

Christoph H. Zimmerer

Automating the Identification of Support Needs in Design: An Eye Tracking-Based Algorithm

Automatische Erfassung von Unterstützungsbedarf im Design: Ein Algorithmus auf Basis von Eye-Tracking

Band 204

Systeme ■ Methoden ■ Prozesse

Univ.-Prof. Dr.-Ing. Dr. h.c. A. Albers
Univ.-Prof. Dr.-Ing. S. Matthiesen
(Hrsg.)

Forschungsberichte



Christoph H. Zimmerer

Automating the Identification of Support Needs in Design: An Eye Tracking-Based Algorithm

Automatische Erfassung von Unterstützungsbedarf im Design: Ein Algorithmus auf Basis von Eye-Tracking

Band 204

Herausgeber Univ.-Prof. Dr.-Ing. Dr. h.c. A. Albers
Univ.-Prof. Dr.-Ing. S. Matthiesen

Copyright IPEK - Institut für Produktentwicklung, 2026
Karlsruher Institut für Technologie (KIT)
Die Universität in der Helmholtz-Gemeinschaft
Alle Rechte vorbehalten

ISSN 1615-8113

Automating the Identification of Support Needs in Design: An Eye Tracking-Based Algorithm

Zur Erlangung des akademischen Grades eines
DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)
von der KIT-Fakultät für Maschinenbau des
Karlsruher Instituts für Technologie (KIT)

angenommene

DISSERTATION

von

M. Sc. Christoph H. Zimmerer

Tag der mündlichen Prüfung:	17.06.2026
Hauptreferent:	Univ.-Prof. Dr.-Ing. Sven Matthiesen
Korreferent:	Prof. Dr.-Ing. Kilian Gericke

Vorwort der Herausgeber

Wissen ist einer der entscheidenden Faktoren in den Volkswirtschaften unserer Zeit. Der Unternehmenserfolg wird mehr denn je davon abhängen, wie schnell ein Unternehmen neues Wissen aufnehmen, zugänglich machen und verwerten kann. Die Aufgabe eines Universitätsinstitutes ist es, hier einen wesentlichen Beitrag zu leisten. In den Forschungsarbeiten wird ständig Wissen generiert. Dieses kann aber nur wirksam und für die Gemeinschaft nutzbar werden, wenn es in geeigneter Form kommuniziert wird. Diese Schriftenreihe dient seit mehr als 20 Jahren als eine Plattform zum Transfer und macht damit das Wissenspotenzial aus aktuellen Forschungsarbeiten am IPEK - Institut für Produktentwicklung Karlsruhe* am Karlsruher Institut für Technologie (KIT) verfügbar. Die Forschung des IPEK ist dabei strukturiert in die Kategorien Systeme, Methoden und Prozesse, um so der Komplexität heutiger Produktentwicklung ganzheitlich gerecht zu werden. Erst die Verknüpfung dieser drei Kategorien ermöglicht die Synthese innovativer Systeme durch Nutzung neuester Methoden und Prozesse. Gleichzeitig werden durch die Systemsynthese die erforschten neuen Methoden und Prozesse validiert und deren Mehrwert für die Praxis abgesichert. Dieses Forschungskonzept prägt nicht nur das IPEK-Leitbild, sondern auch den Charakter dieser Schriftenreihe, da immer alle drei Kategorien und deren Wechselwirkungen berücksichtigt werden. Jeder Band setzt hier individuelle Schwerpunkte und adressiert dabei folgende Forschungsgebiete des IPEK:

- das Entwicklungs- und Innovationsmanagement,
- die Entwicklungs- und Konstruktionsmethodik,
- der Leichtbau von der Ebene des ganzen Systems bis hinunter zur Optimierung des Bauteils,
- die Validierung technischer Systeme auch unter Berücksichtigung der NVH Aspekte (Noise, Vibration, Harshness) mit dem Fokus auf Schwingungen und Akustik an Komponenten und in den Gesamtsystemen sowie deren subjektiver Beurteilung durch den Menschen,
- die Antriebssystemtechnik mit den Schwerpunkten komplette Antriebslösungen für Fahrzeuge und Maschinen,
- das Design, die Tribologie und Erprobung von Kupplungen und Bremsen sowie
- die Gerätetechnik mit dem Schwerpunkt auf Power-Tools.

Die Forschungsberichte stellen Ergebnisse unserer Forschung sowohl anderen Wissenschaftlern als auch den Unternehmen zu Verfügung, um damit die Produktentwicklung in allen ihren Facetten mit innovativen Impulsen zu optimieren.

Albert Albers und Sven Matthiesen

* Eh.: Institut für Maschinenkonstruktionslehre und Kraftfahrzeugbau, Universität Karlsruhe (TH)

Vorwort zu Band 204

Entwicklungsmethoden haben das Ziel, den Menschen bei der Entwicklung von Produkten zu unterstützen. Sie werden von einer großen wissenschaftlichen Community erarbeitet und bereitgestellt. Dabei ergeben sich zwei zentrale Herausforderungen:

Erstens müssen die angebotenen Methoden genau dann wirksam unterstützen, wenn der produktentwickelnde Mensch Schwierigkeiten hat.

Zweitens muss nachgewiesen werden können, dass diese Methoden dem Menschen helfen, die bestehenden Schwierigkeiten zu überwinden.

Immer dann, wenn der Mensch Teil des zu untersuchenden Systems ist, werden klare, deterministische Aussagen schwierig, da der Unterstützungsbedarf zwischen einzelnen Personen stark variiert. Entsprechende Forschung erfordert daher eine ausreichend große Zahl an Probanden. Dadurch entstehen große Studien mit umfangreichen Daten, deren Auswertung mit oft sehr hohem Aufwand verbunden ist. So erfordert die Auswertung einer Stunde erhobener Studiendaten etwa 25 Stunden Analyseaufwand.

An diesem Punkt setzt die Arbeit von Christoph Zimmerer an. Er untersucht, wie der Unterstützungsbedarf von Produktentwicklungsingenieurinnen und -ingenieuren automatisiert erfasst werden kann – mit dem Ziel, einerseits den Zeitpunkt des Auftretens von Unterstützungsbedarf zu erkennen und andererseits Methoden validieren zu können, die diesen Unterstützungsbedarf reduzieren. Zur Identifikation des Unterstützungsbedarfs entwickelt er einen Algorithmus, der automatisiert auf Basis von Eye-Tracking-Daten arbeitet. Die Qualität des Auswerteverfahrens bewertet Christoph Zimmerer in einer explorativen Studie, indem er die automatisiert erhobenen Ergebnisse mit den von den Probanden selbst berichteten Schwierigkeiten vergleicht.

Die Ergebnisse zeigen, dass Eye Tracking in Kombination mit maschinellem Lernen eine skalierbare und objektive Alternative zu klassischen Methoden darstellt. Sie ermöglicht größere Stichproben, reduziert den manuellen Auswerteaufwand und verbessert die Vergleichbarkeit von Studien. Damit leistet die Arbeit einen wichtigen Beitrag zur Entwicklung effizienter und belastbarer Forschungsmethoden in der Konstruktionsforschung.

Juni 2026

Sven Matthiesen

Kurzfassung

Das Identifizieren von Unterstützungsbedarf von Konstrukteurinnen und Konstrukteuren im Konstruktionsprozess stellt in der Konstruktionsforschung eine zentrale Herausforderung für die Entwicklung und Validierung von Konstruktionsmethoden dar.

Derzeit wird dieser Unterstützungsbedarf überwiegend mit etablierten Forschungsmethoden wie Protokollanalysen oder Interviews erhoben. Diese Forschungsmethoden erfordern jedoch einen enormen Zeit- und Arbeitsaufwand in der Datenauswertung aufgrund manueller Kodierungsschritte. Dieser Aufwand wird durch Maßnahmen zur Steigerung der Objektivität, wie beispielsweise die Interrater Reliabilität, noch weiter gesteigert. Das Problem dabei ist, dass der damit verbundene Ressourceneinsatz die Größe empirischer Studien erheblich begrenzt und somit den Nachweis statistischer Zusammenhänge erschwert. Um diese Einschränkung zu überwinden, sind neue Forschungsmethoden und -werkzeuge erforderlich, die es ermöglichen, ressourcenintensive Datenauswertung zu reduzieren, die Objektivität zu erhöhen und damit groß angelegte, statistisch belastbare Studien mit Konstrukteuren durchzuführen.

Das Ziel dieses Forschungsvorhabens ist die Entwicklung eines Algorithmus zur teilautomatisierten und objektiveren Erfassung von Schwierigkeiten als Indikator für methodischen Unterstützungsbedarf im Konstruktionsprozess. Dazu werden Eye-Tracking-Daten mit retrospektiven Selbstauskünften der Teilnehmenden kombiniert, um Schwierigkeiten als Indikatoren für Unterstützungsbedarf systematisch zu erfassen.

Im ersten Schritt werden im Rahmen einer explorativen Studie systematisch Selbstauskünfte von Probanden über auftretende Schwierigkeiten ausgewertet und kategorisiert. Dabei können insgesamt acht Kategorien von Schwierigkeiten identifiziert werden. Diese Kategorien bilden die Grundwahrheit für die Entwicklung eines Algorithmus zur Erforschung auftretender Schwierigkeiten.

Auf dieser Grundlage erfolgen die Entwicklung und Evaluation eines Algorithmus, der die identifizierten Schwierigkeiten automatisch anhand von Eye-Tracking-Daten erkennen kann. Hierfür wurden die zuvor identifizierten Schwierigkeiten mit Eye-Tracking-Daten der Probanden verknüpft. Auf dieser Basis wurden verschiedene Ansätze des maschinellen Lernens angewendet und die Ergebnisse mit den Selbstauskünften der Probanden verglichen. Der leistungsfähigste Ansatz erreichte eine Genauigkeit von 62 % bei einer Fehlklassifikationsrate von 26,6 % für nicht erkannte Schwierigkeiten. Es konnte gezeigt werden, dass unterschiedliche Typen von Schwierigkeiten mit unterschiedlichen Blickparametern zusammenhängen. Folglich stellt die automatisierte Detektion von Schwierigkeiten als Indikatoren für Unterstützungsbedarf durch einen Algorithmus eine vielversprechende Ergänzung klassischer Forschungsmethoden in der Konstruktionsforschung dar.

Ziel dieser Arbeit war es, eine teilautomatisierte und objektivere Erfassung von Schwierigkeiten im Konstruktionsprozess als Indikatoren für methodischen Unterstützungsbedarf zu ermöglichen. Dieses Ziel konnte durch den entwickelten Algorithmus erreicht werden. Die Ergebnisse zeigen, dass ein Algorithmus auf Basis von Eye-Tracking eine skalierbare und objektive Ergänzung zu klassischen Methoden darstellt. So kann der manuelle Auswertungsaufwand reduziert, die Vergleichbarkeit von Studien verbessert und letztlich größere Stichproben ermöglicht werden. Damit leistet die Arbeit einen entscheidenden Beitrag zur Skalierbarkeit und statistischen Absicherung in der Konstruktionsforschung.

Abstract

In design research, identifying the support needs of engineering designers is a key challenge for the development and validation of design methods.

Currently, this need for support is mainly assessed using established research methods such as protocol analyses or interviews. However, these methods require an enormous amount of time and effort to evaluate the data due to manually performed coding steps. This effort is further increased by activities to enhance objectivity, such as interrater reliability. The problem here is that the associated effort significantly limits the size of empirical studies and thus makes it difficult to prove statistical correlations. To overcome these limitations, new research methods and tools are required that make it possible to reduce resource-intensive data evaluation, increase objectivity, and thus conduct methodologically robust studies involving engineering designers with sample sizes large enough to yield statistically significant and reliable results.

The aim of this thesis is the development of an algorithm for the semi-automated and more objective assessment of difficulties as an indicator of methodological need for support in the design process. To this end, eye-tracking data is combined with retrospective self-reports from participants to systematically record difficulties as indicators of need for support.

First, an exploratory study systematically evaluates and categorizes self-reported difficulties provided by participants. A total of eight clusters of difficulties can be identified. These categories serve as the ground truth for developing an algorithm to identify occurring difficulties.

On this basis, an algorithm is developed and evaluated that can automatically detect the identified difficulties using eye tracking data. To this end, the previously identified difficulties are linked to the participants' eye tracking data. Various machine learning approaches are tested and the results compared with the self-reported difficulties provided by the participants. The most powerful approach achieves an accuracy of 62 % with a false negative rate of 26.6 % for unrecognized difficulties. It can be shown that different types of difficulties are associated with different eye tracking parameters. This demonstrates that the automated detection of difficulties as indicators of support needs using an algorithm is a promising addition to traditional research methods in design research.

The aim of this work was to enable automated and more objective assessment of difficulties as indicators for the need of support during the design process. This aim was achieved with the developed algorithm. The results show that an algorithm based on eye tracking is a scalable and objective addition to traditional methods and tools.

This reduces the amount of manual evaluation effort, improves the comparability of studies, and finally enables larger samples. This work thus makes a decisive contribution to scalability and statistical reliability in design research.

Table of Contents

- Kurzfassung.....i
- Abstract.....iii
- List of Figures.....ix
- List of Tables.....xi
- Acronyms.....xiii
- 1 Introduction.....1
 - 1.1 Focus of this Thesis.....2
 - 1.2 Structure of this Thesis.....3
- 2 State of Research.....7
 - 2.1 Designer's Support Needs.....7
 - 2.1.1 The Role of Conceptual Design in the Design Process.....8
 - 2.1.2 Difficulties of Professional Designers During Conceptual Design 11
 - 2.1.3 Design Support.....13
 - 2.1.4 Individual Need for Support 15
 - 2.1.5 Empirical Assessment of Support Needs and Sampling Challenges.....17
 - 2.1.6 Physiological Signals in Design Research 19
 - 2.1.7 Interim Conclusion.....21
 - 2.2 Cognitive Load 22
 - 2.2.1 Cognitive Load Theory.....23
 - 2.2.2 Measure Cognitive Load.....25
 - 2.2.3 Further Containment Based on own Preliminary Work: Cognitive Load as Indicator for Need of Support 28
 - 2.2.4 Interim Conclusion.....37
 - 2.3 Automated Data Collection and Analysis of Cognitive Load 37
 - 2.3.1 Eye Tracking to Measure Cognitive Load 38
 - 2.3.2 Further Containment Based on own Preliminary Work: Measure Cognitive Load in Design via Eye Tracking 41
 - 2.3.3 Basics of Machine Learning.....51
 - 2.3.4 Machine Learning Algorithms Using Eye Tracking to Measure Cognitive Load.....53
 - 2.3.5 Interim Conclusion.....56
 - 2.4 Conclusion on the State of Research 56

3	Motivation and Research Objectives	59
3.1	Motivation	59
3.2	Research Objectives.....	61
4	Research Questions and Research Design.....	63
4.1	Research Questions	63
4.2	Research Design	64
5	Difficulties in Concept Synthesis.....	67
5.1	Methodology	68
5.1.1	Experimental Design	69
5.1.2	Data Analysis	74
5.2	Results.....	74
5.2.1	Difficulties in Detail	75
5.2.2	Clusters of Difficulties.....	76
5.3	Discussion	78
5.3.1	Difficulties of Engineering Design Students	78
5.3.2	Comparison to Experienced Designers	79
5.3.3	Implications	81
5.4	Interim Conclusion	82
6	An Algorithm to Detect Difficulties in Concept Synthesis	85
6.1	Methodology	86
6.1.1	Dataset	86
6.1.2	Algorithm Development.....	88
6.1.3	Scoring Metric	94
6.2	Results.....	95
6.2.1	Parameter Study	96
6.2.2	Binary Classification.....	97
6.2.3	Multibinary Classification.....	99
6.2.4	Summary of Results.....	103
6.3	Discussion	104
6.3.1	Detection Difficulties via Eye Tracking	104
6.3.2	Tailored Feature Selection	105
6.3.3	Algorithm for Automated Evaluation	106
6.3.4	Implications for the Field of Design Research	107
6.4	Interim Conclusion	107
7	Summary and Outlook.....	111
7.1	Summary	111
7.2	Outlook	114

References I

Publications Co-authored by the Author of this Thesis XVII

Use of Tools for assisted Copy Editing XIX

List of Figures

Figure 1.1: Structure of the thesis	4
Figure 2.1: Conceptual design process adapted from Andreasen et al. (2015) ..	10
Figure 2.2: Required sampling sizes for different study-types adapted from Cash et al. (2022).....	18
Figure 2.3: Results of the evaluation of cognitive load using the NASA-TLX questionnaire for the various tasks (n = 12)	32
Figure 2.4: Results of the evaluation of performance for the various tasks	33
Figure 2.5: Correlation between cognitive load values and performance across all tasks and participants displayed as a negative trending line	35
Figure 2.6: Design tasks used in the preliminary work	44
Figure 2.7: Parameter study according to Halverson et al. (2012).	54
Figure 4.1: Research design	65
Figure 5.1: Design Task 1.....	71
Figure 5.2: Design Task 2.....	72
Figure 5.3: Design Task 3.....	73
Figure 5.4: Distribution of the number of difficulties occurring per participant	75
Figure 6.1: Distribution of the 439 difficulties reported by the participants (n = 28) in the data set	87
Figure 6.2: Algorithm training process: data preparation and classification	89
Figure 6.3: Schematic procedure for segmentation.....	90
Figure 6.4: Functionality of a binary classifier	92
Figure 6.5: Graphical comparison of the multibinary approaches one-against-all (OAA) and cascading one-against-all (cOAA)	94
Figure 6.6: Calculation of the f1-score and fn-score.....	95
Figure 6.7: Results of the parameter study	97
Figure 6.8: Highest possible classification result for each class of difficulty using the most appropriate classifiers	99

List of Tables

Table 2.1: Differentiation between various forms of design support.....13

Table 2.2: Description of design tasks for answering PRQ1 29

Table 2.3: Eye tracking metrics and tasks investigated in various studies from different fields 38

Table 2.4: Participant demographics.....42

Table 2.5: Overview of the cognitive load from the NASA-TLX and the mean values of all parameters per task 47

Table 2.6: Results of the correlation analysis between cognitive load and the parameters considered. Here, both tasks were considered.....48

Table 2.7: Results of the statistical analysis regarding differences between the tasks used and the parameters considered.....49

Table 5.1: Results of the evaluation of the participants' statements. The difficulties encountered can be divided into eight different clusters. 76

Table 6.1: Comparison of the tested classifiers and their achieved scores 98

Table 6.2: Used Feature OAA.....101

Table 6.3: Change in the individual classification results due to the cascading one-against-all approach 102

Table 6.4: Final result of the three classification approaches measured by f1-score and fn-score as well as the associated standard deviations. ...103

Acronyms

PRQ	Preliminary Research Question
ORQ	Overarching Research Question
RQ	Research Question
DRM	Design Research Methodology
CLT	Cognitive Load Theory
CL	Cognitive Load
MCH	Modified Cooper-Harper Scale
OWS	Overall Workload Scale
SWAT	Subjective Workload Assessment Technique
NASA-TLX	NASA Task Load Index
ET	Eye Tracking
PD	Pupil Diameter
FD	Fixation Duration
BD	Blink Duration
SR	Saccadic Rate
PERCLOS	Percentage Pupil Closure
EDA	Electrodermal Activity
ECG	Electrocardiogram
HRV	Heart Rate Variability
EEG	Electroencephalography
GSR	Galvanic Skin Response
fNIRS	Functional Near Infrared Spectroscopy
fMRI	Functional Magnetic Resonance Imaging
HCI	Human-Computer Interaction

BC	Binary Classification
OAA	One-Against-All
cOAA	Cascading One-Against-All

1 Introduction

Support in design research must be oriented around the engineering designers (Ehrlenspiel & Meerkamm, 2013). Their benefits can only be fully realized if they are made available to engineering designers according to their individual needs. The central added value of design research lies in the fact that it offers design support where actual difficulties arise in the design process. In order to identify these individual support needs, precise data collection in empirical participant studies is required to show when and where engineering designers encounter difficulties during the design process.

Traditional research methods such as protocol analysis or interviews focus on verbal and behavioral data to explain reasoning and decision processes (Dorst & Cross, 2001; Ericsson & Simon, 1993). However, they require extensive coding and interpretation, making them highly resource-intensive. In order to ensure the quality of the findings obtained by these research methods, the quality criteria of empirical social sciences - objectivity, reliability, and validity - must be met (Hussy et al., 2013). Ensuring objectivity in particular, as a prerequisite for reliability and validity, increases this effort even further, as it can only be improved by multiple coders during evaluation (Cash et al., 2012). This high level of effort during evaluation is an important factor limiting scalability and conducting larger participant studies to prove statistical correlations in design research. As a result, many participant studies remain small in scale and offer only limited opportunities to identify statistically reliable results in design research (Cash et al., 2022; Dinar et al., 2015).

