS and the Factorisation to Usability and Learnability Scales According to [Lewis and Sauro, 2009] (with Values Scaled to 0-100 to Match SUS). The results indicate high usability and learnability.

value of 3. However, it did not seem sensible to do so, since an answer not being fitting in the eyes of raters is not the same as them not having a strong opinion one way or the other. Therefore, I simply omitted missing values and excluded their scores from analysis if the missing value was necessary for the calculation of the individual scale. While not suggested or done by [Lewis and Sauro, 2009], I also presented the resulting Usability and Learnability scales as ranging between 0 and 100 by multiplying with 3.125 and 12.5 respectively. [Lewis and Sauro, 2009] report two slightly different distributions of answered SUS questionnaires which allow for meaningful comparison of study's questionnaire answers to the distribution. Otherwise, it could only be used to compare between clearly separable conditions which, as discussed, is not trivial for UbiComp systems 2.5. Therefore, it is not required to have multiple SUS answers within participants to conduct meaningful SUS analysis.

When comparing the results to the distributions of the SUS presented in [Lewis and Sauro, 2009], the results collected in the study are higher (see Table 3.7). Considering the median of the pessimistic distribution presented, only four scores are bellow the first quartile and seven results are above or at the 3rd quartile. While by no means perfect, this suggests generally high or at least above average usability. More importantly, the learnability scores are even better, with seven participants rating the highest possible learnability score. Interpreting the learnability of the SUS alone, the only claim that can be made, is that whatever learning happened, it was rather easily done.

Combining both the SUS results and the behavioural learnability measures, it seems very likely that not much learning was necessary for the interaction at the workbench in the first place. Therefore not much learning was done and the little that was done, was very easily manageable. Since there was not much learning happening and the learning was very simple, the resulting cognitive load is very likely to have been very low.

Participants did not perceive the other SAR systems as useful or important in the context of their interaction. Therefore, they were filling out the questionnaires with neither the Collaborative Safety Zone Awareness nor the Ergonomic Notification in mind.

Furthermore, they did not interact much with them, resulting in questionably useful behavioural data. As a result, the learnability data is lacking for the other SAR systems.

3.7 Thinking Aloud

The thinking aloud protocol[Rooden, 1998] was developed to gain live impressions of users about the interaction. This way, ideas, joys and frustrations can be voiced the moment they come up and therefore more accurately attributed in analysis. To accomplish this, participants are instructed to voice their thoughts and feelings about how the interaction with the technology is going the moment when they arise. When designing the study, I thought that when participants abruptly stopped thinking aloud that cognitive load spikes must have occured and that they could be pin-pointed in the video recordings. This plan to gather information on the state of working memory resource use fell short due to two main reasons:

1. many participants did not think aloud and
2. when they did, they did not do so constantly.

It is a common and known issue with thinking aloud that not all participants participate in it even when reminded to do so during the study. So even though participants are reminded that they should proclaim their thoughts and feelings about the interaction, some simply do not and it is hard to attribute this to any particular reason. Nonetheless, I thought cognitive load spikes could be detected this way.

While nice in theory, this abrupt stop proved hard to quantify and extract from the video recordings. Because even for participants who did think aloud, abrupt stops never occurred. Since this idea fell flat, I wanted to extract periods of time where participants thought aloud but this would be heavily biased data by talking speed and style alone. Therefore, it could not be used to determine cognitive load.

While also leading to an increase in element interactivity, the thinking aloud protocol yielded impressive yet inconclusive data on the SAR system, study setup and usability problems arising during the interaction. It worked well as a control, especially combined with interview data, as it helped in forming a clear image of what was important for participants during the interaction. It aided in identifying their perception of the SAR systems, as in that they did not care to interact with the policy related SAR systems, since they could not identify their usefulness. We could use it to identify multiple usability issues of the whole interaction and wishes for different use cases. However, it aided nothing in identifying the cognitive load for learning the interaction. Its data could only be used to identify limitations of other measures and inform us that participants rarely considered all SAR systems, also when filling out the questionnaires.

3.8 Cognitive Walkthrough

The decision to include a cognitive walkthrough to gather more data on the cognitive load of the interaction was done post-hoc after first cutting it from drafts of the study plan. The original cognitive walkthrough [Lewis and Wharton, 1997] assumes a clear interaction between user and computer and unambiguous correct steps as well as the possibility for usability experts to play through this interaction step by step. For the SAR systems in question, this does not work for multiple reasons. First, there is no clear one way to achieve the task and it does not require taking actions with a computer to be accomplished. Instead users can opt to use information available to them to aid them in their task. Then, ideally, there is also no clear interaction with the SAR system and the interaction merges into the background of the repeated task providing ambient assistance. Lastly, the experts could not interact with the prototype since it no longer existed. Therefore, the methodology needed to be adapted to be usable for the purposes of this study.

There would be no interaction of the experts with the SAR systems. Instead, they were shown anonymised video footage of interactions of the representative target users recorded during the user study. The footage was 13 minutes and 50 seconds long and was comprised of various successful and unsuccessful interactions chosen to include a broad variety of observed interactions with the SAR systems active.

The standard questions aiming to yield easy answers to complex questions were adapted to the new situation and more nuanced possibilities for accurate evaluation.

1. Will users try to achieve the right result?
2. Will users notice that the correct action is available?
3. Will users associate the correct action with the result they are trying to achieve?
4. After the action is performed, will users see that progress is made toward the goal?

became

1. Did users notice that aiding information for their task is available? Why? (In other words, is the information visible or easily findable when performing the task?)
2. Did users associate the available signals with the information they are trying to communicate? Why? (Even when the signal is visible, will users know how to interpret it correctly?)
3. Did users incorporate the available information when working toward their goals? Why? (Even if visible and interpreted correctly, could it be usefully incorporated into the task?)

4. After an action is performed, did users see if progress is made toward the goal? Why? (Based on what occurs after an action is taken, will users know that this action was correct and helped them make progress toward their larger goal?)

each including an explanation to preempt miscommunication.

The procedure started with showing each HCI expert a sample interaction of the second experimenter showcasing basic interactions with the SAR systems in a dummy run from an over-the-shoulder, third-person perspective. Then, they were asked what they thought the interaction and what the SAR systems were doing. After collecting their perception, they were debriefed on the SAR systems functionality and aims and the purpose of the user study. Thereafter, they were shown the selected interactions from the video recordings made during the user study which had multiple perspectives, one showing the workbench and two showing the back area including the Ergonomic Notification and Collaborative Safety Zone Awareness. During the whole demonstration, the experts were asked to immediately voice their thoughts similarly to the thinking-aloud protocol. Afterwards, they were asked the rephrased questions. In the end, the experts were debriefed about the purpose of the cognitive walkthrough to test a method to estimate cognitive load for this thesis followed by open questioning which was used to clarify previous statements and voice final thoughts.

With these changes, a more accurate name of the method would be "a qualitative video-guided expert evaluation based on the cognitive walkthrough by [Lewis and Wharton, 1997] adapted for use with a UbiComp system". However, it does not make a great section title.

Expert A has a PhD with a focus in HCI and 15 years experience in user experience design and research. Expert B is an associate professor with experience in HCI, user experience research as well as persuasive technologies. Expert C has a PhD in Engineering Sciences with a focus of HRI research in industrial contexts.

Usability issues were easily found, some better backed by other data than others. Only one expert, Expert A, could talk intricately about cognitive load since she incorporates it into her interaction designs. While not explicitly stating a source during the conversation, she used a cognitive load understanding very similar to the newer strands of Cognitive Load Theory covered in chapter 2.3. The other experts, Expert B and Expert C, while providing excellent feedback on the system's usability, did not have additional in-depth knowledge on cognitive load.

All three experts immediately criticised how the Ergonomic Notification was implemented. To present information about lifting by marking a space in the factory with a specific color was identified as being confusing and resulting in an informational disconnect. With users not knowing precisely how the information about the policies would be made available to them, they were irritated when the region from which they retrieved the required parts turned red upon there picking them up. Luckily, most participants promptly ignored the information but were nonetheless disturbed in their workflow. Expert A additionally

criticised this flow disruption as a strain on the mental workload, especially one that is entirely unnecessary.

The information ambiguity was additionally increased, since marking a space red was also used for signalling users that they trespassed the safety zone. So, the same type of information was used to signal two completely different policy transgressions which only have in common that the user did not comply with it (which might not even mean wrong or sub-optimal behaviour depending on the case).

All in all, the Ergonomic Notification was identified as badly designed to inform users on their state of healthy lifting. The information was provided in the wrong place at the wrong time to be effectively included to facilitate better lifting. In addition to not providing much information of use, it added elements to be considered to the workflow and increased the resulting cognitive load of learning the interaction. It was the only SAR system failing all four criteria of the adapted Cognitive Walkthrough questions.

Continuing with a trend in disturbing the workflow, Expert A identified the red flashing of the Safety Zone as another disruption, since participants in the videos stopped within the Safety Zone to turn and see what was flashing red on the edge of their field of view. Expert A called it especially intrusive as moving and flashing objects at the edge of our vision pass unfiltered to our brain for identification which always leads to a disruption and additional mental workload. Instead, she suggested marking it in red permanently without flashing while still moving it when necessary, to allow users to incorporate the information into their workflow subconsciously without disturbing them. This idea was also proposed by Expert B to communicate the Safety Zone more clearly to participants. Before being debriefed about the purposes of the SAR systems, all participants initially identified the Collaborative Safety Zone Awareness as a wayfinding system due to its confusing presentation as an outlined geometric shape being accompanied by a line moving between the user and the Safety Zone. The Collaborative Safety Zone Awareness got very mixed reviews for the four criteria of the adapted Cognitive Walkthrough questions. Expert B identified it as fulfilling the first three criteria at least to an extent since one participant did also incorporate it into his pathfinding. However, this participant in this run was the only instance where a change in route was made due to the Safety Zone. In all other cases, participants waited for the robot to finish moving out of their way or simply ignored the Safety Zone, when the robot was not in their way. Expert A and Expert C rated it as fulfilling the first and second criteria in most cases, but not fulfilling the other two.

The Manual Assembly Assistance was described as being used to great effect by the participants (even though one of the shown participants ignored the placement information completely). Expert B identified constant blinking and inaccuracies of the marked placement errors as incredibly distracting. When comparing it to the previous statements of Expert A on the Safety Zone it would also classify as disrupting the workflow and unnecessarily increasing mental load. Apart from that, all three experts identified it as successfully supporting participants in their goals and even giving them feedback on their

progress. It was therefore identified as the only SAR system successfully fulfilling the four modified Cognitive Walkthrough Criteria.

As Expert C pointed out, even though the SAR systems are by no means designed perfectly, most of the usability errors could simply have arisen from the lack of instruction with the technology. However, since participants did use the Manual Assembly Assistance to great success even without any instruction, it also stands to reason that the other SAR systems simply did not provide their information clearly and usefully enough to the participants.

As it is the case with the previous methods from this chapter, it is hard to make conclusive claims with the data from the cognitive walkthrough alone. New issues and sources of cognitive load could be discovered using this method, but this is not a given. The quality of results depends heavily on the specialisation or interests of the expert in question. To accurately gauge usability issues and evaluate the interaction success, the experts require experience with UbiComp systems. To evaluate cognitive load, the experts need deeper understanding of human perception and basic knowledge of Cognitive Load Theory. In my case, I was lucky enough to have an expert working with cognitive load with an understanding matching the newer strands of theory, but this is not a given. Additionally, it is hard to recruit experts matching the description, especially when considering the ambigous work on cognitive load of UbiComp experts covered in the related work 2. Despite the partial successes I gained by using this method, I would therefore not consider this adapted version of cognitive walkthrough a reliable method to gauge cognitive load. To mitigate the mentioned issues, it might be sensible to instruct the HCI experts with UbiComp system experience in Cognitive Load Theory to work with a compatible vocabulary for the evaluation and better detect transgressions and unnecessary interactivity (leading to extraneous load). However, it would heavily increase the time required to conduct the method.