This presents a key challenge for design research: the high effort required for evaluation not only limits the number of participants investigated but also restricts the ability to generalize findings and systematically identify design support needs. While studies with limited participant numbers are well suited to investigating designers' detailed procedures and formulating hypotheses about design behavior, existing research approaches remain limited in their ability to generate results that are both statistically significant and empirically reliable. To overcome this limitation, approaches are needed that enable the continuous and objective assessment of designers' cognitive processes during design activities and the needs that arise during these activities.

An approach to analyzing cognitive processes in design is cognitive load. Cognitive load reflects the mental effort required to process information and solve problems,

and thus provides an indirect indicator of moments in which designers experience difficulties (Paas et al., 2003; Sweller et al., 2019). To capture cognitive load, physiological signals can be used as indicators for continuous and objective evaluation.

Physiological signals such as eye tracking, heart rate variability (HRV), and galvanic skin response (GSR) enable such objective measurement and allow continuous data collection with minimal influence on participants (Lohmeyer & Meboldt, 2016). Eye tracking seems particularly suitable, as visual perception plays a key role in the design process and the engineering designer's eyes act as a source of information (Ullman, 2010). Initial studies have shown that gaze patterns can be indicators of cognitive load and comprehension problems (Bi & Reid, 2017). However, the lack of standardized evaluation routines limits comparability between studies and prevents the full exploitation of methodological benefits.

In summary, current research methods in design research are limited by high effort for evaluation, limited objectivity, and a lack of scalable, continuous approaches. Overcoming these limitations is essential for identifying support needs objectively and at scale.

1.1 Focus of this Thesis

The focus of this thesis will be on the conceptual design phase. The main activities in the conceptual design phase include analyzing the function, searching for suitable functional principles, and combining them to develop viable concepts. The most suitable solution is selected by evaluating several variants. (Pahl et al., 2007) Within this process, concept synthesis represents the central activity in which these principles are gradually concretized into a final product design (Andreasen et al., 2015). As described by Andreasen et al. (2015), concept synthesis brings together the *“idea with the product”* - such as user needs, market positioning, and intended product value - with the *“idea in the product”* - such as function and embodiment (Andreasen et al., 2015). VDI Guideline 2222 describes it as follows: “A wrong decision in the early conceptual design phase can generally no longer be compensated for in a subsequent phase or work stage, no matter how far-reaching the optimization” (Verein Deutscher Ingenieure e.V., 1997, p. 11). Ullman (2010) also points out that around 70 % of the subsequent product costs are already determined at the end of the conceptual design phase, even though only around 10 % of the development costs have actually been incurred.

In this phase, numerous requirements must be taken into account simultaneously. Design engineers are often faced with conflicting goals between functional, technological, economic, and design aspects, which must be translated into a consistent design concept (Wittel et al., 2013). In addition, there are uncertainties regarding technical feasibility, costs, and market acceptance of the solution concepts, meaning that decisions have to be made with limited information.

For this reason, the thesis focuses on the concept synthesis phase and the difficulties that arise during this phase as the central field of investigation. In this phase, key decisions are made that have a decisive influence on the subsequent course of the product development process. At the same time, individual difficulties and support needs become particularly apparent here. Consequently, it is of special interest to systematically identify these needs and to investigate suitable methodological support in order to effectively assist engineering designers during this critical stage.

In combination with the problem of current research methods regarding effort for data collection and analysis, this leads to the following aim: to develop an algorithm that enables an automatable evaluation of support needs in empirical participant studies. For this purpose, eye tracking data will be used to capture difficulties as indicators of support needs during design synthesis.

This leads to the overarching research question:

How can eye tracking data be used to objectively and automatically measure support needs in the form of difficulties in concept synthesis as part of the design process?

1.2 Structure of this Thesis

This thesis is divided into seven chapters. The structure of the thesis is shown in Figure 1.1.

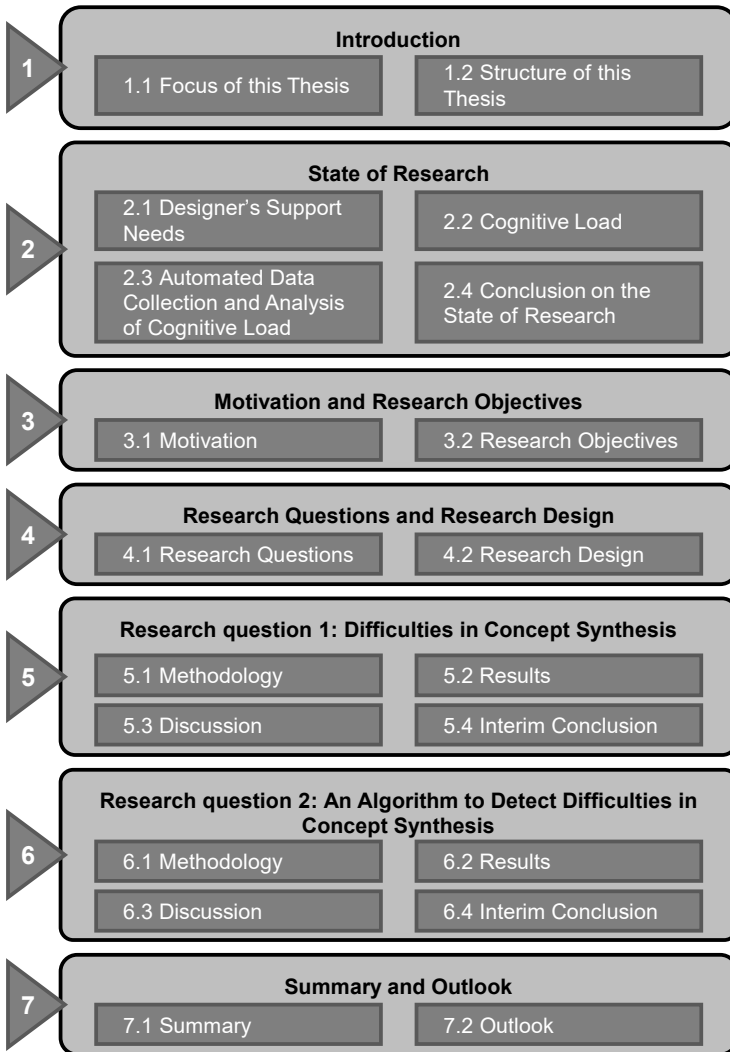


Figure 1.1: Structure of the thesis

Following this chapter, chapter 2 presents the relevant state of research. It is divided into three sub-chapters: the positioning of the conceptual design phase within the design process and the current challenges arising from design research (chapter 2.1), an introduction to cognitive load theory and established survey methods (chapter 2.2), and an overview of approaches to automating data collection and evaluation using physiological signals (chapter 2.3). For the latter two chapters, preliminary work is also presented to further specify the scope of this thesis.

Based on the current state of research, chapter 3 describes the motivation and research objectives. Chapter 4 then presents the research questions and research design to achieve the objectives.

Chapter 5 presents the study conducted to answer the first research question. A qualitative study provides the data basis for the subsequent algorithm development. To this end, difficulties arising in the conceptual design phase are identified using retrospective think aloud protocols and subsequently clustered.

In chapter 6, various approaches to the automated detection of difficulties are presented in order to answer research question 2. The difficulties identified in the first research question are combined with eye tracking data and several automated detection algorithms are developed using machine learning approaches.

Finally, the results of this work are summarized in chapter 7 and an outlook on future research topics for continuing this research is provided.

2 State of Research

This chapter presents the state of research relevant for this work and introduces the essential theoretical foundations. First, the role of design support as an integral part of the product development process is discussed, along with the individual designer and his needs within this process.

Building on this, research methods for investigating designers in conceptual design are examined, with a particular focus on measuring cognitive load to identify support needs.

Subsequently, an overview is provided of approaches and possibilities arising from the use of physiological signals in design research, particularly concerning the use of eye tracking data for automation of data collection and analysis.

Finally, based on the current state of research and prior work, the existing research gap is identified and the problem addressed in this thesis is formulated.

2.1 Designer's Support Needs

The role of conceptual design within the overall design process is outlined in Chapter 2.1.1, followed by a summary of difficulties experienced by professional designers during this phase (Chapter 2.1.2).

In the following, chapter 2.1.3 provides an overview of existing types of design support that aims to support engineering designers during design. Chapter 2.1.4 then addresses the individual need for support in design and discusses its relevance for the development and evaluation of design methods.

Chapter 2.1.5 describes the impact of existing research methods on empirical studies in design research, in particular regarding their potential for identifying support needs. Chapter 2.1.6 presents physiological signals as a data source in design research, highlighting their potential for identifying the need of support.

Finally, chapter 2.1.7 provides an interim conclusion that summarizes the key points and establishes the link to the subsequent chapters.

2.1.1 The Role of Conceptual Design in the Design Process

The conceptual design phase plays a central role in product development. It marks the transition from creative, often unstructured ideation to the systematic development of technical solutions. In this phase, key decisions are made that have far-reaching effects on the further course of development. The decisions made in the conceptual design phase have a significant impact on the entire subsequent process and the overall balance of function, embodiment, and costs.

VDI guideline 2222 states: 'A wrong decision in the early conceptual design phase can generally no longer be compensated for in a subsequent phase or work stage, no matter how far-reaching the optimization' (Verein Deutscher Ingenieure e.V., 1997). According to Ullman (2010), around 70 % of the later product costs have already been determined by the end of the conceptual design phase, although only around 10 % of the development costs have actually been incurred (Ullman, 2010). This finding emphasizes that the quality of content and structural consistency of the conceptual design phase are decisive factors in determining how efficiently and effectively the subsequent phases of the development process can proceed. The need for methodological support in this phase is therefore particularly high in order to be able to make well-founded and sustainable decisions.

To categorize the conceptual design phase within the design process, there are a number of models in the literature that divide the design process into structured phases. An established process model can be found in Pahl et al. (2007) and in a comparable form in Ehrlenspiel and Meerkamm (2017) and Dieter and Schmidt (2013). These authors typically divide the design process into four successive main phases:

- 1) Planning and task clarification,
- 2) conceptual design,
- 3) embodiment design and
- 4) detailed design.

The planning and task clarification phase forms the basis for the design process. Here, the task is analyzed and structured by collecting all requirements and boundary conditions. The result of this phase is a list of requirements that systematically guides the design process and is continuously updated. A thorough clarification of the task is essential to avoid errors during the subsequent process and to enable focused conceptual design. (Pahl et al., 2007)

Based on the clarified task, the development of a basic solution takes place in conceptual design. The main tasks are analyzing the function, searching for suitable functional principles and combining them to create viable concepts. The most suitable solution is selected by evaluating several variants. As subsequent adjustments to decisions on principle solutions are only possible with considerable effort, the conceptual design is a key part of the whole design process. (Pahl et al., 2007)

During embodiment design, the selected concepts are further concretized and elaborated from a technical perspective. Product design, embodiment and initial material and production specifications are defined. During this phase, a preliminary layout is created, followed by a definitive layout in which the solution is finalized in detail. Systematic evaluation steps ensure technical and economic feasibility and reduce risks in the subsequent development stages. (Pahl et al., 2007)

Finally, all product details are finalized during detailed design. Geometries, materials, production, and tolerance specifications are finalized and transferred into production-ready documentation. Errors in this phase can have a significant impact on the quality, costs, and marketability of the product. At the same time, there is still room for final optimization of individual components and assemblies. (Pahl et al., 2007)

A phase model similar in its core structure originates from Andreassen et al. (2015), who divides the design process into five activities:

- 1) Exploration,
- 2) Concept Synthesis,
- 3) Product Synthesis,
- 4) Product Development, and
- 5) Product Life Synthesis.

In contrast to the approach of Pahl et al. (2007), who see a concept as the basis for a realizable design, Andreassen et al. (2015) sees the concept itself as the 'complete' solution in a preliminary state which demonstrates the relevance of concept synthesis.

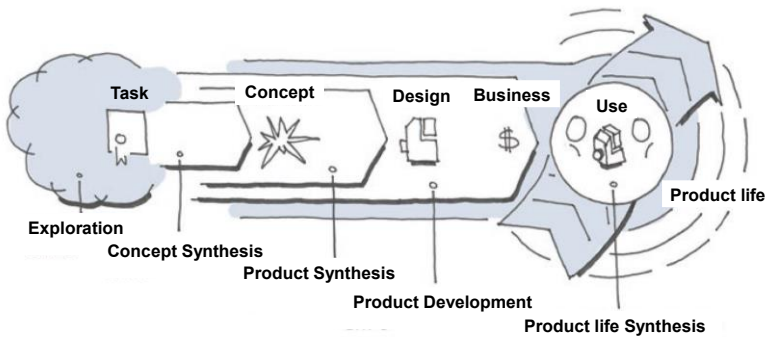


Figure 2.1: Conceptual design process adapted from Andreasen et al. (2015)

For the activities to develop a concept and the methodological support required here, this means that not only design methods for creativity and the development and evaluation of suitable principles of action are required, but also tools for thinking ahead about the embodiment design and detailed design.

However, concept synthesis represents a key challenge in practice. Weisbrod and Kroll (2018) and Tan et al. (2017) emphasize the importance of this phase, where the initial creative openness of conceptual design is increasingly restricted by requirements in terms of function, manufacturing, costs, sustainability, and other framework conditions (Ullman, 2010). Decisions made at this stage, such as the choice of materials, the manufacturing process or the product topology, can often only be revised with considerable effort (Hsu & Liu, 2000).

Pahl et al. (2007) illustrate this with the statement: “Conceptual design results in the specification of a principle solution (concept). Often, however, a working structure cannot be assessed until it is transformed into a more concrete representation.” (Pahl et al., 2007, p. 131) This statement emphasizes the difficulty of being able to conclusively assess the viability of a concept before it is developed - an uncertainty that must be methodically overcome in the product synthesis phase.

This makes it particularly challenging by the fact that a large number of requirements have to be considered simultaneously. Wittel et al. (2013) point out that engineering designers are typically confronted with multiple conflicting goals in this phase: Functional, technological, economic, and design requirements are often in tension with each other and must nevertheless be transferred into a consistent product concept (Wittel et al., 2013). In addition, there is often uncertainty regarding the

technical feasibility, cost development, and market acceptance of the solution concepts in early development phases. The decision in favor of a particular concept variant is therefore made under conditions of limited information, which increases the risk of downstream adaptation costs.

From a methodological point of view, this results in a clear need for structuring, evaluating, and ensuring approaches that enable a well-founded further development of the concept ideas, particularly in the concept synthesis phase. The aim of methodological support must be to accompany the transition from the abstract solution principle to the concrete technical design in such a way that both the requirements of the market and the technical-functional requirements can be systematically taken into account. The models of Pahl et al. and Andreasen provide a suitable basis for this, with Andreasen's model being used in particular in this study. This makes it possible to specifically analyze and further develop methodological support in the phase of concept synthesis.

Design methods and tools of design methodology offer targeted support here: they help to systematically record requirements, explore solution spaces in a structured manner, evaluate concepts comparatively and identify risks at an early stage. However, research shows that design methods are often applied only in very specific situations rather than on a regular basis (Birkhofer et al., 2002). Moreover, it has been observed that there is no consistent use of design methods in practice (Birkhofer et al., 2002; Gericke et al., 2017). There is therefore a need for methodological support in the conceptual design phase, both in terms of generating and evaluating concepts and in terms of decision-making under uncertainty.

To summarize, it can be said that the conceptual design phase, in particular the concept synthesis, plays a critical role in the design process. It is decisive for the technical and economic viability of the product to be developed and is characterized by a high level of complexity and a large number of simultaneous requirements. Due to the scope of the decisions made here, it is essential to provide individualized methodological support to be able to develop well-founded, realizable, and high-quality concepts.

2.1.2 Difficulties of Professional Designers During Conceptual Design

The difficulties faced by designers in practice are profound and complex, as described by Jonassen et al. (2006). Design tasks are characterized as unstructured and complex as they involve conflicting goals, different solution methods, non-technical standards of success, unforeseen problems, and distributed knowledge.

This diversity of requirements makes the effectiveness of design processes a crucial factor that is significantly influenced by the expertise of the designers. (Jonassen et al., 2006)

The experience of designers plays a central role in this. With in-depth knowledge of their specific domain based on many years of experience, they are able to successfully apply specific design strategies. The expertise of experienced designers spans years of practical work in their specific domain, giving them a unique insight and in-depth understanding.

An important limitation that influences the expertise of designers is domain dependency. Experienced designers can work effectively and master complex difficulties in the domain in which they work. However, switching to a different domain poses a particular difficulty. In such cases, designers can no longer use most of their experience. Ericsson and Lehmann (1996) emphasize that the performance of experienced designers is often domain-specific and not easily transferable between domains (Cross, 2005).

The importance of suitable design methods in practice becomes particularly clear in this context. When designers switch to a “foreign domain”, they are faced with the challenge that a large part of their previous experience can no longer be used. The need for design methods becomes even more urgent in such situations. This is also emphasized by Cross (2005), who points out that the adaptation of designers to new domains requires a methodical approach. However, this contrasts with the low application rate of design methods in practice (e.g. Daalhuizen, 2014) even by experienced designers (Dorst, 2008). Automated research methods based on quantitative data is needed to identify the need for support of engineering designers.

Doellken et al. (2020) conducted a study examining difficulties faced by experienced designers in the context of production-oriented design. This study focused specifically on a manufacturing process that was not part of the engineering designer's core expertise. By focusing on the sheet metal manufacturing process, which was not part of the common experience of the participating designers, four central difficulties were identified: First, experienced designers struggled with a lack of know-how regarding the chosen manufacturing process, particularly in the area of sheet metal. Secondly, the focus was on synthesizing a structure in which the design objective had to be reconciled with the manufacturing requirements. Thirdly, analyzing the manufacturability of existing concepts was a challenging task. Fourthly, the lack of knowledge about relevant parameters for the evaluation of concepts to assess the manufacturing effort was emphasized. However, this

assessment is essential to ensure a decision on the final design considering the manufacturing requirements.

2.1.3 Design Support

Design support aims to improve both the effectiveness and efficiency of the design process by providing systematic support that structures thought processes, guides information flows, and informs decisions. Design support contributes to this by offering structured approaches that guide the solution process (Ehrlenspiel & Meerkamm, 2017, p. 173). In design research, the term "support" is understood comprehensively: It encompasses all means, measures and tools that serve to improve the design process, from strategic approaches and methodological procedures to software-based tools (Blessing & Chakrabarti, 2009, p. 4).

Various forms of design support can be found in the literature, which differ in terms of their objectives, scope of application, and formalization. The design research methodology (Blessing & Chakrabarti, 2009, p. 142) distinguishes between design methodologies, design methods, design guidelines, and design tools. A differentiation between the various types of design support is provided in Table 2.1 according to Gericke et al. (2017).

Table 2.1: Differentiation between various forms of design support adapted from Gericke et al. (2017)

design methodology	Design methodologies are comprehensive process models that structure the entire design process. They define phases, activities, their sequence, and the design methods and information flows used. They also reflect fundamental ways of thinking and objectives in design.
design method	Design methods describe specifically how a particular task can be handled in the design process to achieve the desired objective. They define inputs, processes, tools, presentation options, and provide targeted support for individual steps such as finding principles or evaluating variants.

design guideline	Design guidelines are recommendations for how to proceed in certain situations. Deviations from guidelines should be well justified.
design tool	Design tools are means of implementing methods or guidelines. They can be software-supported or physical and are used to support specific tasks, for example through simulation, visualization, or analysis.

The categories described above distinguish between different types of support in the design process. In the following, the focus is on design methods, as they are used in a situation-specific and goal-oriented manner to deal with specific tasks in the design process.

Since every design task has specific requirements, the selection of suitable design methods always depends on the context. Hsu and Liu (2000) and Albers (2010) emphasize the uniqueness of each design process and thus underline the need for flexible, situation-adapted application of design methods (Albers, 2010; Hsu & Liu, 2000).

Despite the large number of design methods available, their actual use is highly dependent on the specific application situation and individual factors. Birkhofer et al. (2002) emphasize that the successful use of design methods is not guaranteed by their description or availability alone. Rather, factors such as the type of task, the experience and knowledge of the engineering designers, the corporate culture, and the organizational context play a central role. (Birkhofer et al., 2002)

Especially in the conceptual design phase, design methods play a central role by providing structured approaches for creative, analytical, and evaluative activities. In this phase, the focus is on different activities, for each of which specific design methods are available. Cross (2005) names central tasks such as ideation, decision making, analysis, and embodiment design (Cross, 2005).

Creativity techniques such as brainstorming or morphological boxes are used to generate ideas. Decision-making is supported by design methods such as decision matrices or utility value analyses. Analytical methods, for example for requirements analysis or functional structuring, help with the systematic recording and evaluation of technical relationships. In the embodiment design phase, design methods such as the contact and channel approach (C&C²-A) (Matthiesen, 2021) or the principles for developing working structures according to Pahl et al. (2007) offer structured approaches to finding solutions.

Accordingly, the question of which design method should be used when, how and by whom cannot be answered in general. Daalhuizen (2014) and Dorst (2008) criticize that the influence of the design context and the role of the designer is often neglected in the discussion of design methods (Daalhuizen, 2014; Dorst, 2008). Instead, methodological support should always be considered in conjunction with the respective context: as a flexible instrument that structures thinking and opens up options. However, this requires a deeper understanding of the situation and the actual need for a design method.

Researchers in the field of design research use various methodological approaches to systematically analyze the need for support in the early phases of the design process. These include observational studies in real development environments, qualitative interviews with practitioners, and controlled experiments that simulate typical design situations. The aim of these studies is to gain a deeper understanding of the decision-making behavior of engineering designers, typical sources of error and cognitive requirements in the design process. The knowledge gained will in turn be incorporated into the development and evaluation of design support.

2.1.4 Individual Need for Support

The context specific use of design methods described in the previous chapter illustrates that the effectiveness of methodological support in the design process depends largely on whether and when there is a need for support. Design methods do not unfold their benefits as universally applicable tools, but always in the context of specific requirements and situational challenges that designers are confronted with during the design process (Daalhuizen et al., 2009).

Difficulties often arise due to the large number of simultaneous requirements and goals. In such contexts, common strategies, and routines often fail. This leads to so-called non-routine situations (Daalhuizen et al., 2009), which require increased cognitive effort, as designers are forced to adopt new perspectives and explore previously unexplored solution spaces. In these moments, the question arises as to whether methodological support is required and, if so, what form it should take. The aim must therefore be to better understand the behavior of engineering designers in such situations and to explore how the perceived source of uncertainty can be specifically addressed with suitable design methods (Daalhuizen et al., 2009).

In this context, Birkhofer et al. (2002) emphasize that the focus in the development and implementation of design methods must be more on the individual needs of engineering designers. A central obstacle to the effectiveness of design methods lies in the fact that every designer is individual and therefore plans and implements

the design process differently (Badke-Schaub et al., 2011). Accordingly, the acceptance and knowledge gain in dealing with design methods is also strongly dependent on the individual: Designers with a positive mindset towards design methods learn significantly more about their application than designers with a neutral or negative attitude (Birkhofer et al., 2002).

This shows that it is not only the quality of a design method that is decisive, but in particular its fit with the respective problem context, the current need, and the individual attitude of the engineering designer (Blessing & Chakrabarti, 2009). In the Design Research Methodology (DRM), the identification of such needs is a central element of methodological development: the development of new, designer-centered design methods requires that existing difficulties and needs for support are first systematically recorded and understood (Blessing & Chakrabarti, 2009; Matthiesen et al., 2022).

A key characteristic of these needs is the individuality and context dependency. Engineering designers differ in terms of experience, cognitive strategies, and domain-specific knowledge (Cross, 2004; Ericsson & Lehmann, 1996). While experienced designers can fall back on well-founded routines in their established domain, their need for methodological support increases significantly as soon as they are confronted with new technologies, unfamiliar manufacturing methods or novel fields of application (Cross, 2004). At the same time, it remains largely unexplored how exactly the sources of uncertainty perceived by engineering designers are related to the selection and effectiveness of specific design methods (Daalhuizen et al., 2009).

This underlines both the challenge and the necessity of considering design methods as adaptive, situational instruments. A design method can only be effective if it is offered precisely at the time and in the situation in which there is a real need (Andreasen, 2011; Blessing & Chakrabarti, 2009). As this need is dynamic and often only becomes apparent during the development process, the identification of corresponding needs must also be closely linked to the individual situations and approaches of the engineering designers.

For design research, this results in the task of differentiating the situational, individual, and context-dependent needs for methodological support in order to develop designer-centered methods. This requires not only a deeper understanding of the cognitive processes in design, but also empirical approaches enabling the continuous identification of needs during the design process from the perspective of the engineering designers themselves.

2.1.5 Empirical Assessment of Support Needs and Sampling Challenges

To systematically identify and understand designers' individual needs for support, empirical investigations play a central role in design research. They are essential to evaluate whether design support actually addresses the intended needs and has the desired effect.

For empirical studies of design processes, a large number of established research methods and tools are available in design research, which were mainly adopted from social research. Ahmed (2007) especially names interviews, observations, and protocol analyses as common research methods with which cognitive or creative aspects of the designer's approach can be operationalized and investigated. The choice of research method used largely depends on the underlying research objective and the variables to be recorded, such as problem-solving behavior, creativity or perception of uncertainty. (Ahmed, 2007)

These research methods provide deep insights into the behavior of engineering designers, but are associated with considerable effort and are based on the researchers' interpretation of the data. For example, analyzing one hour of interview material can require up to nine hours of transcription and interpretation work (Ahmed, 2007), while according to Ahmed et al. (2003), analyzing one hour of observation data can take up to 25 hours of follow-up work (Ahmed et al., 2003).

More recent studies also emphasize the high effort involved in data collection and analysis and the resulting limited sample sizes (Dinar et al., 2015; Escudero-Mancebo et al., 2023). As a result, many empirical studies have focused on qualitative studies with small samples. For example, Björklund (2013), Cramer-Petersen et al. (2019) and Nelius et al. (2020) each investigated fewer than 20 participants. Studies with larger sample sizes, such as Kokotovich (2008) with 33 participants, usually evaluate the outcome of the design process and not the process itself.

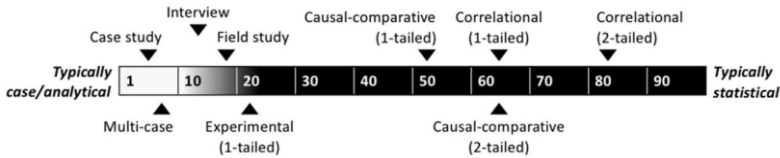


Figure 2.2: Required sampling sizes for different study-types adapted from Cash et al. (2022): The sample size is primarily based on the research objective and the research methods used (Cash et al., 2022).

A central methodological problem also lies in the way in which needs for support are identified. In most empirical studies, the need for support is not stated by the designers themselves but is based on interpretation by researchers (Hay et al., 2017). Around 70 % of the studies examined use protocol analyses in which the participants report retrospectively on their approach, but rarely on individual difficulties or needs. Only around 20 % of the studies used self-reporting by the designers, mostly in the form of retrospective interviews over longer development periods. In addition, these surveys usually relate to entire development projects and less often to the situation-specific dimensions within the conceptual design phase. A differentiated investigation of difficulties in specific phases such as conceptual design rarely takes place. (Hay et al., 2017)

This raises the question of how a better understanding of the individual needs of engineering designers can be gained during the design process, especially considering situational, context-dependent difficulties. Although the use of established research methods such as Concurrent Think Aloud is widespread, its use influences the cognitive load of the participants and can thus impair the validity of the data (van Someren et al., 1994). It is therefore necessary to develop research methods that capture the cognitive processes of engineering designers as uninfluenced as possible and at the same time enable an objective assessment of critical situations.

Overall, it can be said that there is a need for research methods and tools in design research suitable to systematically record individual needs for support. The research methods used to date are either dependent on subjective assessments by researchers or involve so much effort that a broader empirical basis can hardly be realized. Accordingly, there is a need for further research into valid, practical, and designer-centered research methods that allow situational needs to be identified in a reliable manner.

2.1.6 Physiological Signals in Design Research

In addition to conventional empirical research methods for investigating design processes (see chapter 2.1.5), physiological signals are increasingly becoming a focus of research. These enable the design process to be recorded continuously and automatically evaluated, opening up new perspectives on the cognitive and affective states of engineering designers. In literature, a distinction is made between two categories: design physiology, which encompasses physical signals, and design neurocognition, which refers to neural activity (Gero & Milovanovic, 2020).

Both approaches are particularly important for the automation of data collection and evaluation. Although the investigation of design thinking using these measurement methods is still a relatively new approach, it is gaining increasing acceptance (Gero & Milovanovic, 2020). Lawrie et al. (2023) also point to the growing use of physiological, behavioral, and mixed methods, which show initial progress in bringing design research closer to cognitive psychology. Physiological data can provide insights into emotional and cognitive states, while neurophysiological data provides information about brain activation associated with cognitive processes (Hay et al., 2020). Both approaches are briefly introduced below, along with the common measurement methods.

2.1.6.1 Design Physiology

The aim of this approach is to capture the emotional responses of participants in a design context (Balters & Steinert, 2017). Design physiology provides objective data that allows conclusions to be drawn about attention, stress, or emotions. According to Ullman (2010), sensory information from the environment - such as via the eyes, ears, and hands - flows into the cognitive system and can therefore be captured using appropriate measurement methods. The most common physiological signals are eye tracking, electrodermal activity, electrocardiogram, and emotion tracking (facial feature recognition). (Gero & Milovanovic, 2020)

Eye tracking

Eye tracking captures the pupil, gaze movements, and fixations, providing insights into visual attention and information processing (Holmqvist et al., 2011). In design research, it is used to investigate analytical and creative processes (Ruckpaul et al., 2017; L. Sun et al., 2014). In this context, eye tracking data from engineering designers can provide information about different functional interpretations of systems. The data is usually collected using special glasses that record eye movements and pupil positions using infrared technology.

Electrodermal activity (EDA)

EDA measures the electrical conductivity of the skin. It is therefore also known as galvanic skin response (GSR). The signal is considered an indirect indicator of emotional and physiological reactions (Villanueva et al., 2018) and is also an indicator of cognitive load (Nourbakhsh et al., 2013). The measurement is usually taken using sensors on the fingers or palms, where the sweat glands are particularly dense.

Electrocardiogram (ECG)

The electrocardiogram records the electrical activity of the heart and provides the basis for calculating heart rate variability (HRV). This describes the temporal variation between successive heartbeats and allows conclusions to be drawn about mental strain or stress (Lohmeyer & Meboldt, 2016). The measurement can be taken using classic electrode arrangements or chest strap systems.

Emotion Tracking

Emotion tracking captures facial expressions and automatically assigns them to emotional states. Emotions are considered central to understanding human behavior and offer potential for inclusion in the design process (Balters & Steinert, 2017). The data is collected using camera systems and facial recognition algorithms.

2.1.6.2 Design Neurocognition

Design neurocognition examines the connection between cognitive processes and neural activity. The aim is to better understand the neural basis of thinking in the design process and to identify the underlying mechanisms of decision-making, problem-solving, visual processing, and other cognitive processes. Design is viewed as a complex, context-dependent activity that activates several areas of the brain. Methods from cognitive neuroscience provide new measurement possibilities to expand theories of design research. (Grohs et al., 2017; Shealy et al., 2017)

Electroencephalography (EEG)

EEG measures the electrical activity of the brain using electrodes placed on the head skin. It enables the recording of rapid cognitive processes, but has limited spatial resolution, making it difficult to precisely locate activated brain regions (Eysenck & Keane, 2020). EEG is mainly used for tasks involving simple stimuli or basic processes. Processes such as design and problem solving can be examined as well (Grohs et al., 2017). The measurement is performed using an electrode cap

that records electrical activity. EEG has potential in design research, but it only measures cognitive brain activity and does not allow conclusions to be drawn about the design context on its own (Gero & Milovanovic, 2020).

Functional near infrared spectroscopy (fNIRS)

With fNIRS, brain activity is measured through changes in blood flow and oxygen concentration in the brain using imaging techniques. Participants can sit upright and perform tasks during the measurement, which allows for more realistic design situations to be simulated compared to fMRI (Eysenck & Keane, 2020). fNIRS can be used, for example, to investigate differences in design cognition depending on expertise (Shealy et al., 2017). The measurement is performed using a cap that sends light into the head skin and measures reflected light to determine the activity of specific brain regions. In besides its higher resolution, fNIRS has comparable applications to EEG in design research.

Functional magnetic resonance imaging (fMRI)

fMRI indirectly measures neural activity through changes in blood flow, known as the BOLD signal (Blood Oxygen Level Dependent). The method offers very high spatial resolution but is limited to stationary laboratory environments, as participants must remain still in a scanner during the measurement (Eysenck & Keane, 2020). Data is collected by imaging blood flow in the activated brain regions. The severely restrictive scanners used in stationary laboratory settings make fMRI hardly useful for practical investigations in design research.

2.1.6.3 Conclusion on Physiological Signals

The areas of design physiology and design neurocognition presented above provide comprehensive insights into the physical and neural processes of engineering designers. Within physiology, eye tracking is a particularly interesting method as it not only provides insights into cognitive processes, but also captures visually recorded information, including the context of the design task (Ullman, 2010). Due to this potential, eye tracking will be examined in more detail in the following chapter.

2.1.7 Interim Conclusion

The conceptual design phase plays a central role in the product development process, as fundamental decisions are made in this phase that significantly determine the further course of the design process. In this phase, the transition from creative ideation to systematic solution development takes place, whereby decisions about functions, solution principles, and technical feasibility must be made, often

under uncertainty and high complexity. Accordingly, there is a great need for methodological support to systematically manage the multitude of competing requirements and develop viable concepts.

There is a great need for methodological support to overcome these challenges. Design methods offer structured support for central activities such as generating ideas, making decisions or analyzing problems. However, their effectiveness is highly dependent on their suitability for the specific problem situation at hand and on the individual characteristics and preferences of the engineering designers. The actual use of such design methods is therefore context- and designer-specific. Particularly in non-routine, cognitively stressful situations, an individual need for support arises that can often neither be directly observed nor systematically recorded.

Design research is therefore faced with the challenge of identifying this need not only retrospectively, but also situation-based and time continuous. Although established empirical research methods exist, such as interviews, observations or protocol analysis, these have clear limits in terms of their effort and scalability. In addition, methodological needs are usually recorded indirectly through interpretation by the researchers, while direct, continuous self-reporting by the engineering designers is rare and increases the effort even further.

This situation leads to a central need for design research: To advance the development of new design methods and to statistically validate them, new research methods and tools are needed that allow the individual need for support in the conceptual design phase to be recorded. These research methods must be able to:

- consider the need for support individually from the engineering designer's point of view,
- enable time-continuous evaluation of the process,
- to be able to automatically process a large amount of data in a scalable manner, and
- deliver objective and reliable results.

2.2 Cognitive Load

This chapter provides the basics and implications of cognitive load and examines its suitability as an indicator of support needs (chapter 2.2.1). Chapter 2.2.2 then provides an overview of the various possibilities for measuring cognitive load.

Finally, a preliminary study (chapter 2.2.3) examines the extent to which cognitive load is suitable as an indicator of support needs in engineering design. Chapter 2.2.4 draws a preliminary conclusion.

2.2.1 Cognitive Load Theory

Cognitive Load Theory (CLT) originates from psychology and describes the cognitive demands that arise during learning or problem solving. Its central concept is the assumption that humans have a limited capacity for working memory, which restricts the simultaneous processing of information (Paas & van Merriënboer, 1994; Sweller et al., 2011). The CLT is therefore highly relevant not only for learning processes, but also for complex cognitive activities such as design, as these also involve intensive information processing and problem-solving processes (Paas & van Merriënboer, 1994). In the context of engineering design, cognitive load specifically refers to the mental resources required to complete a design task.

Cognitive load consists of three categories: Intrinsic, extrinsic, and germane cognitive load (Paas et al., 2003)

- ***“Intrinsic cognitive load*** is determined by the interaction between the content and type of information available and existing prior knowledge. In the process, comprehension is built up.” (Zimmerer et al., 2023, p. 482)
- ***“The design and structure of the available materials influence the extrinsic cognitive load.*** A high number of repetitions, cross-references as well as many alternating ways of providing information result in an increased extrinsic load.” (Zimmerer et al., 2023, p. 483)
- ***“The germane cognitive load*** represents the learning-related contribution to the cognitive load. It describes the additional effort required to permanently internalize the knowledge currently built up in the form of processes or schemata. If the intrinsic and extrinsic load become too high, there is not enough capacity left to internalize what has been applied.” (Zimmerer et al., 2023, p. 483)

The combination of these three categories results in the total cognitive load, which ideally remains below the capacity limit of working memory. If this limit is exceeded, cognitive performance declines and necessary information can no longer be processed adequately (Sweller et al., 2011). From the perspective of design research, the germane load can be understood as the proportion of cognitive capacity that is devoted to the application of design methods, tools, and processes. Methods and tools in design can contribute to increasing germane load and at the same time, they help to reduce excessive intrinsic and extrinsic load, which can otherwise limit the available capacity.

A commonly used model for describing the relationship between cognitive load and performance is the inverted U-shaped relationship first formulated by Yerkes and Dodson (1908): up to a certain point, increasing cognitive load improves performance because it has a stimulating effect. However, if a critical level is exceeded, cognitive overload occurs, which in turn leads to a decline in performance. (Bruggen, 2015; Yerkes & Dodson, 1908) Even too little cognitive load, such as being underchallenged or performing monotonous tasks, leads to a reduction in performance.

In addition to these three categories, Charles and Nixon (2019) also introduced task load to describe the complexity of a task itself. Task load refers to the structurally determined requirements of a task, independent of individual factors such as prior knowledge or experience (Charles & Nixon, 2019; Hancock et al., 2021). It thus represents an external influencing factor that can have a significant impact on intrinsic load in particular. Reducing task load, for example, through targeted task structuring or simplification, can help to lower overall cognitive load (Sweller et al., 2019).

In engineering design, it has been shown that cognitive load depends heavily on the specific type of design activity. Different tasks in the design process are associated with specific cognitive demands, such as mental effort, physical and temporal demand, or stress (Nolte & McComb, 2021). Analyzing such demands provides a better understanding of how cognitive load changes over the course of the design process, and in which phases targeted support can be particularly effective.

Based on the relevance of the conceptual design phase described in chapter 2.1.1 - a large number of requirements and a high level of uncertainty -, increased cognitive load is expected in early stages of the product development process. This is due to the fact that complex information must be processed, open problems structured, and creative solutions generated. Several studies suggest that creative performance in particular is influenced by cognitive demands (Howard et al., 2008; Nelson & Menold, 2020; G. Sun & Yao, 2012). For example, G. Sun and Yao (2012) investigated the cognitive load of experienced and inexperienced designers solving a conceptual design task and showed that there is a correlation between cognitive load and creativity. It can therefore be concluded that targeted methodological support can focus cognitive load and thus increase the quality of results (Howard et al., 2008).

CLT thus provides a theoretically sound basis for analyzing and developing design methods. By taking into account the interaction between cognitive load, expertise, and method, situation specific design methods can be developed and provided that

increase both the efficiency and quality of design processes (Sweller et al., 2019). A deeper understanding of cognitive load in the context of design research - including its situational characteristics - is therefore key to developing effective, adaptive design methods.

In this context, design methods offer a possibility for managing the balance between the three types of cognitive load. By structuring tasks, reducing irrelevant information, and providing targeted assistance, they can reduce cognitive load, especially extrinsic load. This frees up cognitive capacity that can be used for creative or strategic problem-solving processes (Lavrsen & Daalhuizen, 2024). This means that well-designed methods not only increase efficiency in the design process, but can also strengthen the creative or problem-solving potential of engineering designers.

In summary, cognitive load theory provides a framework for describing and evaluating cognitive demands in engineering design. Its application allows design tasks to be systematically analyzed in terms of their mental strain and suitable design methods to be derived. In order to fully utilize this potential in design research, it is necessary not only to describe cognitive load theoretically, but also to measure it empirically and relate it to performance parameters.

2.2.2 Measure Cognitive Load

There are basically three methodological approaches available for measuring cognitive load (Longo, 2018):

- 1) survey-based methods based on subjective self-reports,
- 2) task-performance-based methods, and
- 3) methods based on physiological signals that measure physical reactions.

Survey-based methods rely on the participant's self-assessment and capture their individual perception of cognitive load. They are considered particularly meaningful, as only the participants themselves can provide a well-founded assessment of their experience of cognitive load. (Longo, 2018)

Task-performance-based methods assume that cognitive load is particularly relevant when it affects performance in a task. Depending on the approach, either the primary task itself or a supplementary secondary task is evaluated to record changes in cognitive capacity. Therefore, they cannot measure cognitive load consistently. Typical indicators include reaction times or error rates. (Longo, 2018) Since Task-performance-based methods measure cognitive load only implicitly via

performance results and thus relate more to the area of performance measurement, this approach will not be discussed in more detail below.