Additional problems arise due to the basically required post-hoc video reconstruction of an interaction. It cannot recreate the perception of interacting with the system in question in a real-life setting. While creating a test setup on the fly is trivial for traditional human-computer interaction, it is the opposite for UbiComp systems. Having the space and resources to create even an artificial test setup temporarily requires many resources. Then, fitting experts willing to come to the location and users into the time frame where the setup is available is anything, but simple. And it cannot be done remotely which will drastically decrease the number of experts willing to participate.

3.9 Control Data

The data in this section was used to put all findings into perspective and include factors otherwise not included in the study. There are multiple variables we wanted to control for. First, we wanted to know, how participants perceived the individual SAR systems and which functionality they attributed to them based on their interaction (see table 3.8). This information was mainly elicited in the post trial interviews by asking participants

Participant	MAA	CSZA	EN
T01	Full	Partial	None
T02	Full	Full	Partial
T03	Full	None	Partial
T04	Full	Full	None
T05	Full	None	None
T06	Full	Partial	None
T07	Full	Full	Partial
T08	Partial	Full	Partial
T09	Full	Full	None
T10	Full	None	None
T11	Full	Partial	None
T12	Full	Full	Full
T13	Full	Partial	None
T14	Full	Partial	Full
T15	Full	Partial	None

Table 3.8: Recognition of Systems by Participants. Columns from left to right: Manual Assembly Assistance, Collaborative Safety Zone Awareness and Ergonomic Notification. There is a clear decline of recognition from left to right.

what the projections were doing and how. It was also supported by the thinking-aloud data during the trial. As can be clearly seen, participants did not consciously perceive the Ergonomic Notification. There was some understanding and perception on the Collaborative Safety Zone Awareness, but only a third of participants recognised it is a zone they should not enter because the robot was moving there. The Manual Assembly Assistance was understood by all but one participant. This also explained, why participants reported filling out the surveys with the Manual Assembly Assistance in mind, since it was the only SAR they were confident in understanding correctly.

The total experiment time was collected with the aim of correlating it to other measures if findings were difficult to explain using other data. Since both SAR systems of interest even failed in being recognised and were not used by participants, correlation with their success criteria was not deemed very useful. For the Manual Assembly Assistance, time data with less noise and bias was chosen in evaluation.

Planning the experiment, a random choice of paths by participants was expected, if no route was planned. But even though there was more reason to choose the right path based on the location of parts communicated in the instructions, the left path was chosen 204 times, while the right path was chosen only 104 times. Therefore, the location of the parts and the resulting slightly shorter route cannot have been the deciding factor in route planning. Why exactly this happened is unclear, but the left path was a little wider and was used to show participants the experiment area.

Participant	Total Experiment Time
T01	15:48
T02	14:15
T03	14:37
T04	16:14
T05	16:12
T06	19:04
T07	37:59
T08	11:50
T09	17:44
T10	15:10
T11	17:51
T12	14:48
T13	12:47
T14	20:37
T15	16:14

Table 3.9: Total Experiment Time per Participant. Participant T07 recounted all placements at one point during the experiment and was very meticulous about placing correctly in general, resulting in this outlier time.

CHAPTER 4

Discussion

4.1 Cognitive Load - A Holistic View

As can be seen by the data reports and arguments of the previous chapter, it is hard to make definitive claims on the cognitive load of learning the interaction using a single method shown. Each method only alludes to parts of the whole interaction and is unable to capture every strand. Using the behavioural data, it is possible to say, that the Manual Assembly Assistance had a positive impact on efficiency and effectiveness and the other SAR systems probably did not. As to how and why cannot be answered using this method without testing multiple different scenarios and setups. Therefore, it is impossible to make definitive claims about the cognitive load of learning each system. But one could assume, that the cognitive load for the Manual Assembly Assistance was not high enough to result in cognitive overload, since participants were able to use it successfully.

Considering the control data, that most participants did not recognise the other SAR systems, possible reasons can be argued for. The SUS data tells us that participants report little trouble learning the interaction and most felt confident in their ability with the Manual Assembly Assistance. With the minimal changes in time-per-run over time (Fig. 3.5), the learning is very likely to have peaked very quickly. Both of this suggests low cognitive load for learning how to interact with the Manual Assembly Assistance, since lower cognitive load measures have until now resulted in easier learning [Duran et al., 2022]. This is additionally supported by the NASA-TLX results being on the low end of spectrum and also mainly yielding data for the Manual Assembly Assistance. Looking at the learnability data and the technology in question, it also very likely that not much learning had to be done in the first place which made cognitive load effectively a non-factor for successful interaction with the Manual Assembly Assistance.

The secondary task of recounting how many runs the SAR systems were active, was accomplished very well, which supports the view of overall low mental workload during

the trial even though the experiment setup included elements artificially increasing the mental workload. Mainly, these are the secondary task and the experiment structure. This increase in load is indicated by interview responses of participants prioritising to suggest improvements of the task structure and laboratory setup for a better workflow and less opportunity for misunderstanding. Additionally adding to the baseline mental workload, further sources could be identified in the cognitive walkthrough. Flashing red lights which were often perceived as random, indirect location information with geometric shapes, and the double attribution of a red zone used to communicate two entirely different circumstances all contributed in increasing the overall mental workload of perceiving in the lab environment.