Physiological measurements are based on the analysis of physical reactions associated with cognitive load. These include heart rate, eye movements, brain activity (e.g. via EEG), and muscle activity (EMG). These methods enable continuous recording without actively interrupting the participant's actions. (Longo, 2018)

The following section first presents the questionnaire-based methods for assessing cognitive load. The questionnaire-based methods are particularly relevant for the preliminary work in chapter 2.2.3, which examines the influence of cognitive load on performance. This will be carried out using established questionnaire-based methods initially. The methods based on physiological signals are discussed in more detail in chapter 2.3.1.

2.2.2.1 Survey-based Assessment of Cognitive Load

Even though rating scale techniques have repeatedly proven to be reliable and sensitive (Paas et al., 1994), there is no established gold standard for measuring cognitive load (de Winter, 2014; Hart & Staveland, 1988). As a result, there are a variety of questionnaire instruments in the literature for subjectively measuring cognitive load. de Winter (2014) provides a structured overview, compiling a detailed evaluation of the frequency of use of various assessment methods (de Winter, 2014). Grier (2015) provides a complementary quantitative perspective, analyzing the frequency of use of various methods based on 237 studies (Grier, 2015).

The common methods can be divided into one-dimensional and multidimensional survey methods, which differ in terms of data collection and interpretability of the results:

One-dimensional questionnaires for assessing cognitive load summarize perceived load in a single global value. Two established methods in this area are the Modified Cooper-Harper Scale (MCH) (Wierwille & Casali, 1983) and the Overall Workload Scale (OWS) (Vidulich & Tsang, 1987). The MCH is based on a ten-point rating using a decision tree, while the OWS uses a linear scale from 0 (low stress) to 100 (high stress).

The main advantage of one-dimensional methods is their ease of use, the short time required to complete the survey, and the low cognitive load on respondents. They are particularly suitable as screening instruments when a quick assessment of

cognitive load is necessary or when other survey methods are used in parallel. (Hill et al., 1992)

However, one-dimensional scales are limited in terms of their sensitivity and diagnostic depth. They do not provide any information about the causes or composition of the perceived load and are therefore only of limited use for differentiated analyses of cognitive demands.

Multidimensional questionnaires differentiate cognitive load along several sub-aspects. Among the best-known methods are the Subjective Workload Assessment Technique (SWAT) (Reid & Nygren, 1988) and the NASA Task Load Index (NASA-TLX) (Hart, 2006; Hart & Staveland, 1988). Both capture several dimensions such as mental strain, time pressure, physical exertion, and perceived performance.

The multidimensional analysis allows both the causes and characteristics of cognitive load to be identified more precisely. This not only enables a more accurate analysis of the respective cognitive load factors, but also allows well-founded conclusions to be drawn about the causes of the cognitive load. In addition, studies show that multidimensional scales are more sensitive and more readily accepted by participants (Grier, 2015; Hill et al., 1992).

The effort required for implementation and evaluation is higher compared to one-dimensional methods, and the assessment is less intuitive. Nevertheless, in many applications, the advantages in terms of sensitivity and interpretability of the results outweigh the disadvantages. (Hill et al., 1992)

Within these categories, it is evident that, despite the existing diversity of different methods, the NASA Task Load Index (NASA-TLX) by Hart and Staveland (1988) has established itself as by far the most frequently used instrument (de Winter, 2014; Grier, 2015). Thus, it can be regarded as a kind of gold standard for measuring cognitive load (Dias et al., 2018).

The NASA Task Load Index (NASA-TLX) according to Hart and Staveland (1988) and Hart (2006) is based on a multidimensional assessment of six aspects of workload: mental demand, physical demand, temporal demand, effort, performance, and frustration level. All dimensions are rated on a scale from 1 to 20. (Hart, 2006)

There are two possible calculation methods for determining the overall cognitive load. In the original version, the individual dimensions are weighted based on pairwise comparisons, whereby individual weighting factors are determined and then used to calculate a weighted total value. Alternatively, an unweighted variant can be used, in which the six individual ratings are averaged directly. This method,

known as NASA Raw TLX (RTLX) (Hart, 2006), has also become widely used. Since no methodological advantage of either variant has yet been empirically proven (Hart, 2006), both approaches are considered equally applicable.

2.2.3 Further Containment Based on own Preliminary Work: Cognitive Load as Indicator for Need of Support

Cognitive load theory suggests that cognitive load may also has a central influence on performance in the design process. The survey-based methods described above - especially the NASA-TLX - offer a reliable way to measure this cognitive load. At the same time, research shows that every design activity creates a unique pattern of cognitive load (Nolte & McComb, 2021).

The effect of cognitive overload is particularly relevant here, as there is great potential for the use of design methods that can increase performance in the design process by specifically reducing cognitive load (Lavrsen & Daalhuizen, 2024). Against this background, the following preliminary work investigates the connection between cognitive load and performance in engineering design. However, the connection between cognitive load as an indicator of decreasing performance and the need for support in design research still needs to be demonstrated. To this end, the preliminary research question 1

PRG1: “How does cognitive load affect performance in engineering design?”

is examined.

The following subchapters 2.2.3.1 to 2.2.3.3 are based on the publication “Study on the impact of cognitive load on performance in engineering design” (Zimmerer & Matthiesen, 2021). The text has been taken from the publication with minor modifications, especially with regard to numeration, formatting, and terminology. For the sake of clarity, the texts have not been marked separately.

2.2.3.1 Experimental Design

The following chapter describes the study that was conducted to answer the preliminary research question 1. The structure of the study design and the evaluation procedure is described in detail.

Participants

A total of 15 participants took part in the study. A requirement for participation was that the participants were students in the fields of mechanical engineering or mechatronics and successfully completed a mechanical design course. This was to ensure that all participants had sufficient and comparable basic technical knowledge.

Tasks

Each participant was asked to work on six different tasks. The tasks spanned different fields of engineering design. A more detailed description of each task can be found in Table 2.2. In addition, there were two small introductory tasks at the beginning to familiarize the participants with the type of data acquisition. However, these introductory tasks were not included in the evaluation.

The different tasks are briefly described in Table 2.2:

Table 2.2: Description of design tasks for answering PRQ1 (Zimmerer & Matthiesen, 2021)

task 1	analysis of technical drawing	In this task, participants should analyze a technical drawing of a gearbox, including identification of various design details such as the bearing concept and gear stages.
task 2	functional analysis 1: cartridge press	In this task a cartridge press represented by a technical drawing should be analyzed. The aim of the task was to evaluate the influence of several parameters on the clamping effect of the reset mechanism.
task 3	functional analysis 2: snap-fit assembly	A snap-fit connection is a snap connection with a combination of non-positive and positive locking. Subject of the task was to assess the influence of various design parameters. In each case, the effect of an increase in the respective parameter on the retention force of the snap connection was to be evaluated.

task 4	considering manufacturing processes in design	This task deals with the manufacturing of a sheet metal bending part. For the unfolding of a given sheet metal part, the different bending steps should be considered in order to obtain the 3D part that is also given.
task 5	failure analysis	In this task various failure scenarios of bearings were to be assessed. For this purpose, the load on the shaft leading to the respective damage scenario should be identified for each given case.
task 6	load analysis in machine elements	In this case, a section of a technical drawing of a shaft was given. For several marked machine elements, the load that is applied to the component by clamping it with a shaft nut should be assessed here.

The purpose of the different tasks was to cover a similar range on the cognitive load scale as described in the results of Grier (2015). Both the total range and the range in which the majority of data sets (80 %) are located were compared.

Data collection

The data collection was carried out through an online setup of the study. The university's own learning platform was used.

The basic structure of the study was designed in a way that after each task a query of the cognitive load via the NASA-RTLX (see chapter 2.2.2) was carried out. The questionnaire was based on the dimensions described by Hart (2006). Only the dimension physical demand was excluded, as the study was conducted solely on computer. Therefore, no differences in physical load were expected between tasks. Previous research has also shown that the design process should be considered more as mental and not as physical activity (Nikulin et al., 2019). The NASA-RTLX questionnaire thus provides a combined value for cognitive load between 1 and 100.

The tasks were shortly introduced in both text and image form to the respective problem, followed by several questions to which the participants could enter their solution.

Data analysis

The data analysis was carried out in four consecutive steps:

- 1) First, the results from the NASA-RTLX questionnaires were evaluated. For each participant and task, the values of the five individual dimensions were normalized to obtain the cognitive load on a scale from 0 to 100. Subsequently, the values were grouped according to the tasks.
- 2) In a second step, the tasks were evaluated in terms of performance. For this purpose, a defined number of different items were determined for each task in advance of the execution. In the analysis, the participants' answers for each item were evaluated for every task and classified as correct or incorrect with regard to the pre-defined solution. At the end of the performance evaluation, the quotient of correct items in relation to the total number of items available was calculated for each task. Thus, the performance for each task can be determined specifically in [%] and compared with the values from the other tasks.
- 3) The results from the first two steps were then combined to investigate possible relationships between the two parameters. For this purpose, the performance values of the individual tasks were set in relation to the result of the respective NASA-RTLX questionnaire. This way, one data point, consisting of the two parameters cognitive load and associated performance, is created for each participant and task. The resulting data set is used to search for a correlation between the two parameters.
- 4) A statistical evaluation conducted in a fourth step completed the analysis. The SPSS Statistics 26 software by IBM was used for evaluation. The confidence interval for the determination of significant results was set at 95 %.

Two tests were used for the actual statistical evaluation. The Spearman correlation test was used to search for correlations between the cognitive load and performance. This is to be used because the data was not normally distributed.

For the investigation of differences between the individual performance values or the cognitive load values, the Mann-Whitney U test was used, since normal distribution of data could not be assumed.

2.2.3.2 Results

Of the 15 participants, only 12 fully completed the study. Therefore, only these 12 were considered for further evaluation. In the following chapters, the results from the NASA-RTLX questionnaires and the performance in the tasks are presented individually and in relation to each other.

Cognitive load in the different tasks

In the first step the NASA-RTLX questionnaires were evaluated as described in chapter 2.2.3.1 and the results for the individual tasks were examined. Figure 2.3 gives an overview of the results from the evaluation of the NASA-RTLX questionnaires separately for each task.

Across all tasks, the cognitive load scatters in a range between 7 and 95, which means that the values are very comparable with the range of 6 to 89 determined by Grier (2015). Furthermore, the range in which the majority of the data (80 %) is located is also very close to the values reported by Grier (2015). In this study, 80 % of the data points are between 23 and 71 and the comparative area is between 26 and 68.

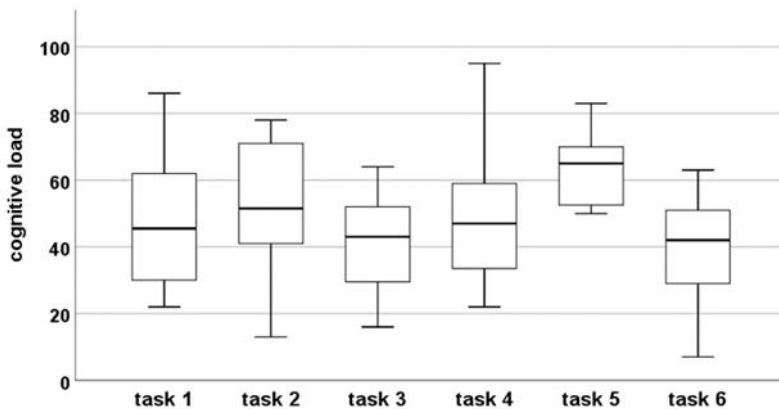


Figure 2.3: Results of the evaluation of cognitive load using the NASA-TLX questionnaire for the various tasks (n = 12) (Zimmerer & Matthiesen, 2021)

Looking at the tasks individually, it is visible (see Figure 2.3) that they also produce different values in the cognitive load. Task 5 (failure analysis) is by far the most difficult task for the participants with the lowest scatter (median = 65; SD = 10.2). The remaining tasks are relatively close according to the median cognitive load but differ in their scatter. Task 3 (functional analysis; median = 43.0; SD = 15.8) and task 6 (load analysis in machine elements; median = 42.0; SD = 16.9) have a lower scatter than task 1 (analysis of technical drawing; median = 45.5; SD = 20.3), task

2 (functional analysis; median = 51.5; SD = 19.4), and task 4 (considering manufacturing processes in design; median = 47.0; SD = 21.3).

Task 5 (failure analysis) also produces statistically significantly higher values in the cognitive load than task 1 (analysis of technical drawing; two-tailed Mann-Whitney U test, $U = 36$, $p = 0.037$), task 3 (functional analysis; two-tailed Mann-Whitney U test, $U = 16$, $p = 0.001$), task 4 (considering manufacturing processes in design; two-tailed Mann-Whitney U test, $U = 34$, $p = 0.028$), and task 6 (load analysis in machine elements; two-tailed Mann-Whitney U test, $U = 13.5$, $p = 0.001$).

Performance in the different tasks

Following the cognitive load, the performance of the participants in the different tasks was evaluated as a second step. Figure 2.4 provides a closer look at the performance of the participants in the different tasks.

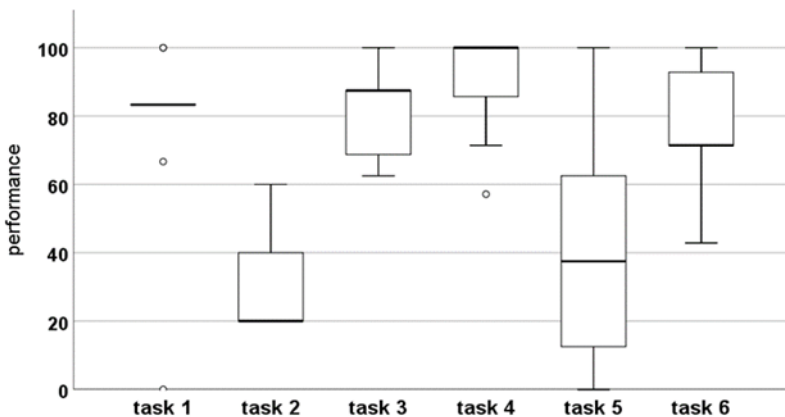


Figure 2.4: Results of the evaluation of performance for the various tasks (n = 12) (Zimmerer & Matthiesen, 2021)

It can be seen that the participants performed considerably better in task 1 (analysis of technical drawing; median = 83.3; SD = 26.0), task 3 (functional analysis; median = 78.5; SD = 13.6), task 4 (considering manufacturing processes in design; median = 95.0; SD = 14.2), and task 6 (load analysis in machine elements; median = 71.4; SD = 19.4) than in task 2 (functional analysis; median = 20.0; SD = 15.9) and task 5 (failure analysis; median = 37.5; SD = 32.8).

The scatter in the individual tasks is rather low with values between 14.2 and 26.0 with the exception of task 5 (failure analysis) where the scatter is significantly higher with 32.8.

Correlation between cognitive load and performance

Combining the results from the two datasets cognitive load and performance results in the scatter plot shown in Figure 2.5. Each data point represents one task processed by one participant. A correlation analysis was conducted to determine whether there is a correlation between the two parameters. The result is shown as a linear line in the figure. As the data is not normally distributed, a correlation analysis according to Spearman was used. The cognitive load indicated by the participants correlates significantly with the performance achieved in the corresponding tasks, $r = -0.243$, $p = 0.040$, $n = 72$. According to Cohen (1992) this value is attributed to a weak effect.

A correlation analysis based on the assumption of a quadratic relationship similar to the curvilinear (or inverted-U) relationship did not yield a statistically viable result.

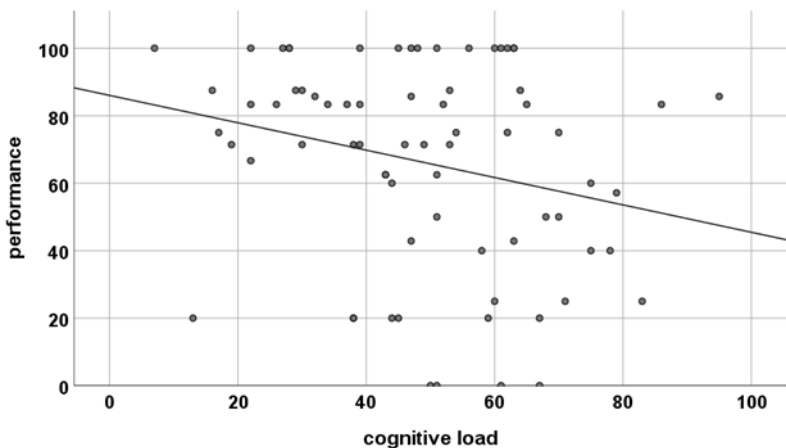


Figure 2.5: Correlation between cognitive load values and performance across all tasks and participants displayed as a negative trending line ($r = -0.243$, $p = 0.040$, $n = 72$) (Zimmerer & Matthiesen, 2021)

2.2.3.3 Discussion and Conclusion

The aim of this preliminary study was to investigate the relationship between cognitive load and performance in design. A general decrease in performance with increasing cognitive load could be detected. This performance decrease can be interpreted as cognitive overstraining caused by difficulties in processing the given tasks. However, an additional drop in performance at low levels of cognitive load could not be shown.

Suitability of the tasks

Since there have been very few investigations of the cognitive load in the field of engineering design so far and these are limited to conceptual design, the suitability of the applied study design must be evaluated. For this purpose, the cognitive load values are compared with common values from other disciplines, which were investigated and summarized by Grier (2015). This comparison showed that the tasks used in this study achieved very similar values to those obtained in investigations from other areas.

Altogether, the six different tasks covered a wide range on the cognitive load scale. The distribution of the values is generally very close to the values that were identified in the investigation of studies from other areas which were using the NASA-RTLX

(Grier, 2015). This leads to the conclusion that the tasks as a whole are well suited to investigate the cognitive load in engineering design.

One limitation of the tasks used, however, is that there is very little difference between the different tasks. As can be seen from Figure 2.3, the only task that seemed to be significantly more difficult than the others was task 5 (failure analysis), which makes it hard to derive generally statements on the context investigated. However, differences can be seen in the distribution of the individual tasks. Especially tasks 1, 2 and 4 show a strong scattering of cognitive load values. From this can be derived that the individual difficulties of the participants in certain areas are more important than a difficulty induced by the individual task.

When looking at the performance in the individual tasks, a contradictory situation emerges. While the tasks that scattered very much in the cognitive load show a very low degree of scatter in the performance. Task 5 (failure analysis), which has the highest values with the lowest scatter in the cognitive load, scatters the most in performance.

Correlation between cognitive load and performance

When the two data sets from cognitive load and performance are combined in a single diagram (Figure 2.5), it is difficult to establish a relationship between the two variables based on the individual data points. Although a correlation analysis has shown a significantly decreasing correlation, it is rather weak. This is also illustrated by the weak effect according to Cohen (1992).

From this, it can be concluded that although the available data provide a weak indication of the right part of the inverted-U curve according to Yerkes and Dodson (1908) and Bruggen (2015). This effectively represents the area of overstrain. The left part of the inverted-U curve, which stands for understrain, could not be mapped. A possible explanation for this situation is that the tasks used can be described as rather “simple” based on the NASA-RTLX used, but were not so trivial that an actual understraining occurred.

Limitations

The effect of the statistical outcome is limited by the relatively small number of participants in this study ($n = 12$). Furthermore, due to the decision for an online setup for the study, it was not possible to fully control the study conditions during the processing. This may have led to deviating influences on the individual dimensions of the NASA-RTLX.

2.2.4 Interim Conclusion

Cognitive Load Theory emphasizes that high cognitive load can impair designers' performance and, consequently, serve as an indicator of potential need for methodological or tool-based support within design research. Previous studies have demonstrated that high cognitive load can negatively affect problem-solving efficiency and creative output. Furthermore, survey-based methods for assessing cognitive load are well-established and can provide valuable insights into designers' subjective perceptions of cognitive demands. However, these methods exhibit significant limitations regarding time-continuous measurement and automated analysis, which restrict their applicability for automatable and real-time investigations of design activities.

The preliminary study conducted confirmed a significant correlation between cognitive load and design performance. In addition, the study revealed only minor differences in the average cognitive load between distinct design tasks. This homogeneity provides a favorable condition for subsequent machine learning-based investigations, as it enables the generation of sufficiently large and consistent datasets. Consequently, the results of the preliminary work suggest that avoiding situations of high cognitive load could represent a promising approach to enhancing design performance, while also indicating that the prerequisites for developing approaches to automatically assess cognitive load in design activities are promising.

2.3 Automated Data Collection and Analysis of Cognitive Load

Chapter 2.1.5 already highlighted the need for new research methods and tools, as established procedures such as protocol studies require a great deal of time and resources, thereby limiting the size of the samples (Ahmed, 2007). Automated, computer-based methods for data collection and analysis offer the potential to conduct studies more objectively, continuously over time, and with larger sample sizes (Dinar et al., 2015; Hay et al., 2017). Physiological signals are particularly suitable as a measurement variable in this context (Gero & Milovanovic, 2020).