Therefore, while the cognitive load was low overall, many unnecessary sources of extraneous cognitive load of learning the task could be identified. Without combining the findings from the multiple methods used, neither argument could be made confidently. Relying on only one of the measures to determine cognitive load for the interaction e.g., the NASA-TLX, would not allow any definitive claims at all.

Comparing this to the methods discussed in the related work 2, the same issues are present. Authors are trying to find a single method to determine absolute cognitive load in real time with the goal of adaptive automation. However, as argued for in the related work, there is currently no means to meaningfully interpret their data. As could be seen in the previous chapter 3, it would be the same, if I were to choose a single method applied in the case study to determine cognitive load. It would be very difficult to formulate convincing answers, especially when trying to use the same method for every UbiComp system expecting similar and comparable results.

Maybe, the goal of adaptive automation is too far a stretch for the current knowledge of mental workload, the state of the human mind and cognitive load as a measure. Instead, I suggest the alternative goal of developing a toolkit of mixed methods to convincingly determine cognitive load for flexible circumstances of technology adoption. This toolkit of methods would then be used in user studies of new technologies like the one conducted in the case study. It could however not be used for the goal of adaptive automation. But in development and testing of this toolkit, a better understanding of Cognitive Load Theory and its related measures could be gained to further refine the theory and in the end it might be a necessary step on the way to adaptive automation.

While using something similar in this work, the methods tested do not fit the criteria of this optimal toolkit I would be looking for. Rather, doing the case study and working intricately with Cognitive Load Theory showed me that a similar but more refined and tested approach might be well suited to determine cognitive load, especially for evaluating usability and adoption of a new technology.

In general, after evaluating the data of the case study, I think that it is actually not as important which methods one uses to estimate cognitive load, but that it is important to look at it from many avenues. Otherwise, oversimplified and likely wrong assumptions will be made which are not contradicted in the data. Especially, quantitative data which

yields very simple and rather definitive claims can be combined with qualitative data which includes a broader context at the cost of less definitive claims. Looking at the data from the methods tested in this study and the methods reviewed in the related work chapter 2, there is no obvious right path, but there are some wrong choices.

Starting with the ones I conducted, the adapted cognitive walkthrough is presumably one of these wrong choices. If it were not for Expert A, it would have yielded little data of value for cognitive load evaluations, since the experts require knowledge of Cognitive Load Theory and UbiComp systems to conduct accurate analysis. Such experts seem hard to come by, considering the questionable understanding of Cognitive Load Theory evident of experts publishing in the field of Ubiquitous Computing covered in the related work 2.2. In addition, the material need to be adapted and preparing so the method can be conducted or the experts need to be brought to the study setup. Even then with the experts recruited, it only yielded data which required further vantage points to be meaningfully interpreted. Therefore, I would not recommend the use of this method to estimate cognitive load.

While the thinking-aloud protocol does not add a lot of work to study preparation and even though it was very valuable to control for the participants experiences and priorities, I would not rely on its use as a cognitive load measure as proposed in the previous chapter 3.7. It added nothing of value to determine cognitive load for the case study and is undependable even as a control, since not every participant participates equally. But since it does not cost much in terms of resources to conduct, it can still be included in a study, just not as a reliable cognitive load measure.

Intrusive methods, like [Saha et al., 2018] using EEG, [Murata and Suzuki, 2015] measuring blood flow by taping sensors to the neck, and [Gavas et al., 2017, Pillai et al., 2022] using wired eye-trackers all have the same issue: they are very distracting or cumbersome to apply. [Fridman et al., 2018] solved this issue by using a simple camera to extract eye-features. Apart from possibly being distracting and demanding too much attention from the participants, I have few objections about including objective measures into the toolkit. However, relying on one objective measure to estimate cognitive load while calling its results objective cognitive load will not yield interpretable, comparable and meaningful findings with current methods.

On the use of secondary tasks, I am torn. They inherently add to the mental workload of the participant introducing bias to all other measures. On the other hand, when applying standardised methods like n-back, they can be somewhat controlled to quantify left-over working memory by triggering and detecting cognitive overload for some but not all participants and comparing load between participants. However, interpretation of this data is not as clear cut, as one could ask for, since cognitive load is highly individualistic requiring many participants to make quantitative evaluation reliable, which again would increase resource requirements for conducting the study. After having applied my own (compared to n-back, less demanding) secondary task, I would not use one again with my current knowledge, if the goal is to estimate cognitive load for technology adoption as part of a user study evaluating usability. I see their greatest application (especially

n-back) in accurately determining working memory limits to work on refining Cognitive Load Theory, as laid out in section 2.3.2.

All other methods used in the study were easy to conduct and did not add any complexity during the interaction and therefore did not increase the baseline mental workload. Combined, their results could be used to formulate a convincing interpretation of all findings explaining the observed phenomena (as can be seen in the beginning of the chapter). Therefore, I would deem them useful as possible additions of the aforementioned toolkit.

Looking at the partial and limited insight any used method allows, it quickly becomes clear why researchers are trying to find a one-size-fits-all solution to determine cognitive load. But considering their questionable interpretability and the complexity of cognitive load that can be seen even with the simple interactions of the case study, the methods covered in chapter 2 do not try to do the complexity justice. It is understandable not wanting to account for the intricacies of human perception and learning when evaluating new technologies or learning material. But if the aim is to find holistic and conclusive findings, chasing easy answers with cheaply available technology to gather bodily measures without trying to account for influences will not yield the required understanding.

Now, I investigate the negative findings of the holistic view on the interaction. As participants reported ill-timed flashing lights that they could not attribute meaning to, if they even noticed them at all, eludes to a major usability issue: that the information is presumably not presented in the position at the right time. As information positioned visually in the environment is the only information available for users, this would make successful and effective interaction functionally impossible. Since participants either did not notice or attributed little meaning to the Safety Zone and the Ergonomic Notification, the resulting mental workload was presumably not very high. But as Expert A pointed out in the cognitive walkthrough, blinking red lights perceived as random or which could not be attributed with much meaning are disruptive and need to be worked through by working memory.