In the following, chapter 2.3.1 provides existing possibilities for measuring cognitive load using eye tracking data. In chapter 2.3.2, a preliminary study examines the extent to which these findings can already be applied to design processes. Finally, initial approaches to the use of algorithms for determining cognitive load using eye tracking data are introduced. To this end, the necessary basics are first introduced (chapter 2.3.3) and then the various approaches are explained (chapter 2.3.4).

2.3.1 Eye Tracking to Measure Cognitive Load

Eye tracking has already been used in other areas of research to measure cognitive load. Its use is particularly well established in areas such as human-computer interaction (HCI) (S. Chen et al., 2011; Zagermann et al., 2018), driving simulations (Benedetto et al., 2011; Pedrotti et al., 2014), and flight simulations (Halverson et al., 2012), but also in medical contexts such as surgery (Zheng et al., 2012). It has been shown that cognitive load is reflected in various metrics of eye movements. Nevertheless, there is still no consensus in research on which parameters are best suited to reliably measure cognitive load (Halverson et al., 2012).

The parameters most frequently examined in this context can be traced back to four basic types of eye movements: pupil diameter, fixations, saccades, and blinks (Holmqvist et al., 2011). Pupil diameter describes the size of the pupil and is usually measured in millimeters. Fixations refer to phases in which the eyes rest on a point and can be quantified in terms of duration and frequency. Saccades are rapid eye movements between fixations, which are characterized by angle or speed. Finally, blinks can be recorded in terms of their frequency and duration. (Holmqvist et al., 2011)

Table 2.3 provides an overview of central studies and the eye tracking metrics used in each case. It offers a compact overview of the field, tasks, and the most frequently analyzed eye metrics.

Table 2.3: Eye tracking metrics and tasks investigated in various studies from different fields (Zimmerer et al., 2023)

source	field	tasks used	eye metrics investigated
Benedetto et al. (2011)	driving simulators	simulated lane change test (LCT) in a driving simulator	blinks, pupil
Buettner (2013)	HCI	job interviews	pupil, fixation, saccade
F. Chen et al. (2016)	HCI	arithmetic tasks	blinks, pupil
J. Chen et al. (2019)	HCI	flight tasks with different difficulties	blinks, pupil, fixation, saccade

S. Chen et al. (2011)	HCI	search task in computer-based training application,	blinks, pupil, fixation, saccade
Halverson et al. (2012)	flight simulators	aircraft monitoring on screen	blinks, pupil, fixation, saccade
Pedrotti et al. (2014)	driving simulators	different driving tasks	pupil
Petkar et al. (2009)	design	stroop-tests	blinks, pupil
Zagermann et al. (2018)	HCI	stroop-tests	blinks, pupil, fixation, saccade
Zheng et al. (2012)	medicine/surgery	two simulated cases of laparoscopy	blinks

In the field of HCI, eye tracking data has often been used to record strain during computer-based tasks. S. Chen et al. (2011) investigated search tasks in training applications and showed that blinking, pupil reactions, fixations, and saccades react sensitively to task requirements. In later work, F. Chen et al. (2016) demonstrated that both arithmetic tasks and flight control tasks in the HCI context lead to characteristic patterns in pupil diameter, fixations, and saccades (F. Chen et al., 2016; J. Chen et al., 2019). Buettner (2013) also found changes in all three metric types during cognitively demanding interview tasks, when considered individually. Zagermann et al. (2018), who used Stroop tests, also showed that blinking, the pupil, fixations, and saccades are related to cognitive load. The results suggest that a wide range of metrics are used in the HCI field and that there is no clear focus on individual metrics.

Driving simulators are another field for investigating cognitive load using eye tracking. Benedetto et al. (2011) showed in the lane change test that changes in pupil size and blink rate correlate with cognitive load. Pedrotti et al. (2014) were also able to demonstrate, using various driving tasks, that pupil diameter increases systematically with mental stress. Another important parameter is PERCLOS (Percentage pupil closure), which was originally developed to measure fatigue but is also associated with increased mental stress (Halverson et al., 2012; Marquart et al., 2015). This makes it clear that pupil metrics and blinking in particular provide reliable results in the context of driving simulation.

Measuring cognitive load also plays a central role in aviation. Halverson et al. (2012) investigated screen monitoring activities in flight simulators and identified PERCLOS in particular as a very precise metric, in addition to pupil diameter and fixations. These studies particularly emphasize the importance of continuous eye tracking data, as discrete averaging over longer tasks often hides important variations. (Halverson et al., 2012)

In the medical field, Zheng et al. (2012) used eye tracking to measure the cognitive load of surgeons during laparoscopic training tasks. The blink rate proved to be particularly relevant, as a direct correlation with the increase in cognitive load was identified. Such findings are highly relevant in medicine, as they open up possibilities for identifying difficulties in real time and deploying targeted assistance systems in training scenarios (Zheng et al., 2012), which can also be transferred to the design context.

In addition to considering single parameters, several studies also used combinations of different eye tracking metrics to increase the predictive accuracy with regard to cognitive load. In HCI, J. Chen et al. (2019) used pupil measurements, fixations, saccades, and blinks in parallel and were able to show that this increased predictive accuracy compared to individual metrics. This increased the prediction accuracy from around 30 % for individual metrics to 40 % for combinations of metrics. (J. Chen et al., 2019). On the one hand, this shows that combining several metrics increases prediction accuracy, but at the same time, a clear and precise determination of cognitive load has not yet been achieved. This makes it clear that while combining multiple metrics increases the information content, the measurement quality still has limitations.

On the other hand, the design lacks findings on the relationship between eye metrics and cognitive load, as there is a lack of studies in this domain. Although Petkar et al. (2009) can be assigned to the design domain, the study relied on Stroop tests rather than actual design tasks and therefore provides only limited insight into cognitive load during real design processes.

The present studies show that eye tracking has been repeatedly used to measure cognitive load in various disciplines, ranging from driving simulators to HCI, aviation, and medicine. Pupil metrics and blink parameters emerge as particularly frequently used indicators, supplemented by fixation and saccade measures. At the same time, the results indicate that the sensitivity of the metrics depends quite considerably on the respective field of application or the activity performed. In contrast, there is a need for reliable studies in the field of design research, indicating a need for further research in this area.

In summary, existing research demonstrates the potential of eye tracking for identifying cognitive load across various fields. However, the absence of studies in the design domain leaves open whether comparable relationships between eye movement parameters and cognitive load can be observed during design-specific activities. To further explore this, the following preliminary study examines the suitability of eye tracking as an indicator of cognitive load in the context of conceptual design.

2.3.2 Further Containment Based on own Preliminary Work: Measure Cognitive Load in Design via Eye Tracking

Previous research shows that eye tracking metrics have been successfully used in various disciplines to measure cognitive load (see chapter 2.3.1). However, the results also show that the sensitivity and significance of individual parameters depend considerably on the respective field of application (Halverson et al., 2012). In the field of design research, especially conceptual design, there aren't any solid studies yet that systematically check which eye tracking metrics are good for measuring cognitive load. With this in mind, this work investigates the preliminary research question 2:

PRQ2: *“How applicable are established eye parameters for capturing cognitive load in the design process?”*

The aim is to examine the transferability of common eye tracking metrics to conceptual design and thus create a basis for future studies in design research.

The following subchapters 2.3.2.1 to 2.3.2.3 are based on the publication “Using Eye Tracking to Measure Cognitive Load of Designers in Situ” (Zimmerer et al., 2023). The text has been taken from the publication with minor modifications, especially with regard to numeration, formatting, and terminology. For the sake of clarity, the texts have not been marked separately.

2.3.2.1 Experimental Design

Participants

A total of 27 participants attended the study. Since engineering design students and design engineers differ in the way they analyze and process perceived information (Ruckpaul, Nelius, & Matthiesen, 2015), participants from both groups were considered.

The group of students consisted of 13 male and 2 female participants. They were on average 24.1 years old (standard deviation: 2.4 years) and in their 10th semester (average: 9.1 semesters, standard deviation: 3 semesters).

The 12 participating design engineers consisted of 10 males and 2 females and were on average 33.2 years old (standard deviation: 7.2 years). The designers came from different companies and had between one and 20 years of professional experience (average: 7.6 years, standard deviation: 5.3 years). All participants had normal or corrected to normal vision.

A detailed overview of all participants is shown in Table 2.4:

Table 2.4: Participant demographics (Zimmerer et al., 2023)

	Type	Age	Gender	Semester	Work experience [in years]
P 01	Design engineer	34	Male		6
P 02	Design engineer	45	Male		20
P 03	Design engineer	30	Male		3.5
P 04	Student	27	Male	13	
P 05	Student	26	Male	7	
P 06	Student	26	Male		11
P 07	Design engineer	30	Male		4
P 08	Design engineer	49	Female		15
P 09	Student	21	Male	3	
P 10	Design engineer	26	Male		1
P 11	Student	20	Female	3	
P 12	Student	22	Male	10	
P 13	Student	26	Male	12	
P 14	Student	22	Female	9	
P 15	Student	29	Male	11	
P 16	Student	24	Male	7	
P 17	Student	22	Male	9	
P 18	Student	25	Male	11	
P 19	Student	24	Male	11	
P 20	Student	23	Male	9	
P 21	Student	25	Male	13	

P 22	Design engineer	29	Female	3
P 23	Design engineer	25	Male	2.5
P 24	Design engineer	26	Male	6
P 25	Design engineer	37	Male	7
P 26	Design engineer	33	Male	11
P 27	Design engineer	35	Male	9

Tasks

The participants were given a conceptual design task from the field of design for manufacturing. The objective was to redesign a given fastening bracket using sheet metal forming process. In addition, the manufacturing effort was to be reduced as much as possible. The task consisted of two phases following the activities stated by Nolte and McComb (2021). In the first phase, the participants had to create variant concepts considering the new manufacturing process. For this purpose, they were provided with an initial assembly of a holding bracket. By changing the manufacturing process, the participants were asked to make the manufacturing of the assembly more cost-effective. For example, the number of parts could be reduced or welds could be replaced by bends. They were asked to sketch several alternatives on paper. In addition to the new manufacturing process, a number of boundary conditions were specified, such as a constant material and an unchanged drilling pattern.

In the second part of the task, created concepts were then to be evaluated first against each other and then in comparison with a selection of predefined concepts. For this purpose, the participants were asked to identify the various properties of the concepts such as the number of parts and the number of welds and bends. Subsequently, the various properties were to be weighed against each other in terms of their influence on manufacturing costs. The participants evaluated the concepts in terms of manufacturing effort. The participants were given predefined criteria (number of parts, number of manufacturing steps, etc.) for evaluation.

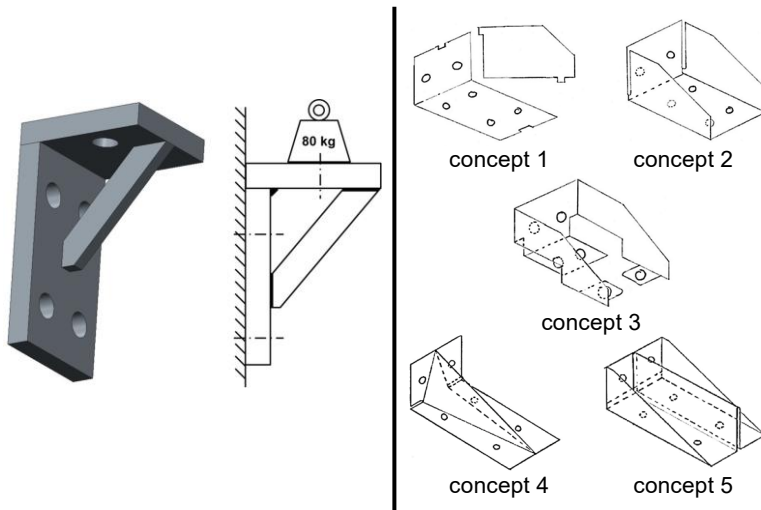


Figure 2.6: Design tasks used in the preliminary work. On the left is the given fastening bracket, for which various concepts were to be created as sheet metal bending designs in Task 1. On the right is the selection of predefined concepts to be evaluated in Task 2.

Procedure and data collection

The study procedure was designed such that the participants took part in the study consecutively and alone. In preparation for the task, the eye tracking technology was set up and calibrated, and the recording was started. Tobii glasses 2 (sampling rate 100 Hz) were used to record the eye tracking data. The calibration was done using the related calibration card. Subsequently, the participants were instructed to read through the task description. For the completion of the first task, the participants were given a target time of 10 minutes. Task 1 was given to them on the computer, but the actual processing of the task then took place using pen and paper. Upon completion, the participants' cognitive load during conceptualization was assessed by a NASA-TLX questionnaire. The dimensions of Hart and Staveland (1988) were used, with the exception of the physical demand. This was not considered, since the entire study took place at a desk and thus no physical demand was expected. The dimensions of mental and temporal demand, effort and frustration were assessed on a scale from low to high. Performance was surveyed ranging from good to poor. This procedure has already proven successful in previous studies (e.g. Zimmerer and Matthiesen (2021), see chapters 2.2.2 and 2.2.3).

No time limit was set for the second part, the evaluation. This task took place entirely on the computer. After the participants had evaluated their concepts against each other and against the predefined concepts, the cognitive load for the concept evaluation was assessed using the NASA-TLX.

Data analysis

First, the results from the NASA-TLX questionnaires were analyzed. Thereby, the variant without weighting factors (NASA Raw TLX) proposed by Hart (2006) was used. Accordingly, the values of the individual dimensions were added and thus a cognitive load between 1 and 100 was determined for the first and second task part respectively.

The analysis of the eye tracking data was limited to the phases of task processing starting after the reading of the task description. The focus was on the eye tracking metrics introduced in Chapter 2.3.1, which include fixation-based, saccade-based, and pupil-based measures. Therefore, the eye tracking data was first analyzed using the Tobii Pro Lab analysis software provided by Tobii. This involved event detection using the I-VT Fixation Filter developed by Tobii. Based on pre-defined parameters, fixations and saccades are detected in the raw data. The data is first smoothed by moving median filters. Subsequently, fixations and saccades are defined via a classifier. This classifier consists of a limit for the eye movement speed, which is set to 30 °/s in this case. Fixations are then defined as the contiguous regions where the eye velocity remains below this limit. A more detailed explanation of the choice of the individual parameters in the filter settings is described by Olsen and Matos (2012). Situations in which the eyes could not be detected were classified as blinking.

The pupil diameter was evaluated separately from the other data since it was measured continuously over the evaluation frequency, while the other parameters always have a certain event length.

For the statistical analysis, the confidence interval was set to 95 %. Because of the level of measurement available, Spearman's rank correlation test was used to determine correlations. For the identification of differences, the Mann Whitney U test was applied.

The SPSS 26 software from IBM was used for the analysis.

2.3.2.2 Results

Out of the 27 participants, 6 could not be considered for the evaluation due to errors in the recording of the data (P2, P8, P9, P15, P21, P22). Thus, 12 engineering design students and 9 design engineers were considered for the following analysis.

Table 2.5 combines all data from the NASA-TLX questionnaires and the mean values of the considered parameters over the total task duration.

Table 2.5: Overview of the cognitive load from the NASA-TLX and the mean values of all parameters per task (CL: cognitive load, PD: pupil diameter [mm], FD: fixation duration [ms], BD: blink duration [ms], SR: saccadic rate [1/s]). The higher value in the comparison between the tasks is highlighted in dark grey (white indicates equal value). (Zimmerer et al., 2023)

	Task 1					Task 2				
	CL	PD	FD	BD	SR	CL	PD	FD	BD	SR
P 01	16	3,49	290,39	108,55	2,48	21	3,20	228,49	81,13	3,08
P 03	36	4,08	207,67	73,61	3,13	38	3,41	135,19	66,36	5,28
P 04	23	3,16	419,45	64,28	2,31	42	2,82	313,83	76,46	2,59
P 05	65	3,73	213,73	141,51	3,82	69	3,36	197,37	274,03	2,75
P 06	87	3,43	242,75	73,58	3,80	88	2,92	217,22	75,42	3,82
P 07	33	3,09	281,07	95,98	3,27	53	2,95	246,12	60,49	3,35
P 10	42	3,42	324,68	97,96	2,51	39	3,04	193,13	50,83	4,20
P 11	50	4,66	303,64	99,51	2,62	49	4,03	234,90	121,14	2,92
P 12	47	3,72	331,04	58,15	2,26	48	3,40	242,32	54,31	2,90
P 13	49	3,33	394,73	99,98	2,53	58	2,91	264,21	117,00	3,22
P 14	41	3,83	203,65	62,37	3,78	57	3,58	156,90	64,96	4,85
P 16	40	2,91	234,66	106,71	2,83	45	2,58	170,59	110,38	3,75
P 17	41	3,84	222,26	63,65	2,58	62	3,54	223,95	68,65	3,36
P 18	28	3,53	193,74	70,59	2,73	80	3,16	164,75	71,53	3,76
P 19	39	4,87	412,93	72,09	2,11	64	4,19	341,89	88,64	2,68
P 20	59	3,55	367,69	153,23	2,15	52	3,36	230,29	3,78	3,01
P 23	36	4,67	366,02	1488,28	2,36	38	3,41	267,83	3,55	3,28
P 24	5	4,18	231,94	98,93	3,56	5	3,32	250,22	1,49	3,19
P 25	22	2,99	433,68	62,91	1,74	29	2,76	279,61	1,65	3,12
P 26	39	3,72	290,98	77,52	2,72	52	3,43	267,07	4,86	2,81
P 27	28	2,74	263,59	164,93	3,24	28	2,58	175,82	2,63	4,93

The analysis of the NASA-TLX questionnaires (column 'CL') shows that most participants (n = 16) reported a higher cognitive load in the second task, the concept evaluation, than in the concept generation. Only three participants rated the first task as more challenging. For two participants, both scores were equal.

There are also no differences between the engineering design students and the design engineers. Two engineering design students indicated a higher cognitive load for the first task. In the case of the design engineers, this applies to one participant and, in addition, there are two participants in this group who rated both tasks with exactly the same value.

In the next section, statistical correlations between the cognitive load and the considered parameters shown in Table 2.5 are examined in more detail. The results of the correlational analysis are summarized in Table 2.6.

Table 2.6: Results of the correlation analysis between cognitive load and the parameters considered. Here, both tasks were considered. (Zimmerer et al., 2023)

	pupil diameter	fixation duration	blink duration	saccade rate
correlation coefficient	0.106	-0.156	0.184	0.173
P	0.502	0.324	0.244	0.274

As can be seen from the table, no significant correlations can be found between the cognitive load and any of the parameters considered. Although statistically no correlation can be detected, it is possible to read a trend from the correlation coefficient. This trend indicates the direction in which the parameters change with increasing cognitive load.

While the pupil diameter, blink duration, and saccade rate increase slightly with increasing cognitive load, fixation duration slightly decreases.

Since no significant correlations to the cognitive load could be found by the correlation analysis, the next step is to evaluate if there are any differences in the parameters between the two tasks concept generation and concept evaluation.

The results of this analysis are shown in Table 2.7.

Table 2.7: Results of the statistical analysis regarding differences between the tasks used and the parameters considered. (Zimmerer et al., 2023)

	pupil diameter [mm]	fixation duration [ms]	blink duration [ms]	saccade rate [1/s]
median task 1	3.55	290.39	95.98	2.62
median task 2	3.32	230.29	66.36	3.22
p	0.01	0.007	0.024	0.003
U	117.5	114	131	338.5
Cohen's d	0.872	0.908	0.741	1.031

When looking at the results, it can be seen that there is a significant difference in all four parameters between the two tasks. The pupil diameter, as well as the fixation duration and the blink duration, are significantly larger during concept generation than during concept evaluation. According to Cohen (1992), there is a large effect in each case.

For the saccade rate, it is exactly the opposite. Here, the value is significantly larger for concept evaluation than for concept generation. Here, too, a large effect is observed (according to Cohen (1992)).

2.3.2.3 Discussion and Conclusion

The aim of this study was to examine the transferability of eye tracking parameters that can be used to determine cognitive load to conceptual design, thereby providing a first step toward enabling in situ investigation of design activities. Building on the gaps identified in previous research, the study focused on examining the applicability of established eye tracking metrics within the context of conceptual design, in order to assess their potential for future use in design research.

For this purpose, the preliminary research question (PRQ2) was stated to what extent established eye parameters are applicable to capture cognitive load during the design process.

In order to answer the preliminary research question “How applicable are established eye parameters for capturing cognitive load in the design process?”, four parameters were examined in more detail, which should show a correlation to cognitive load based on the current state of the art. For this purpose, the cognitive load was measured using the NASA-TLX questionnaire, which was developed by Hart and Staveland (1988) and is a widely used and effective way to measure the

cognitive load of participants (Hart (2006), see chapter 2.2.2). Since different activities from the design process were used for the data collection, it was evaluated in a second step whether an influence of the tasks on the eye parameters can be observed.