In the video recordings chosen for the cognitive walkthrough alone, participants forgot what they were doing due to being alerted after entering the Safety Zone. But since there was no real danger present, the participants were simply confused, their workflow was interrupted, and they had to retrace their steps. The same confusion by unclear information, while not as frequent (since only 6 participants noticed them at all), was reported in the interviews about the Ergonomic Notification. Therefore, the other SAR systems added mental workload while not providing much of value. The higher baseline mental workload allows less resources to be used for cognitive load of actual learning increasing the likelihood of cognitive overload, which would prevent the learning and in turn the successful adoption. Considering element interactivity and the blinking red lighting, the Collaborative Safety Zone Awareness and the Ergonomic Notification did therefore not only provide nothing of value to the interaction but were an active detriment to task completion, albeit only a minor one. While undoubtedly increasing

task performance at the the workbench, all three SAR systems are neither optimised for cognitive load nor usability in general.

4.2 Reflection on the study structure for a UbiComp system

If the technology is useful easily and transparently helps the user in task completion, it will be adopted even without any explanation, if the task is easy enough. With a task as simple as the one of the case study, this could be seen with the adoption of the Manual Assembly Assistance. If, however, the technology does not present useful information on what users deem important at any given point during the interaction, users will not use it. This is what happened with the Collaborative Safety Zone Awareness and the Ergonomic Notification.

While confirming my suspicions about the SAR systems not being visible and placed correctly, I would always explain the functionality of tested technology to participants for future studies. While it might still be required to elicit the users perception of systems and explanation does not need to be done at the beginning of the experiment, I would measure usage success only after explaining how the technologies worked in the future. A lot of the data was unusable due to me not explaining all systems before starting the experiment, resulting in confusion and possibly influencing how valuable the other SAR systems might have been to the participants. The bad visibility could have easily been elicited by a blind trial run before starting the other runs.

However, especially when evaluating UbiComp systems, I would give participants the informed option of not using the system to accomplish the task if possible. For UbiComp systems aiming to streamline workflows rather than replacing existing workflows with improved ones, it makes little sense to force participants to use the technology for user testing. If participants opt not to use the technology in the user study, the technology brings no apparent improvement to them. It is immediate and clear feedback on low satisfaction and is a red flag that requires resolution in e.g., post-trial interviews.

As already mentioned when reporting the NASA-TLX data from the user study, when researching UbiComp systems it easily can become unclear for which part of the system quantitative data is gathered. In our case, this was only known because participants told us that they were filling out the surveys with Manual Assembly Assistance in mind. Because of UbiComp systems' increasing subtlety and interweaving with the environment it is hard to determine what gathered quantitative data is measuring. As mentioned in the chapter on the usability of UbiComp systems 2.5, it is not trivial to which interaction between environment, UbiComp system and user the gathered data has to be attributed to. This is especially true without further data clarifying the participants' perception and experience of interaction.

While being worse for self-reporting data which is inherently dependant on the participants' experience, this problem also transfers to objective data for cognitive load

measuring the participant or their behaviour. Cognitive load depends heavily on the previous knowledge and how the presented information is already available as schemata which in turn determines how many elements have to be kept in mind at one time to successfully complete the interaction/task [Duran et al., 2022]. To know which elements are interacting to demand the levels of working memory resources, it is therefore necessary to know the participants' perception and abstraction of the situation. Any data measuring cognitive load, is therefore depending on the participants' perception to be correctly interpretable. Otherwise, it is not possible to know what the gathered data is referring to due to the necessity of interweaving parts for accurate representation which cannot be controlled for in an experiment.

That being the case, the only way quantitative data for UbiComp systems evaluation can be meaningfully interpreted, is when the whole UbiComp system including its environment is evaluated. And even then, only performance data can be confidently used as otherwise it cannot be known if the whole system was perceived as such. Shallow quantitative data on participants' state of mind alone, while statistically analysable, will not yield the data required to formulate adequate changes.

This is where qualitative methods for evaluation come into play. While not as easily generalisable and assessable using mathematical methods, they allow for capturing the levels of nuance required to identify issues. This again supports my call for developing a method toolkit to evaluate cognitive load. I would go one step further and would advise any evaluation of UbiComp systems to alleviate itself from questionable findings by incorporating qualitative methods into their study design to allow for more nuanced approaches to interpret the context and findings. Maybe a similar call to build a comparable, flexible method toolkit to evaluate UbiComp systems is appropriate as well. As already mentioned in the chapter on usability for UbiComp systems 2.5, there is increasing work with e.g., [Rocha et al., 2017], [Carvalho et al., 2018] and [Bezerra et al., 2014] trying to establish new criteria which encompass these arising issues. The adaption of methods, as done by [de Souza Filho et al., 2020], and development of new methods encompassing these criteria could be the next step, to improve the evaluation of UbiComp systems. If attention is put on replicability as well by streamlining the application of the proposed methods, it might additionally aid in combating the replication crisis.

As argued before, little cognitive load data could be gathered about the Safety Zone and the Ergonomic Notification, since there was little interaction with the systems. Cognitive load starts playing a role for evaluation after the used and tested technology can be used by representative participants to improve current workflows or to create new and improved workflows. Only then it makes sense to optimise for an improved learning (low cognitive load/mental workload) or less demanding continued use (low mental workload). For traditional human-computer interaction, participants are practically required to use the technology to fulfill their task. So despite being difficult to adopt and use (high cognitive load/mental workload) and using it being tiresome or unnecessarily confusing (low satisfaction, high mental workload), the usage of the new system might still be evidently more performant than the earlier way of accomplishing the task (high

effectiveness/efficiency).