The results of this study show that there is no significant correlation between the cognitive load of the participants and the eye parameters used. Although the correlation analysis yields a trend for each of the parameters considered, none of them is significant. One possible reason for this could be the lack of randomization of the tasks. However, since two different types of tasks were involved, the influence should be rather small. Nevertheless, it has to be considered for future investigations.

As mentioned in the literature review by Marquart et al. (2015) an increasing cognitive load leads to enlarged pupil diameter and an increased fixation duration. Blink duration, in turn, decreases with increasing cognitive load. (Marquart et al., 2015) Comparing this with the results from this study, it can be seen that only the trend in pupil diameter is in line with these findings. For the other parameters, the opposite can be observed. This may be due to the fact that the results by Marquart et al. (2015) refer to driving simulators, which are not comparable to the design process, or that the eye parameters were more influenced by factors other than the cognitive load.

Therefore, the second step of the evaluation examined whether the task itself had an influence on the eye parameters. The results showed that for all parameters there was a significant difference between the tasks. This result may be explained by the fact that the two tasks induce a different gaze behavior.

While concept generation tends to evoke a serene gaze pattern with long fixations and fewer saccades, this is the other way around for concept evaluation. This could be due to the fact that the different concepts are considered in rotation during the evaluation.

The difference in pupil diameter can also be explained since the participants had to look on paper for the concept generation and not on a monitor as in the evaluation. The significant differences can be explained by the changes in brightness that occur during this process. According to Klingner (2010), these are considerably stronger than the influence caused by changes in the cognitive load.

However, in the case of blink duration, unlike the three parameters before, no direct influence can be seen by the way the tasks were provided.

In summary, the preliminary research question can be answered as followed: Established eye tracking parameters show limited applicability for capturing cognitive load across various design activities. While trends in pupil diameter are consistent with findings from related fields, the lack of significant correlations and the strong influence of task characteristics indicate that these metrics cannot yet be used as reliable standalone indicators. However, the observed task-specific effects suggest that, once contextual influences are better understood and controlled, eye tracking metrics hold considerable potential as part of a multimodal approach to assessing cognitive load in design research. So, the findings indicate that limiting the analysis to a single design activity provides a necessary foundation for using eye tracking parameters for cognitive load measurement in the conceptual design context.

2.3.3 Basics of Machine Learning

Due to the large volumes of eye tracking data generated during design activities, conventional analysis methods may be limited in capturing complex patterns. Machine learning techniques provide a way to process these data efficiently and to combine multiple metrics, supporting a more comprehensive assessment of cognitive load.

Machine learning is a subfield of artificial intelligence that aims to enable algorithms to learn from data and describe it in a mathematical model. This method is particularly useful when rule-based decisions become too complicated (Alpaydin, 2020). The central idea is to develop models that can recognize patterns in examples and are able to generalize them to new data sets. In the field of machine learning, there is a common distinction between supervised and unsupervised learning (Bishop, 2006). In supervised learning, training data with known dependent variables is available; the model learns to assign inputs in the form of independent variables to the corresponding values of the dependent variables. In unsupervised learning, on the other hand, there are no dependent variables; the main focus here is on identifying unknown structures or clusters in the data. A further differentiation is made between classification, in which data points are assigned to discrete classes, and regression, which continuously predicts a dependent variable. (Bishop, 2006) In the context of this work, supervised learning is used, as the dependent variable, cognitive load, is already known from questionnaires and self-reports, providing the necessary labels for training the models.

The steps involved in developing the machine learning algorithm can be described using established procedures for standardized machine learning pipelines. Such pipelines structure the training process into clearly defined phases of data

preparation, feature processing, and classification to ensure reproducibility and methodological traceability (Meisenbacher et al., 2022). The process applied here includes the steps of segmentation, standardization, feature selection, and classification. In addition, the training data set can be upsampled before classification to compensate for possible class imbalances and achieve a more balanced model fit (He & Garcia, 2009). The individual steps are explained in more detail below.

2.3.3.1 Segmentation

Segmentation describes the decomposition of raw data into smaller, coherent segments in order to focus the analysis on relevant structures and reduce complexity. Examples include dividing a signal into time windows or extracting sequences in text data. Classic methods such as thresholding or clustering are used, but also more complex approaches such as semantic segmentation. (Bishop, 2006)

2.3.3.2 Standardization

Since raw data often has different scales and value ranges, standardization is necessary to ensure comparability. Typical methods include standardization to a fixed range or z-transformation, which takes into account the mean and standard deviation of the data. This prevents characteristics with large value ranges from dominating the analysis. (Alpaydın, 2020; Hastie et al., 2009)

2.3.3.3 Feature selection

In many applications, large quantities of potential input variables are available, but only some of them are actually relevant for the prediction. Feature selection serves to exclude irrelevant or redundant variables in order to avoid overfitting (Reunanen, 2003), reduce computational effort, and increase the interpretability of the models. Common methods include filter methods based on statistical indicators, wrapper methods that take model quality into account, and embedded methods such as decision trees that integrate variable selection directly into the learning process. (Guyon & Elisseeff, 2003; Li et al., 2018)

2.3.3.4 Upsampling

Learning with imbalanced datasets poses a particular challenge. When rare classes are severely underrepresented, many algorithms tend to favor more heavily represented classes. Upsampling methods compensate for this imbalance by duplicating instances of the minority class or generating synthetic examples. As an alternative to upsampling, the imbalance can also be addressed by combining it with downsampling of the majority class. (Chawla et al., 2002; He & Garcia, 2009)

2.3.3.5 Classification

Classification is a central approach in supervised learning: data points are assigned to a predefined class based on their feature vectors. Classification methods can be divided into different model families. Linear models such as logistic regression offer robust baseline methods, while support vector machines (SVMs) can map more complex decision boundaries. Decision trees and methods based on them, such as random forests or gradient boosting, are particularly common because they combine high predictive accuracy with interpretable structures. Finally, neural networks enable the modeling of highly nonlinear relationships, but require larger amounts of data and longer training times. (Alpaydın, 2020; Bishop, 2006; Jain et al., 2000)

2.3.4 Machine Learning Algorithms Using Eye Tracking to Measure Cognitive Load

In addition to combining multiple eye tracking parameters (chapter 2.3.1), there are also approaches that use machine learning algorithms to increase prediction accuracy. There are various approaches, ranging from the use of individual metrics (Halverson et al., 2012) to the use of multiple eye tracking metrics (Wu et al., 2020) to the combination with other physiological signals such as EEG signals (Borys et al., 2017) or GSR (Nourbakhsh et al., 2013).

Halverson et al. (2012) provide a particularly detailed overview of this topic by examining the usefulness of various eye movement and pupil metrics for determining cognitive load. To this end, they conducted an extensive parameter study. They broke down time series into overlapping windows of 1 to 30 seconds to smooth the data and classified them using support vector machines (SVM). A total of ten different metrics were used. Pupil diameter and PERCLOS proved to be particularly effective, achieving average accuracies of up to 81 %. However, there was a high degree of variation between participants, with deviations of up to 16 % in accuracy. In addition, the quality of the results depends heavily on the choice of window size: short intervals were particularly suitable for metrics such as pupil diameter, while longer intervals achieved better results with PERCLOS. For other metrics, the optimal conditions shifted again. An overview of the results from Halverson et al. (2012) is shown in Figure 2.7.

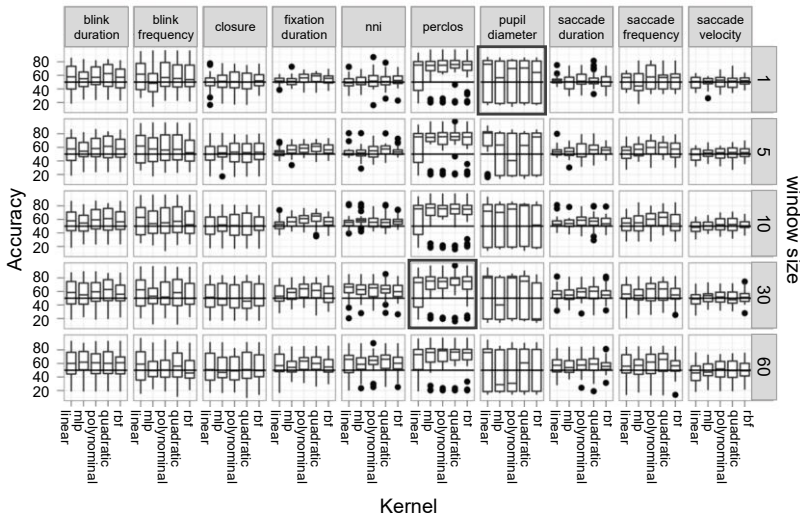


Figure 2.7: Parameter study according to Halverson et al. (2012). This diagram shows the accuracy of cognitive load prediction depending on various eye tracking parameters, evaluation periods, and machine learning kernels. The two highest accuracies were achieved with a 5s window for pupil diameter (linear kernel) and with a 30s window with PERCLOS (quadratic kernel).

At the same time, they point out that combining several metrics, even those with weak individual performance, could be promising. However, a significant limitation arises from the lack of temporal resolution: the short intervals were only retrospectively assigned to the overall cognitive load of the task, so that continuous recording of cognitive load is not possible. This limits the transferability of the results to contexts where continuous measurement is necessary.

Wu et al. (2020) investigated cognitive load in surgery simulations. They used the pupil diameter and gaze entropy, which describes the irregularity of eye movements. Both metrics increased with increasing task difficulty. A classification model based on this achieved an accuracy of 84.7 % (Wu et al., 2020). However, this work is also limited to the analysis of discrete task blocks, which excludes continuous cognitive load measurement and thus limits its transferability.

J. Chen et al. (2019) explicitly took inter-individual differences into account. In flight simulator tasks of varying difficulty, the participants were first segmented into two groups using k-means clustering before a separate SVM classifier was trained for

each group. This approach was intended to integrate physiological differences between individuals into the modeling. (J. Chen et al., 2019) Despite this extension, however, the study only achieved 43.7 % classification accuracy. In addition, there was also a lack of continuous temporal resolution, which limits the usefulness of the results.

Nourbakhsh et al. (2013) combined eye blink metrics with galvanic skin response (GSR) to classify cognitive load during arithmetic tasks at four different difficulty levels. SVM and Naïve Bayes were used as classifiers. This achieved 75 % accuracy in binary classification, while four-class classification achieved only 53.6 %. (Nourbakhsh et al., 2013) Here, too, there was a lack of temporal resolution, as cognitive load was only considered at the level of complete task blocks.

Borys et al. (2017) combined eye tracking data with EEG signals to classify cognitive load during arithmetic tasks. The aim was to distinguish between states of cognitive load and a control condition without a task (no cognitive load). Various classification methods were tested, with SVM with six eye tracking features delivering the best results: 90 % accuracy for binary classification, up to 73 % for multi-class classification. (Borys et al., 2017) The results illustrate the potential of multimodal approaches. However, the results have a significant limitation, as the differentiation is made between task processing and complete rest phases.

In summary, these studies show that various approaches to use machine learning algorithms for measuring cognitive load with eye tracking data have become established. These range from the use of individual metrics (e.g. pupil diameter or gaze entropy) to the combination of several eye tracking parameters to multimodal methods in which eye tracking is combined with other physiological signals such as EEG or GSR. In particular, the work of Halverson et al. (2012) provides important insights into the suitability of different metrics through its comprehensive parameter study and at the same time shows the dependence of the results on the choice of window size. However, all approaches share a central limitation: data is collected exclusively in relation to entire tasks, which means that continuous measurement of cognitive load is not possible. This limitation represents a significant hurdle for the transferability of results to contexts such as design research, where continuous cognitive load measurement is required.

2.3.5 Interim Conclusion

Eye tracking is a widely used method for assessing cognitive load in various domains, with parameters such as pupil diameter, fixations, saccades, and blinks frequently used. However, their applicability to actual design activities remains largely unexplored.

The preliminary work on which eye tracking metrics can be transferred to conceptual design from other research fields showed that established eye tracking parameters follow trends consistent with findings from the state of research. However, no significant correlations were found between subjective cognitive load and the eye tracking parameters. Additionally, substantial differences in visual behavior between the design activities concept generation and concept evaluation demonstrated that task-specific visual behavior strongly influences the measured parameters. These results imply that concentrating on specific design activities is crucial for accurately measuring cognitive load and establishing a basis for further investigations.

In addition to using single eye tracking metrics, algorithmic approaches have demonstrated that combining multiple eye tracking metrics or integrating additional physiological signals can improve the accuracy of cognitive load assessment compared to single parameters. However, most methods still rely on aggregated, task-level data, which limits their ability to capture cognitive load continuously.

2.4 Conclusion on the State of Research

Concept synthesis as a part of conceptual design is a particularly challenging phase with a major impact, as defining the concept significantly restricts the available design space. However, this can create a high degree of uncertainty as to which restrictions are actually necessary and what the consequences of these are. As a result, there are a large number of difficulties that engineering designers can face in this phase. Therefore, there are already a variety of design methods, such as decision methods, to support engineering designers to overcome this uncertainty during conceptual design. One problem with existing design methods is that especially in industry the acceptance towards and the effectiveness of the design methods is considered to be low. One reason for this is that the design methods are not provided according to the individual need.

This need for support has to be determined beforehand to enable design methods to be effective. If in turn the need for the design method is not present in the given situation, the design method may also have negative influences due to lack of

acceptance or not recognizable usefulness. The needs-oriented providing of design methods should not begin to emerge only in the application in practice. Rather, this must already be taken into account by researchers in empirical investigations of concept syntheses as well as in the development and validation of design methods. This issue is a central concern in design research.

In the area of design research, researchers tend to conduct empirical studies of design processes with few participants followed by detailed data analysis. Established research methods and tools, such as protocol analysis or interviews, have the disadvantage of requiring an enormous amount of time and effort per participant for data analysis. This high level of effort is an important limiting factor in conducting larger studies with participants to demonstrate statistical relationships in the context of design methods. Moreover, these research methods have the disadvantage that they are very dependent on the evaluating person and thus do not provide objective insight. The only way to solve this is to work with multiple evaluators, which increases the effort involved dramatically.

To overcome these limitations, new research methods and tools are needed to help researchers improve resource-intensive data preparation and analysis in empirical studies of engineering designers. These research methods must be capable of taking the perspective of the engineering designer into account, providing time-continuous measurement during the process, automatically processing large amounts of data in a scalable manner, and delivering reliable and objective results. In this way, the limitations of traditional research methods could be overcome, thereby allowing larger, statistically grounded studies while reducing the dependence on subjective assessments.

In this context, Cognitive Load Theory offers a promising foundation for identifying individual support needs caused by difficulties in design research. High cognitive load can impair designers' performance and creativity and therefore serves as a measurable indicator of situations in which methodological or tool-based support becomes necessary. While established questionnaires such as the NASA-TLX enable subjective assessment of cognitive demands, their limited temporal resolution and manual data processing restrict their suitability for automated and time-continuous investigations. Accordingly, objective and scalable measures of cognitive load are required to advance empirical research in conceptual design.

This can be achieved with physiological signals such as EMG, ECG, or eye tracking. These are a good starting point to identify difficulties during design. Eye tracking has particular potential here, as this method scarcely influences the participant and is therefore well suited for use in everyday working life. However, when developing

such novel research methods, it is important to combine physiological signals with traditional design research methods such as interview techniques and protocol analysis. In doing so, the suitability of new methods can be evaluated in relation to an established standard. At the same time, the integration of physiological measurements leads to large and complex data sets that can no longer be efficiently analyzed using conventional statistical approaches. This creates a need for algorithmic and machine learning-based research methods capable of identifying patterns and correlations within multidimensional data. Only in this way can the potential of quantitative physiological signals be fully exploited.

In summary, there is a research gap in the field of design research for new types of research methods and tools which, on the one hand, enable an individualized assessment of needs for the use of design methods and, on the other hand, also enable an automated evaluation of the data collected.

3 Motivation and Research Objectives

In the following chapter, the motivation of this thesis is described based on the previously presented state of research and the identified research gap. The aim of this work is then derived from this motivation.

3.1 Motivation

The development of new products in product development is a complex process that constantly challenges engineering designers. Early on in the development process, all further steps in the process have to be thought out in advance and taken into account during concept development. This is further increased by ever more complicated products and interdisciplinary problems. Consequently, engineering designers are more dependent than ever on appropriate support such as design methods.

The development and validation of such design methods is a central topic in design research. However, the use of an unnecessary or not suitable design method, even if its usefulness has generally already been demonstrated, can also have negative effects on the design process. Therefore, understanding when and where engineering designers require design methods for support is essential for the targeted development and application of design methods.

In the state of research, the relevance of an individualized assessment of needs for the use of design methods based on difficulties of engineering designers has been derived in detail. Situations in which engineering designers find themselves in a situation of uncertainty due to a lack of knowledge or experience, and thus have too high cognitive load, play a central role. In these situations, design methods can achieve significant added value through the targeted transfer of suitable knowledge or methodical procedures. However, these relevant situations of high cognitive load, in which there is actually a need for methodological support, are very difficult to identify using established research methods. Through the use of interviews, video recordings, and protocol analyses, the identification of the occurrence and nature of support needs is very time-consuming. In addition, these research methods have the disadvantage that they are very dependent on the evaluating person. To avoid this, evaluation by several persons is necessary to ensure the objectivity of the evaluation.

This can be illustrated by the example of the evaluation carried out for the publication of Nelius et al. (2020). In this study, eye tracking and think aloud protocols were evaluated to identify situations in which designers had a bias in their reasoning (Nelius et al., 2020). The study design applied here included a 20-minute task processing by the participants. Verbal protocols and eye tracking glasses were used for data collection. The data processing consisted of the transcription of the CTA protocols and semantic-gaze-mapping. In the latter, contextual information is assigned to the individual gaze points in the eye tracking video, for example, which component of a system is currently being viewed. The time required for these steps can be roughly quantified as follows for this example: Transcribing and proof reading CTA transcripts took about five times the video duration. The semantic-gaze-mapping required 15 to 20 times the video duration. Thus, the time required for the data processing of a 20-minute video can be estimated at about one working day.

In addition, there is the evaluation of the data, which takes about another day per participant. If the evaluation is not only to be based on the evaluation of one analyzer, the accordance between two or more analyzers can be determined with the help of key values such as the interrater reliability. In this way, the quality of the evaluation can be made more independent of the subjective opinion of a single analyzer. However, this again significantly increases the effort required.

Overall, the effort in a study can be 2-3 days per participant. In a study with 20 participants, this means a total effort of more than 2 months for the generation of results. Scaling up to even larger groups of participants to test statistical correlations is simply not feasible.

To overcome these disadvantages, existing findings from psychology can be transferred and adapted to design research. By quantifying the cognitive load of participants by means of special questionnaires, an objective determination of the cognitive load is already possible, but only for longer periods or separable tasks. Initial approaches to automated assessment of difficulties by measuring cognitive load using eye tracking data already show the feasibility in principle of further automation.

Automating the assessment of need for methodological support by means of algorithms applied to physiological signals makes it possible to overcome limitations imposed by established data collection methods. On the one hand, it can drastically reduce the time needed for the evaluation, and on the other hand, the result is always the same regardless of the researcher, which results in a high objectivity of the evaluation.

Thus, the underlying motivation of this work is to improve objectivity and reduce effort in the identification of difficulties as indicators for support needs for engineering designers during concept synthesis while at the same time enabling larger numbers of participants.

3.2 Research Objectives

As previously stated, there is a need for novel research methods and tools to support design researchers in the evaluation of empirical studies. To this end, an algorithm will be developed that can be used to detect difficulties as indicator for the need for methodological support. In this way, the effort involved in data evaluation and the influence of the researchers on the results will be reduced. To achieve this, eye tracking data will be used as a data basis. Eye tracking is particularly suitable because it provides a quantitative data basis that can be continuously recorded and further processed automatically. In addition, eye tracking data is also very valuable from a design methodological point of view, as they provide an insight into the visual recording of information.

The aim of this work can therefore be formulated as follows:

The aim is to develop a new algorithm that enables an automatable evaluation of support needs in empirical participant studies in engineering design. For this purpose, eye tracking data will be used to survey difficulties in concept synthesis as indicators of need for support.

In doing so, this work seeks to enhance the objectivity, reproducibility, and scalability of empirical design research, thereby contributing to more reliable insights into engineering designers' cognitive challenges and methodological support needs.

4 Research Questions and Research Design

In the following, the overall research question and the research questions are derived from the previously formulated aim. The research questions provide the necessary building blocks for achieving the aim of this thesis. The procedure for answering the research questions is then explained in the research design.