For UbiComp systems mostly aiming for minimal workflow improvements which add up over a longer period of time (small increases in effectiveness/efficiency, especially the proposed policy compliance assisting SAR systems) and their improvement no longer being immediate or clear, users will no longer abide tiresome usage and difficult adoption. This in turn means that, when given the opportunity, potential users will opt not to use it. And since most measures used to evaluate usability and cognitive load require usage of the technology, it might be more sensible to apply cognitive load measures after successful usage becomes evident.

For the case study conducted in this work, this means that a lot of cognitive load measures could be gathered and evaluated for the Manual Assembly Assistance which, while not evaluated formally, was already previously used successfully used by target users. For the experimental SAR systems which were neither developed with target users nor their workflows in mind and were not previously exposed to target users, not much data on cognitive load could be gathered, especially using the quantitative measures deployed.

CHAPTER 5

Conclusion

To summarise, in this thesis I give an overview of the current standings of cognitive load estimation in the field of computer science, especially human-computer interaction. The thesis starts with an examination of current methods mostly improving on objective measures to determine cognitive load. They do this with the clearly stated goal of advancing adaptive automation of interfaces, especially in UbiComp systems. I begin by comparing them according to their self-reported findings: classification accuracy of a reproduction of non-standardised and ill-controlled induced working memory use which is called cognitive load. Within this framework of success, self-reporting measure like [Chen et al., 2011], [Fridman et al., 2018] and [Yin et al., 2007] performed best by a heavy margin. However, it is hardly possible to meaningfully compare their reported percentages of “cognitive load level classification" due to methodological flaws. Afterwards, I argue at length three major issues of this current way to determine “cognitive load". According to newer strands of Cognitive Load Theory, cognitive load is the amount of working memory used by a specific learner with their specific prior knowledge during a given learning activity [Duran et al., 2022]. In the works determining new means to measure “cognitive load" included in the corpus,

- there is no learning activity per se (instead there are only means to induce mental workload),
- the prior knowledge and cognitive ability are hardly controlled for,
- and the means to induce mental workload are not guaranteed to induce similar levels of working memory use for each individual, especially if neither prior knowledge nor cognitive ability are determined.
- Finally, the results are called “cognitive load" while not being actively interpreted based on their limited control for populations and there is no incorporation of Cognitive Load Theory in placing the results.

Afterwards, I go into detail about possible uses in HCI of the cognitive load defined in the theory, which yet requires empirical solidification. I argue that adoption of a new technology is little but a learning activity, since new interactions of the technology have to be learned for successful and continued usage. Therefore, I suggest the usage of Cognitive Load Theory as additions to usability evaluations instead of pursuing adaptive automation given the current unproven stand of the theory and the methodological flaws of current measures.

Then, I proceed to apply the knowledge of Cognitive Load Theory to compare different methods to determine cognitive load to the ones covered in the corpus which I used in a user study evaluating the usability of three Spatial Augmented Reality systems. The results were pretty clear: any method alone could not be used to convincingly answer the cognitive load of learning the interaction with the systems. Only by examining the interaction with multiple cognitive load data sources and the control variables of post-trial interviews and thinking-aloud data could convincing arguments be made that relied little on speculation. The control variables and multiple data sources were especially necessary due to the increased interaction complexity inherent to UbiComp systems introducing interactions between human-environment and computer-environment.

Based on the strengths and flaws identified for each individual method, I now give my recommendation for their future use to determine cognitive load in user studies of UbiComp systems. The thinking-aloud protocol could not be used to determine cognitive load in any meaningful way, since abrupt breaks in speaking could not really be quantified and were very infrequent in the first place due to mixed participation and participants not talking constantly. However, it was a very valuable and easy to apply control variable. My adapted version of the cognitive walkthrough is unreliable requiring much preemptive material adaption or organisation and very specific experts of UbiComp systems with an accurate understanding of Cognitive Load Theory (which seem rare due to the work in the field covered in the related work chapter 2). For these two reasons, I would not recommend its use to determine cognitive load.

While its data was valuable, any secondary task increases the baseline mental workload resulting in increased working memory use. It skews all other results to higher cognitive load due to element interactivity while leaving less working memory for the learning to happen. Therefore, I would no longer use it in the future in usability evaluation since its data is not valuable enough to warrant the negative consequences in my opinion. The mental workload data from the NASA-TLX, the learnability data from SUS and completion time over time, and the behavioural data were all valuable in determining cognitive load while being easy or necessary to apply in a usability evaluation in the first place. Therefore, I can recommend their usage.

Reflecting on the method use, I called for the development of a method toolkit which aims to determine cognitive load convincingly and flexibly depending on circumstances while being transparent and easily replicable. Since it was not as important which method specifically was used, but rather that multiple data sources were included in the analysis, there is a lot of room for experimentation and potential development of new

methods. Even the methods covered in the related work can easily and successfully be deployed if they are not used to claim the one and true objective cognitive load without reflection and active analysis. However, I would refrain from using methods which require cumbersome and distracting sensors to gather their data and rather use methods like [Fridman et al., 2018] and [Yin et al., 2007] which do not even require wearables for participants.

Finally, while not in my field of expertise, enabling the further development and post-positivist interpretation of the methods covered in the corpus would require determining base-line levels of minimal and maximal working memory use as well as means to confidently achieve them. Future work could therefore entail the furthering of either those goals.

Bibliography

[Arshad et al., 2013] Arshad, S., Wang, Y., and Chen, F. (2013). Analysing mouse activity for cognitive load detection. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, OzCHI '13, page 115–118, New York, NY, USA. Association for Computing Machinery.

[Bainbridge, 1983] Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6):775–779.

[Bezerra et al., 2014] Bezerra, C., Andrade, R. M. C., Santos, R. M., Abed, M., de Oliveira, K. M., Monteiro, J. M., Santos, I., and Ezzedine, H. (2014). Challenges for usability testing in ubiquitous systems. In *Proceedings of the 26th Conference on l'Interaction Homme-Machine*, IHM '14, page 183–188, New York, NY, USA. Association for Computing Machinery.

[Brooke, 1995] Brooke, J. (1995). Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189.

[Carvalho et al., 2018] Carvalho, R. M., Andrade, R. M. d. C., and de Oliveira, K. M. (2018). Aquarium - a suite of software measures for hci quality evaluation of ubiquitous mobile applications. *J. Syst. Softw.*, 136(C):101–136.

[Chen et al., 2011] Chen, S., Epps, J., and Chen, F. (2011). A comparison of four methods for cognitive load measurement. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference*, OzCHI '11, page 76–79, New York, NY, USA. Association for Computing Machinery.

[Crabtree and Rodden, 2009] Crabtree, A. and Rodden, T. (2009). Understanding interaction in hybrid ubiquitous computing environments. In *Proceedings of the 8th International Conference on Mobile and Ubiquitous Multimedia*, MUM '09, New York, NY, USA. Association for Computing Machinery.

[de Souza Filho et al., 2020] de Souza Filho, J. C., Brito, M. R. F., and Sampaio, A. L. (2020). Comparing heuristic evaluation and maltu model in interaction evaluation of ubiquitous systems. In *Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems*, IHC '20, New York, NY, USA. Association for Computing Machinery.

[Duran et al., 2022] Duran, R., Zavgorodniaia, A., and Sorva, J. (2022). Cognitive load theory in computing education research: A review. *ACM Trans. Comput. Educ.*, 22(4).

[Frankfurt, 2005] Frankfurt, H. G. (2005). *On Bullshit*. Princeton University Press.

[Fridman et al., 2018] Fridman, L., Reimer, B., Mehler, B., and Freeman, W. T. (2018). Cognitive load estimation in the wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–9, New York, NY, USA. Association for Computing Machinery.

[Fujiwara and Suzuki, 2020] Fujiwara, T. and Suzuki, S. (2020). Cognitive load-strength estimation for nirs with self-organizing particle filtering. In *2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages 756–761.

[Gavas et al., 2017] Gavas, R., Chatterjee, D., and Sinha, A. (2017). Estimation of cognitive load based on the pupil size dilation. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1499–1504.

[Grossman et al., 2009] Grossman, T., Fitzmaurice, G., and Attar, R. (2009). A survey of software learnability: Metrics, methodologies and guidelines. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 649–658, New York, NY, USA. Association for Computing Machinery.

[Haapalainen et al., 2010] Haapalainen, E., Kim, S., Forlizzi, J. F., and Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, page 301–310, New York, NY, USA. Association for Computing Machinery.

[Hart, 2006] Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908.

[Hart and Staveland, 1988] Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Hancock, P. A. and Meshkati, N., editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, Amsterdam, Netherlands.

[ISO 9241-11:2018, 2018] ISO 9241-11:2018 (2018). Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. Standard, International Organization for Standardization.

[Kelleher and Hnin, 2019] Kelleher, C. and Hnin, W. (2019). Predicting cognitive load in future code puzzles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

[Klingberg, 2010] Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences*, 14(7):317–324.

[Lewis and Wharton, 1997] Lewis, C. and Wharton, C. (1997). Chapter 30 - cognitive walkthroughs. In Helander, M. G., Landauer, T. K., and Prabhu, P. V., editors, *Handbook of Human-Computer Interaction (Second Edition)*, pages 717–732. North-Holland, Amsterdam, second edition edition.

[Lewis and Sauro, 2009] Lewis, J. and Sauro, J. (2009). The factor structure of the system usability scale. In *Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International*, volume 5619, pages 94–103, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Li and De Cock, 2020] Li, X. and De Cock, M. (2020). Cognitive load detection from wrist-band sensors. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp-ISWC '20, page 456–461, New York, NY, USA. Association for Computing Machinery.

[Murata and Suzuki, 2015] Murata, Y. and Suzuki, S. (2015). Artifact robust estimation of cognitive load by measuring cerebral blood flow. In *2015 8th International Conference on Human System Interaction (HSI)*, pages 302–308.

[Pillai et al., 2022] Pillai, P., Balasingam, B., Kim, Y. H., Lee, C., and Biondi, F. (2022). Eye-gaze metrics for cognitive load detection on a driving simulator. *IEEE/ASME Transactions on Mechatronics*, 27(4):2134–2141.

[Rocha et al., 2017] Rocha, L. C., Andrade, R. M. C., Sampaio, A. L., and Lelli, V. (2017). Heuristics to evaluate the usability of ubiquitous systems. In *Distributed, Ambient and Pervasive Interactions: 5th International Conference, DAPI 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings*, page 120–141, Berlin, Heidelberg. Springer-Verlag.

[Rooden, 1998] Rooden, M. (1998). Thinking about thinking aloud. *Contemporary Ergonomics*, pages 328–332.

[Saha et al., 2018] Saha, A., Minz, V., Bonela, S., Sr, S., Chowdhury, R., and Samanta, D. (2018). *Classification of EEG Signals for Cognitive Load Estimation Using Deep Learning Architectures: 10th International Conference, IHCI 2018, Allahabad, India, December 7–9, 2018, Proceedings*, pages 59–68.

[Tomitsch et al., 2018] Tomitsch, M., Wrigley, C., Borthwick, M., Ahmadpour, N., Frawley, J., Kocaballi, A. B., Núñez-Pacheco, C., and Straker, K. (2018). *Design. think. make. break. repeat. A handbook of methods.* Bis Publishers, The Netherlands.

[Tullis and Albert, 2008] Tullis, T. and Albert, W. (2008). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics: Second Edition*, pages 216–218.