4.1 Research Questions

From the research gap derived in chapter 2.4, it can be derived that new research methods and tools are needed in design research. This need builds on the findings from the two preliminary studies presented in the state of research, which demonstrated that cognitive load can serve as an indicator of support needs and that it can be captured through eye tracking data. The focus will be on the phase of concept synthesis. Derived from the aim of the thesis, the development of an algorithm that enables an automatable evaluation of support needs in empirical participant studies in engineering design, the following overarching research question is derived for this thesis:

How can eye tracking data be used to objectively and automatically measure support needs in the form of difficulties in concept synthesis as part of the design process?

To answer the overall research question, the following research questions are defined:

RQ1: “What difficulties do engineering designers encounter in concept synthesis that can be used as an indicator of support needs?”

RQ2: “How can difficulties in concept synthesis be detected automatically based on eye tracking data?”

RQ1 deals with the nature of difficulties encountered in concept synthesis. To specifically address the difficulties of engineering designers, these difficulties need to be better understood. In particular, the perspective of the engineering designers

themselves is necessary to build this understanding. Therefore, chapter 5 investigates the question of what difficulties engineering designers face during the phase of concept synthesis. To this end, difficulties in concept development are identified through retrospective think aloud. Thus, by answering this research question, the ground truth for the development of an automated evaluation is established.

Based on this, an algorithm is developed and validated in RQ2 (chapter 6) that can identify these difficulties in an automated way using eye tracking data. By distinguishing between different types of difficulties, as investigated in RQ1, it is also considered whether there are differences in measurability based on the type of difficulty encountered. Thus, both the difficulties as a whole and each one individually is studied. In this way, an algorithm is developed that allows for a reduced-effort evaluation of empirical studies in design research by means of computer-based data collection and subsequent automated evaluation.

4.2 Research Design

This chapter describes the procedure for achieving the overall objective. To achieve this objective, an algorithm will be developed that can automatically identify difficulties based on eye tracking data. Therefore, a first study will be conducted in chapter 5 to answer RQ1. The aim of this study is to identify difficulties during concept synthesis. Then, in the second study the data basis for the development of an algorithm is presented in chapter 6.1.1. Based on this data, an algorithm is trained and validated in chapter 6 to answer RQ2. The designs of the individual studies to answer the research questions are presented in the respective chapters 5 and 6.

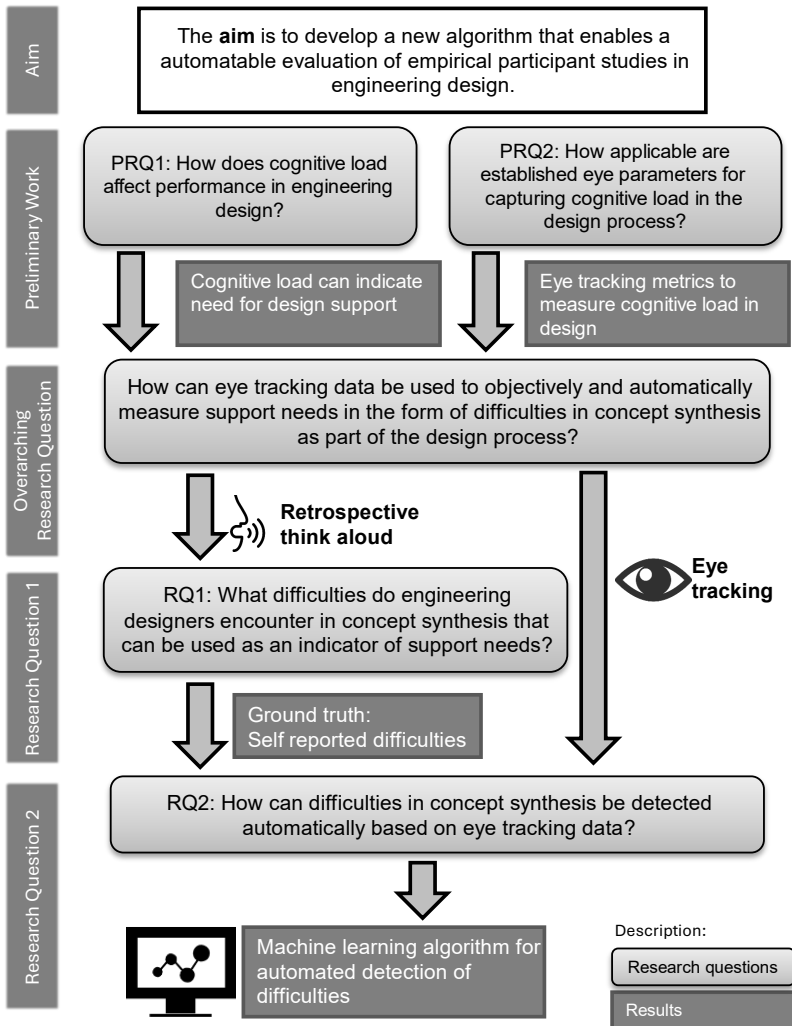


Figure 4.1: Research design to develop a machine learning algorithm for the automated detection of difficulties based on eye tracking data, validated using self-reported difficulties as ground truth

The data needed to answer the research questions is obtained from two studies that follow an identical study design and only slightly adjusted data collection. In the first step, a qualitative study is conducted to answer RQ1 by identifying difficulties that arise during concept synthesis. For this purpose, the participants completed different tasks from the field of concept synthesis. Retrospective think aloud is used for data collection here. The protocols were then analyzed using established research methods to determine which difficulties they recognized in the process. The difficulties identified in this way serve as a baseline for the subsequent algorithm development. The study is conducted with students because their greater homogeneity reduces the confounding effects caused by different professional backgrounds, providing a solid basis for algorithm training. To ensure that the identified difficulties are not limited to this group, an additional comparison is carried out with difficulties reported by experienced engineering designers from the state of research. This approach ensures the reliability of the findings while maintaining the methodological advantages of using a student-based sample for algorithm development.

The second study builds directly on these results and provides the necessary data basis for algorithm development, which addresses RQ2. Here, the previously identified difficulties are combined with eye tracking data to develop an algorithm capable of detecting such difficulties automatically. Furthermore, the findings on the pertinent eye-tracking metrics from the PRQ2 (chapter 2.3.2) are incorporated. For algorithm development, the collected dataset is divided into two parts: one part is used to train the algorithm, while the other is employed for validation in order to assess its accuracy. For this purpose, the difficulties identified by the algorithm are compared with the self-reported difficulties of the participants. In this way, the two studies are connected in a continuous process: the qualitative insights from the first study inform the second study, which in turn delivers the empirical foundation for algorithmic development.

5 Difficulties in Concept Synthesis

Findings obtained through preliminary work have shown that cognitive load can serve as an indicator of when and where engineering designers experience difficulties during design activities and need design support (chapter 2.2.3). Furthermore, eye tracking has proven to be a suitable method for capturing such cognitive load objectively and continuously. However, cognitive load is strongly influenced by the nature of the task and the specific demands of the design situation (chapter 2.3.2). Consequently, before automated research methods can be developed, it is necessary to understand which types of difficulties actually occur during concept synthesis and how they manifest in designers' cognitive processes. Therefore, the first research question focuses on understanding the nature of the difficulties that arise specifically during concept synthesis. The perspective of the engineering designers themselves is particularly important here, as only they can provide direct insight into the challenges and barriers that occur during their work in this early phase of the design process. Such an understanding forms the basis for deriving requirements for new methods and tools as well as for developing automated evaluation procedures.

Therefore, the research question to be answered here is:

RQ1: "What difficulties do engineering designers encounter in concept synthesis that can be used as an indicator of support needs?"

The development of algorithms, however, requires extensive data sets that are difficult to generate with professional engineering designers alone. Against this background, it is proposed to make greater use of data from engineering design students, as this group allows for larger samples and greater homogeneity (Kroll & Weisbrod, 2020). To answer the research question, a laboratory study was chosen as the study environment. According to Kroll and Weisbrod (2020), laboratory studies have several advantages compared to field studies, depending on the research objective. On the one hand, it can be better ensured in laboratory studies that the participants concentrate fully on the tasks set during the processing time and are not distracted by other activities (Blessing & Chakrabarti, 2009). Furthermore, this group allows for larger samples and greater homogeneity (Kroll & Weisbrod, 2020). In field studies, on the other hand, it is much more difficult or even impossible to exclude potential distractions. In addition, control over the process under investigation is much more difficult in field studies and they usually consider

longer periods of time. Due to these long phases of data collection, biases may occur more likely, as participants here usually report retrospectively over a longer period of time, which may lead to inaccuracies.

However, this raises the key question of whether students face similar difficulties in the design process as experienced engineering designers. Answering this question is crucial to ensuring the validity of future studies with sample sizes large enough to yield statistically significant and reliable results.

Considering this and to answer the research question in the context of the research objective, the following chapter focuses to first identify the difficulties students experience during concept synthesis. These are collected using retrospective think aloud protocols. To ensure that the identified difficulties are not only characteristic of the student sample but also relevant for practitioners, a comparison with difficulties reported by experienced engineering designers is part of the investigation. For this purpose, the identified difficulties of the students are compared with findings on difficulties experienced by experienced engineering designers based on the state of research as examined in chapter 2.1.2. This comparison allows assessing the extent to which difficulties observed in professional practice can be replicated in a controlled laboratory environment with students, thereby validating the use of student data for future studies with larger sample sizes.

Chapter 5.1 describes the procedure followed in the study. Chapter 5.2 then presents the results of the study, which are discussed in chapter 5.3.

The following subchapters 5.1 to 5.3 are based on the publication “Challenges in conceptual design: A comparison between engineering design students and experienced designers” (Zimmerer & Matthiesen, under revision). The text has been taken from the publication with minor modifications, especially with regard to numeration, formatting, and terminology. For the sake of clarity, the texts have not been marked separately.

5.1 Methodology

This chapter describes the experimental design of the study (chapter 5.1.1) that was carried out, together with the procedures for data collection and data analysis (chapter 5.1.2).

5.1.1 Experimental Design

A total of 12 participants from the engineering courses at KIT - Karlsruhe Institute of Technology took part in the study. The participants were at the end of their bachelor's or in their master's degree and each had already completed the basic lectures on engineering design.

The participants were between 20 and 27 years old (mean = 23.8, SD = 1.9). The participants had a background in three different design courses but participated in the same lectures and exercises in engineering design (5 engineering designers, 6 mechatronics engineers, and 1 engineering pedagogue). The total duration of study ranged from 6 to 14 semesters. Given their background in mechanical engineering or mechatronics, as well as their advanced status in the course of study, these students possess a solid foundation in the principles of conceptual design and engineering design. This prior knowledge ensures that the participants are well-equipped to engage with complex design tasks, similar to those encountered by young professionals, providing a reliable basis for comparing their difficulties with those faced by experienced practitioners.

At the beginning of the study, all participants had to confirm their participation in the study in a declaration prepared by the KIT data protection office. Approval was obtained from the local institutional review board (IRB) prior to the study. This included measures to ensure the data protection and privacy of the participants. Cooperation with the data protection office of the KIT was an essential part of establishing protocols to protect participant data. In addition, the study was registered in KIT's electronic processing directory (eVV) to ensure transparency and traceability. A risk analysis was conducted to identify and mitigate potential concerns and ensure that the safety and rights of all participants were secured throughout the research process.

Subsequently, the participants were accompanied to their place and the procedure of the study was explained to them. This consisted of the processing of the tasks in randomized order. This should ensure that there were minimal learning effects between the different tasks.

Sheet metal design was chosen as the context for the tasks. This manufacturing process had the advantage that it had been covered in the mechanical engineering lectures, but not in great detail. In addition, sheet metal design contains a large number of individual process steps, such as laser cutting, bending and welding. As a result, many production-related boundary conditions must already be considered in the conceptual design phase.

Three different tasks were used in the study, with an existing system in each task presented to the participants. The task in each case was to come up with alternative concepts for the given assembly. The individual tasks are described in more detail below.

- **Design Task 1:**

The first design task focused on redesigning an existing anchoring foot. The given prototype (see Figure 5.1) had been realized as a welded design consisting of nine individual components. For future large-scale production, various concepts for a cost-effective sheet metal bending design should be generated based on this initial design.

To ensure functional equivalence, several boundary conditions had to be maintained: the hole pattern on the base plate and the bolt holes had to remain unchanged in terms of position and quantity, while the load-bearing capability in all three spatial directions had to be preserved.

Participants were asked to sketch three alternative design concepts, each including 3D representations and corresponding flat patterns, to translate the given welded assembly into a manufacturable sheet metal solution. The optimal design was expected to consist of only two parts, and participants were encouraged to approximate this target as closely as possible.

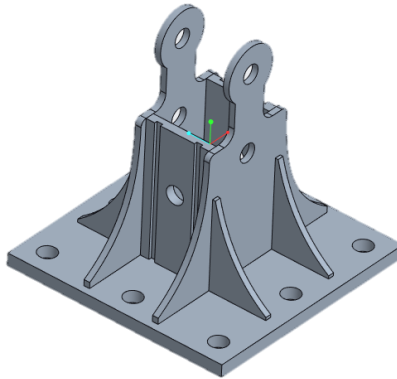


Figure 5.1: Design Task 1: Development of alternative concepts for the realization of the shown bracket as a sheet metal bending design. The task setup aims at the generation of alternative solutions during the conceptual design stage.

- **Design Task 2:**

The second task addressed a subassembly of a commissioning unit that had been previously prototyped using laser-cut high-density fiberboard (HDF) shown in Figure 5.2. During initial functional testing, operational issues were observed: due to the geometry of the boundary surfaces (Figure 5.2 number (2)), multiple cubes were occasionally dispensed simultaneously, resulting in an incorrect sorting sequence.

For the next design iteration, the participants were instructed to redesign the unit as a sheet metal bending design while maintaining the existing motor interface. In addition, the new design was to be cost-efficient and fully manufacturable from sheet metal.

The task required participants to generate conceptual solutions addressing the jamming problem while transforming the prototype into a functional, production-oriented design. Motor representations were not required, but the mechanical interface had to remain consistent with the given prototype.

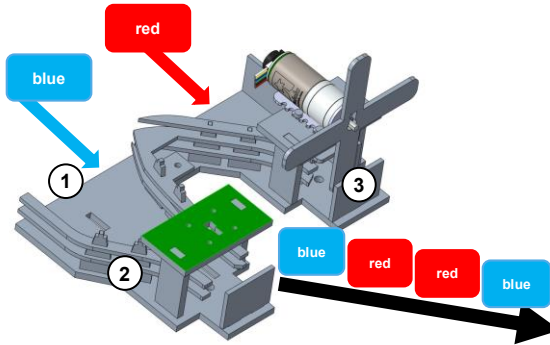


Figure 5.2: Design Task 2: The assembly shown should be redesigned as a sheet metal bending design. The cubes (blue and red) are fed in (1), lined up (2), and then transported further in a specific order by two motor-driven rotors (3).

- **Design Task 3:**

The third design task involved improving an existing latch mechanism used for drawer locking. The initial design shown in Figure 5.3, already realized as a sheet metal bending design, exhibited significant wear due to multiple moving components. To address this issue, the goal of the task was to reduce the number of parts that move relative to one another by means of a geometric and structural redesign.

The reference design consisted of five sheet metal parts with a material thickness of 2 mm, nine bending operations, and no welded joints. The required stroke of the latches was approximately 5 mm.

Participants were asked to develop and sketch alternative latch concepts that achieve the same locking functionality with fewer parts, thereby improving durability and manufacturability.

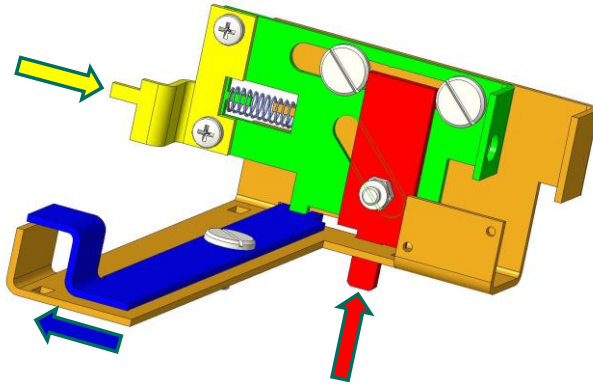


Figure 5.3: Design Task 3: The number of individual parts in the latch shown should be reduced. The movements indicated by arrows should be retained. By moving the blue lever in the direction shown, the two locking pins (yellow and red) should slide in.

Each of the three tasks had a defined sequence. First, the participants were fitted with a head mounted camera system. The purpose of this was to get a video recording of the task processing from the participant's point of view. Subsequently, the participants received a printed version of the task.

They were able to read through the task without time constraints. After the participants confirmed that they had understood the task, the processing phase began and the recording was started. This lasted 10 minutes for each task and consisted of the participants drawing alternative concepts for the respective task on paper.

At the end of the processing time, the recording was stopped, and the participants were asked to stop drawing the concepts. To identify difficulties, the participants were now shown their previously recorded video again on a screen using retrospective think aloud (RTA) techniques.

They were asked to comment on the video what they had just thought. The focus here was on situations in which they had difficulties. Meanwhile, the difficulties mentioned were written down by the researcher. In addition, a supplementary audio recording was made. In this way, the exact wording of the participants could still be accessed during data analysis. In addition to the type of difficulties mentioned, the time stamps for the beginning and end of the respective situation were also recorded.

5.1.2 Data Analysis

For the analysis of the data, the statements regarding difficulties encountered noted during the retrospective commenting of the recording were available for each participant. In the first step of the analysis, the statements of all participants were collected.

The categorizing of the difficulties was done in an iterative procedure. In particular, attention was paid to similar key words, on the basis of which the difficulties of the participants could be connected according to different topics.

Subsequently, in consecutive steps, all similar difficulties were summarized and an overarching term for each kind of difficulty was chosen. Only categories that contained at least three difficulties named by the participants were considered. Difficulties that had already been assigned were no longer considered, so the database of difficulties decreased continuously during the analysis. Finally, the categorization was validated during a structured review session with independent researchers. Ambiguous or borderline cases were discussed in detail, and adjustments were made to refine the assignment of difficulties. This iterative validation process established a clear, consistent, and reproducible categorization scheme, providing a solid foundation for further analyses.

5.2 Results

The data of all 12 participants could be used for the evaluation. In total, 126 difficulties were encountered by the participants across all three tasks. All participants were able to work out a solution in each task.

Most difficulties occurred in task 1 (50) followed by task 3 (39) and task 2 (37). The individual participants had different numbers of difficulties. While participant 9 reported a total of only 2 difficulties during the three tasks, participants 4, 5 and 11 had reported considerably more situations with 15 difficulties each.

The distribution of the number of difficulties encountered across all participants is shown in Figure 5.4.

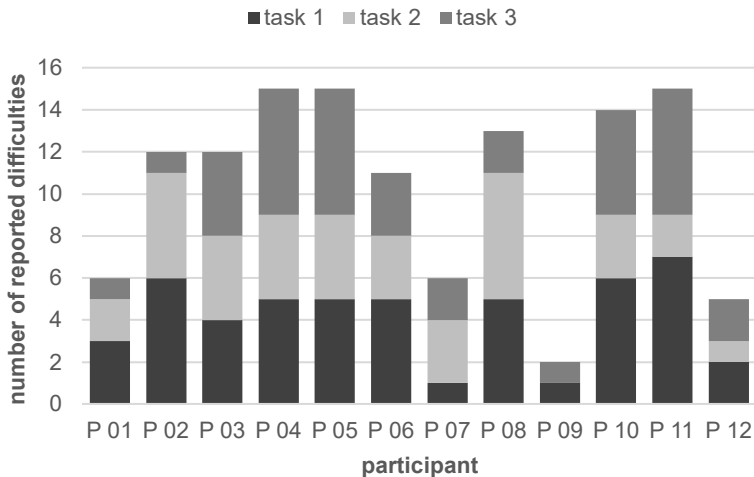


Figure 5.4: Distribution of the number of difficulties occurring per participant (Zimmerer & Matthiesen, under revision)

5.2.1 Difficulties in Detail

The identified difficulties were very different in their characteristics. This also concerns the duration of the respective difficulties as well as their implication.

While some participants had found a solution to their problem themselves after a very short time, there were other situations in which the difficulties had greater implications. To illustrate this in more detail, a few statements from participants are shown in detail below.

Example 1 (Participant 2): (7:25, duration: 02:25 min)

“The next major problem is with the 3D sketch. To find out, to imagine spatially how exactly something runs, um, because it’s a bit more complicated geometry from the base area.”

Example 2 (Participant 3): (0:55, duration: 1:15)

“Then I started to draw something, but then I was unsure whether I should continue there and then I discarded it. At that time, I was still thinking about

whether I could put something together, so how these components can be bent.”