[Wedral et al., 2023] Wedral, A., Vrecar, R., Ebenhofer, G., Pönitz, T., Führer, P. H., Weiss, A., and Stübl, G. (2023). Spatial augmented reality in the factory: Can in-situ projections be used to communicate dangers and health risks? In Abdelnour Nocera, J., Lárusdóttir, M. K., Petrie, H., Piccinno, A., and Winckler, M., editors, *Human-Computer Interaction - INTERACT 2023 (Part II), the 19th IFIP TC13 International Conference*, pages 574–596, York, UK. Springer Nature Switzerland, LNCS 14143.

[Yin et al., 2007] Yin, B., Ruiz, N., Chen, F., and Khawaja, M. A. (2007). Automatic cognitive load detection from speech features. In *Proceedings of the 19th Australasian Conference on Computer-Human Interaction: Entertaining User Interfaces*, OZCHI '07, page 249–255, New York, NY, USA. Association for Computing Machinery.

[Zilz, 2011] Zilz, R. (2011). Specifying concurrent behavior to evaluate ubiquitous computing environments. In *Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, EICS '11, page 295–298, New York, NY, USA. Association for Computing Machinery.

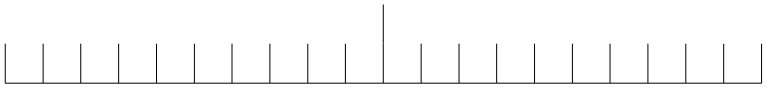
Appendix

NASA-TLX

Bitte beantworten Sie die folgenden Fragen indem Sie die passende vertikale Linie auf der jeweiligen Skala markieren.

Mentale Anforderung

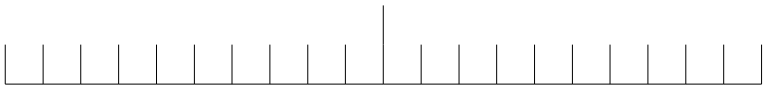
Wie geistig anspruchsvoll war die Aufgabe?



Sehr wenig — Sehr stark

Körperliche Anforderung

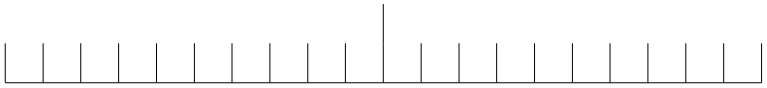
Wie körperlich anspruchsvoll war die Aufgabe?



Sehr wenig — Sehr stark

Zeitliche Anforderung

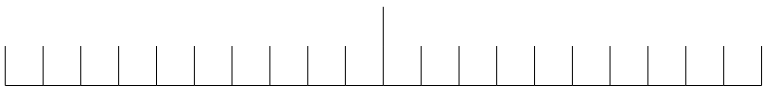
Wie eilig oder gehetzt war das Tempo der Aufgabe?



Sehr wenig — Sehr stark

Leistung

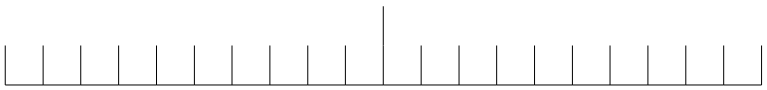
Wie erfolgreich waren Sie darin, das zu erfüllen, was von Ihnen verlangt wurde?



Sehr wenig — Sehr stark

Aufwand

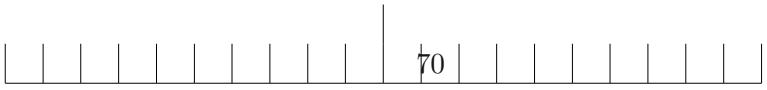
Wie hart mussten Sie arbeiten, um Ihr Leistungsniveau zu erreichen?



Sehr wenig — Sehr stark

Frustration

Wie unsicher, entmutigt, gereizt, gestresst und verärgert waren Sie?



Sehr wenig — Sehr stark

Fragebogen zur System-Gebrauchstauglichkeit

1. Ich denke, dass ich das System gerne häufig benutzen würde.

Stimme überhaupt nicht zu 1	2	3	4	Stimme voll zu 5
○	○	○	○	○

2. Ich fand das System unnötig komplex.

Stimme überhaupt nicht zu 1	2	3	4	Stimme voll zu 5
○	○	○	○	○

3. Ich fand das System einfach zu benutzen.

Stimme überhaupt nicht zu 1	2	3	4	Stimme voll zu 5
○	○	○	○	○

4. Ich glaube, ich würde die Hilfe einer technisch versierten Person benötigen, um das System benutzen zu können.

Stimme überhaupt nicht zu 1	2	3	4	Stimme voll zu 5
○	○	○	○	○

5. Ich fand, die verschiedenen Funktionen in diesem System waren gut integriert.

Stimme überhaupt nicht zu 1	2	3	4	Stimme voll zu 5
○	○	○	○	○

6. Ich denke, das System enthielt zu viele Inkonsistenzen.

Stimme überhaupt nicht zu 1	2	3	4	Stimme voll zu 5
○	○	○	○	○

7. Ich kann mir vorstellen, dass die meisten Menschen den Umgang mit diesem System sehr schnell lernen.

Stimme überhaupt nicht zu 1	2	3	4	Stimme voll zu 5
○	○	○	○	○

8. Ich fand das System sehr umständlich zu nutzen.

Stimme überhaupt nicht zu 1	2	3	4	Stimme voll zu 5
○	○	○	○	○

9. Ich fühlte mich bei der Benutzung des Systems sehr sicher.

Stimme überhaupt nicht zu 1	2	3	4	Stimme voll zu 5
○	○	○	○	○

10. Ich musste eine Menge lernen, bevor ich anfangen konnte das System zu verwenden.

Stimme überhaupt nicht zu 1	2	3	4	Stimme voll zu 5
○	○	○	○	○

List of Figures

List of Tables