Example 3 (Participant 3): (2:30, duration: 0:20 min)

“I had then looked again at the drawing. Whether I could maybe save what I had drawn. Yes, I had then at 2:10 I crossed it out, so discarded the attempt. Then I started with a 3D sketch, but then it was easier for me to start with a 2D sketch, which is why I discarded the 3D sketch.”

Overall, it can also be observed that the individual difficulties differ to some extent in their duration. However, the majority of the difficulties had a duration of less than one minute.

5.2.2 Clusters of Difficulties

As a result of the evaluation, eight different clusters of difficulties were identified by clustering the participants’ statements. A detailed description of the various clusters and the frequency with which they occurred can be found in Table 5.1:

Table 5.1: Results of the evaluation of the participants' statements. The difficulties encountered can be divided into eight different clusters.

Title	description	frequency
Visualization	Difficulties frequently arose when visualizing concept ideas on paper, both in 2D and 3D representations. These challenges were primarily related to translating abstract ideas into sketches.	45
no concept idea	The difficulty lack of concept ideas occurred when the participants had already created initial concepts but then could not get any more ideas about another alternative.	24
embodiment design	Difficulties in embodiment design occurred when the participants already had an initial idea for an alternative concept in mind but had	15

		difficulty translating this idea into an embodiment.	
manufacturing process		Difficulties regarding the manufacturing process were - also due to the context of the tasks used - situations in which the participants did not know exactly what a design suitable for sheet metal design looks like. These included, for example, problems in evaluating the bendability of the designed concepts.	14
general overload	overload	General overload was defined as situations in which the participants had a fundamental problem. This could be, for example, that the participants did not know how to approach the task at all or how to proceed to come up with a solution in general.	5
assembly		The assessment of assembly feasibility emerged as a recurring challenge. Participants examined whether the designed components could be assembled as intended. The evaluation often focused on criteria such as the accessibility of screw joints or weld seams.	3
function unclear	unclear	The difficulty function unclear was attributed when the participants had difficulty understanding how the given design worked. For example, when the connections of the movements of the individual parts were not understood.	3
concept selection		Difficulties in concept selection are a special case since it was not part of the tasks to decide on a favorite concept at the end of the task. Nevertheless, there were some participants who tried to select the best of the concepts, which they had created. However, the participants sometimes had difficulties in doing so because they did not know the relevant criteria for evaluating the concepts.	3

5.3 Discussion

In this chapter, the findings regarding difficulties encountered during concept synthesis are analyzed and discussed. The discussion is structured into three parts. First, the difficulties reported by the engineering design students are presented and categorized (chapter 5.3.1). Next, these findings are compared with the difficulties reported by experienced designers (see chapter 2.1.2) to identify similarities, differences, and potential areas for targeted support (chapter 5.3.2). Finally, implications for design research and the development of support methods are derived (chapter 5.3.3).

5.3.1 Difficulties of Engineering Design Students

The aim of this study was to compare the individual difficulties of engineering design students with those of experienced designers. For this purpose, the participants were asked to retrospectively report their approach to three tasks from the conceptual design area after completing the task. Different types of difficulties were identified, which can serve as indicators of support needs in concept synthesis. Specifically, eight distinct categories were identified based on retrospective self-reports: *'assembly'*, *'concept selection'*, *'embodiment design'*, *'function unclear'*, *'general overload'*, *'manufacturing process'*, *'no concept idea'*, and *'visualization'*. There were some, like *'no concept idea'* that are already extensively addressed in the state of the research. One example here is the ideation phase, in which the goal is to increase the quantity and quality of concepts during ideation (e.g. Shah et al., 2003). For conceptual design, design methods are predominantly developed that support creativity. This is intended to increase the number of concepts developed and thus also the quality of the product to be developed (Jin & Benami, 2010; Shah et al., 2003). However, the results of this study show that not only creativity plays an important role in the conceptual design phase, but also other aspects. Here, the ability to visualize the developed ideas (*'visualization'*) and the knowledge about the requirements of manufacturing processes (*'manufacturing process'*) to be used are to be mentioned above all. This supports the statement that each engineering designer should be supported individually and according to the situation (Badke-Schaub et al., 2011).

The requirements of manufacturing processes also play an important role. Here the argumentation is that only a well thought out concept can be implemented well. If, on the other hand, the subsequent steps such as manufacturing and assembly are not sufficiently considered in a developed concept, this leads to expensive iterations in these phases.

5.3.2 Comparison to Experienced Designers

Doellken et al. (2020) identified as key challenges in sheet metal design: Creating new concepts and designs, considering manufacturability, and evaluating and detailing designs. To align with these challenges faced by experienced designers, the study employed three tasks related to sheet metal design. These tasks covered conceptual design as well as the optimization of existing systems. For the latter, both transfer a component into a sheet metal design and to further optimize a sheet metal design were covered. By focusing on these tasks, it was ensured that the students encountered problem-solving situations similar to those reported by the experienced designers, allowing for a meaningful comparison of the difficulties faced.

The discussion about the difficulties in the design process reveals both similarities and differences between students and experienced designers. Both groups face difficulties, especially if the domain-specific knowledge is not sufficiently available. Here, the difficulties identified for students can be brought together very well with those four kinds of difficulties of the experienced designer (Doellken et al., 2020). The following description follows the four types of difficulties presented in Chapter 2.1.2.

First, a lack of know-how regarding the chosen manufacturing process emerged as a central difficulty. Similar to the experienced designers, who reported missing process knowledge, particularly concerning sheet metal manufacturing, the students also struggled to select and apply suitable manufacturing methods. Difficulties such as '*manufacturing process*' and '*assembly*' reflect comparable challenges in anticipating how individual parts could be feasibly produced and assembled.

Secondly, transferring design objectives into a structure reconciled with the manufacturing requirements was a challenge. Doellken et al. (2020) describe how experienced designers must weigh up objectives with manufacturing constraints, which often leads to iterative redesign. Likewise, students reported difficulties such as '*no concept idea*', '*function unclear*', and '*embodiment design*'. These issues demonstrate uncertainty in linking abstract functions to physical structures while simultaneously considering manufacturability and cost constraints.

Thirdly, analyzing manufacturability and predicting feasibility proved to be another shared challenge. While experienced designers reported problems in assessing the manufacturability of their concepts, the students similarly struggled to anticipate assembly feasibility ('*assembly*') or detect potential manufacturing conflicts ('*manufacturing process*') early on. These parallels suggest that the evaluation of

manufacturability remains a cognitively demanding and error-prone process that depends on both expertise and contextual understanding.

Fourthly, both groups experienced difficulties in '*concept selection*'. According to Doellken et al. (2020), a key reason for this is the lack of knowledge about relevant parameters for evaluating concepts in terms of manufacturing effort and performance.

In contrast, certain difficulties such as '*general overload*' and '*visualization*' appear to be more specific to students. These challenges are less related to domain-specific knowledge and more to a lack of routine in applying design strategies and externalizing ideas effectively.

Although the tasks in this study do not reflect the full complexity of large product development projects in practice, they represent specific, well-defined subtasks integral to such processes. The study focuses on a specific aspect of product development to enable individual feasibility and create a controlled, focused level of difficulty. Task difficulty was ensured by precise formulation tailored to the participants' qualifications and the defined research framework. This approach takes into account basic quality criteria in order to increase the empirical relevance and reliability of the results (Bernd Bender, 2003). In this way, a realistic but controllable environment was created that made it possible to investigate and compare the challenges of prospective engineering designers during concept synthesis.

To summarize, it can be said that the difficulties in the design process of students and experienced designers are quite comparable, especially when it comes to the lack of domain-specific knowledge and the difficulties in dealing with different manufacturing processes. This lack of knowledge is not unique to engineering design students; even experienced designers can face similar difficulties when confronted for example with new manufacturing techniques or unfamiliar technologies. For instance, when new manufacturing processes or materials are introduced, designers may lack the specific experience or expertise to fully assess their impact on the product design. This is in line with the uniqueness of design processes (Albers, 2010), which means that knowledge must always be acquired anew in every design process. Such situations highlight that the challenge of dealing with novel processes is not confined to those with less experience but is rather a common difficulty across all levels of design practice. The similarities also extend to the complexity of concept development and the synthesis of structures, where design goals must be reconciled with manufacturing requirements. Nevertheless, there are specific nuances, such as the general excessive demands and

visualization difficulties, which are more closely related to the students' level of experience.

5.3.3 Implications

Considering the similarities in difficulties between students and experienced designers, a promising approach for the development of automated research methods opens up. The similarities between the two groups allow the advantages of both (the relevance to practice of designers and the comparability and high availability of students) to be combined. A stepwise approach is important. Here, the initial focus was on one specific area, challenges in sheet metal design, to identify and examine the difficulties in detail. By starting with a focused scope, solid insights and methods can be developed. Once these challenges are understood, the research can be expanded to other design domains, allowing the models to be adapted and refined to capture a broader range of difficulties across different contexts. This incremental approach ensures that the methods developed are robust, scalable, and applicable to various design problems, ultimately contributing to a deeper understanding of the design process as a whole.

Also, advanced AI and data analysis techniques could be used to recognize precise patterns and isolate specific difficulties in the design process. Similar approaches have already been successfully applied in other research fields to analyze eye-tracking data (see chapter 2.3.4) and can provide valuable methodological insights for this context. The advancing development of automated research methods requires extensive data sets that comprehensively capture the design behavior of students and experienced designers in different domains. These datasets serve as the basis for training models, which in turn can help extract patterns and thus gain deeper insight into common difficulties. The diversity of these datasets is crucial to ensure the adaptability and transferability of the developed models.

Overall, the consideration of similar difficulties in the design behavior of students and experienced designers illustrates the promising direction for future research in the field of automated research methods in the design process. The continuous collection and analysis of extensive, diversified data sets will make a decisive contribution to the development of effective models that not only improve design practice but can also serve as a basis for further studies and innovative teaching approaches.

5.4 Interim Conclusion

This chapter established the data basis for the development of automatable research methods. To this end, the creation of a large and homogeneous database should be facilitated. The focus was on answering the research question:

RQ1: *“What difficulties do engineering designers encounter in concept synthesis that can be used as an indicator of support needs?”*

To answer this question, a study was conducted in which participants completed tasks related to concept synthesis. Their subjective experiences were then recorded using retrospective think aloud (RTA) techniques, as this allowed for a systematic approach to access the difficulties encountered without interfering with the task completion process itself. On this basis, the reported difficulties could be analyzed, categorized, and compared with results from studies with experienced engineering designers (see chapter 2.1.2). Three key findings can be derived from this:

- the difficulties reported by the participants can be clustered into 8 categories,
- the difficulties encountered by students are comparable to those faced by engineering designers.
- the results provide a consistent data basis for algorithm development,

The first finding concerns the nature of the difficulties encountered. Eight different categories were identified with the help of retrospective self-reports:

- *‘assembly’*,
- *‘concept selection’*,
- *‘embodiment design’*,
- *‘function unclear’*,
- *‘general overload’*,
- *‘manufacturing process’*,
- *‘no concept idea’*, and
- *‘visualization’*

Systematizing the difficulties is significant as it enables comparisons between groups of participants and provides a consistent setting for analyzing future studies.

Secondly, the comparability of students and practitioners and the categorization of the reported difficulties now provide a consistent data basis that can be used for the

development of automated procedures. Although the RTA method proved to be suitable, the survey remains very time-consuming and limited in its applicability to larger samples. The categories developed in this chapter form the basis for enabling automated identification using eye tracking data. This creates the basis for continuing the findings of this chapter algorithmically.

The third finding relates to the comparability of the reported difficulties. The results show that this is largely the case. This proves that students are a useful alternative for generating data on a larger scale. This is important as engineering designers are difficult to recruit, while students are more readily available and more homogeneous in terms of their educational background. This homogeneity increases internal validity, as confounding factors due to different professional experiences are reduced. On this basis, new opportunities for obtaining large, consistent data sets are opening up.

By identifying different clusters of difficulties that arise and comparing them with engineering designers, research question 1 “What difficulties do engineering designers encounter in concept synthesis that can be used as an indicator of support needs?” can thus be considered answered.

6 An Algorithm to Detect Difficulties in Concept Synthesis

Building on the difficulties identified in chapter 5, this chapter examines how these difficulties can be detected automatically. The starting point is the problem introduced in chapter 2.1: established methods in design research, such as protocol analyses, interviews, or the evaluation of eye tracking videos, rely on manual interpretation and evaluation to identify difficulties during the design process. This effort limits the number of participants who can be included in studies, making it difficult to statistically validate research results. This is particularly evident in studies that analyze engineering designers' cognitive processes during concept synthesis, as these require detailed, timely, and objective recording.

The research question to be answered is:

RQ2: “How can difficulties in concept synthesis be detected automatically based on eye tracking data?”

In this context, RQ2 is developing an algorithm that automatically identifies participants' difficulties based on eye tracking data. The aim is to identify situations with high cognitive load as indicators of difficulty in the design process. This approach reflects the subjective stress experienced by participants while avoiding interpretation bias by external evaluators.

To achieve this aim, various classification approaches based on machine learning are being tested. This way, an algorithm is to be developed that allows empirical studies in design research to be evaluated with reduced effort, the number of participants to be increased, and the robustness of the results to be enhanced.

Chapter 6.1 describes the data set used (chapter 6.1.1), the algorithm development process (chapter 6.1.2), and the scoring metrics used (chapter 6.1.3).

The results of the algorithm development are then presented in chapter 6.2. It starts with a parameter study conducted beforehand (chapter 6.2.1), followed by the results of the binary (chapter 6.2.2) and multibinary (chapter 6.2.3) classification.

Finally, the results are discussed in relation to the research question in chapter 6.3.

The following subchapters 6.1 to 6.3 are based on the publication “Towards an automated evaluation of design processes - an algorithm to predict critical situations during concept synthesis” (Zimmerer et al., 2025). The text has been taken from the publication with minor modifications, especially with regard to numeration, formatting, and terminology. For the sake of clarity, the texts have not been marked separately.

6.1 Methodology

In the following chapter, the data set, consisting of eye tracking data and difficulties of the participants, and its creation are presented first (chapter 6.1.1). This is followed by a description of the procedure used to develop machine learning algorithms to answer the research question (chapter 6.1.2).

6.1.1 Dataset

The data set used to train the algorithm was obtained in advance as part of a participant study. The participant study design based on the study described in Chapter 5, involving more participants and including eye tracking. Therefore, the actual study will only be briefly described below, with the focus instead on presenting the data set used for algorithm development.

6.1.1.1 Participants

In the dataset used, a total of 28 student participants were involved, with an average age of 23.5 years ($SD = 2.3$), ranging from 20 to 29 years old. Regarding their experience in the field of design, the mean value was 9.3 semesters ($SD = 3.4$), ranging from 5 to 14 semesters of experience.

Among the participants, 21 had a background in mechanical design, while 7 were in the mechatronics course. 13 participants already had their bachelor's degrees. In terms of vision, all participants either had normal vision or vision corrected to normal. Prior to their involvement in the study, all participants were required to affirm their participation through a declaration provided by the KIT data protection office.

6.1.1.2 Tasks

The same tasks as in the first study were used. These are described in detail in section 5.1.1. Each participant was given all three different tasks to complete.

6.1.1.3 Data Collection and Data Set

Again, the three tasks were completed in a randomized order to minimize sequence effects. First, the participants read through the respective task. Once they had understood them, they began working on the task, drawing alternative concepts on paper while this time their eye movements were recorded using eye tracking glasses, the Tobii glasses 2. Each task took 10 min to complete. After completing the task, the participants watched their own video recording and commented retrospectively on the situations in which they had difficulties and the types of difficulties they encountered.

The participants themselves documented the difficulties they encountered on a standard form. The participants received this form after completing each task and filled it out while watching their own video recording of the task again. They could pause and skip back and forth as they wished. These self-reported difficulties served as the primary data source.

The categories of difficulties identified in chapter 5 were used. Figure 6.1 provides an overview of the number and distribution of the individual categories of difficulties that made up the final data set for algorithm development.

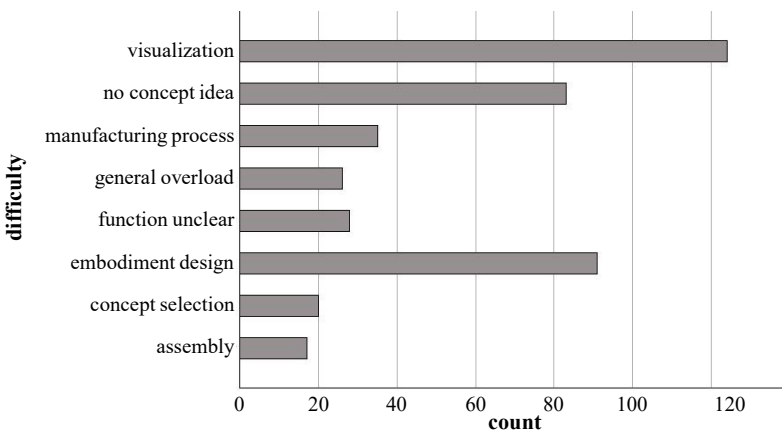


Figure 6.1: Distribution of the 439 difficulties reported by the participants (n = 28) in the data set (Zimmerer et al., 2025)

The raw data, recorded by the Tobii eye tracking glasses at a sampling rate of 100 Hz, were first checked for completeness. Data sets with more than 25 % missing data points were removed entirely. This resulted in a data set of 81 usable recordings.

In the initial analysis of the time series recorded by the Tobii glasses, segments with more than 50 consecutive missing values, equivalent to half a second, were deleted. If the segments of consecutive missing data were shorter than half a second, the missing values were linearly interpolated. The data were then divided into two categories: pupil-based and event-based.

The pupil-based features, recorded directly by the glasses, included gaze point, gaze direction, pupil position, and pupil diameter. First and second derivatives were also calculated to resolve spatial dependencies of individual metrics.

The event-based features encompassed eye movements, categorized as fixation, saccade, and blink. These were identified using the Tobii I-VT (fixation) algorithm (Olsen & Matos, 2012), which specifies the duration of the event and calculates the fixation point in the case of fixation. A detailed description of the algorithm can be found in Olsen and Matos (2012). Several motion-based features were derived from the event classification. For each motion type, the number of events, the occurrence rate, and the variance were calculated for the given windows. In addition, the percentage of time the eyes were closed [PERCLOS see (Dinges & Grace, 1998; Wierwille et al., 1994)] was calculated for each time segment.

Building on the features described, their applicability to design studies lies in their ability to capture nuanced aspects of designers' cognitive and visual behavior during complex tasks. Pupil diameter and PERCLOS are generally known as indicators of cognitive load, helping to identify moments when designers are experiencing high workload. Fixation metrics, such as duration and variance, provide insights into how designers allocate their visual attention and saccade metrics can be linked to the evaluation of specific design elements or problem areas. By analyzing these variables in the context of specific design phases or activities, researchers can map physiologic responses to cognitive difficulties or decision-making moments, thus making the data directly relevant and interpretable for design studies.

6.1.2 Algorithm Development

This chapter describes the procedure for the algorithm development as it is described in chapter 2.3.3, consisting of data preparation followed by classification. In data preparation, the combined data from eye tracking and RTA is processed

further. First, the data is segmented, the time series is divided into windows and the relevant features are calculated. The data is then standardized to ensure consistent scaling. Feature selection is used to identify the most important features to reduce the complexity of the model. To counter the imbalance of the distribution of classes, upsampling is used next. After data preparation, classification takes place, which includes both binary and multibinary classifiers. Figure 6.2 summarizes the procedure below.

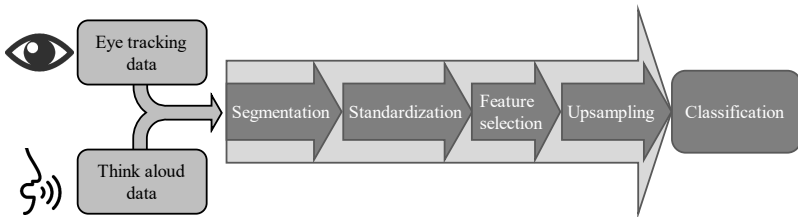


Figure 6.2: Algorithm training process: data preparation and classification (Zimmerer et al., 2025). The process involves sequential training steps: segmentation, standardization, feature selection, upsampling to address class imbalance, and classification.

6.1.2.1 Data Preparation

During segmentation, the time series were prepared so that the features for the respective windows could be extracted. In this study, the recording time series are divided into shorter windows for which the respective features can then be calculated. Variable parameters here are the window length in seconds and the overlap of the windows in percent. This procedure enables the classification of individual segments and the determination of the exact times at which participants encounter difficulties. A schematic representation of this procedure is shown in Figure 6.3.

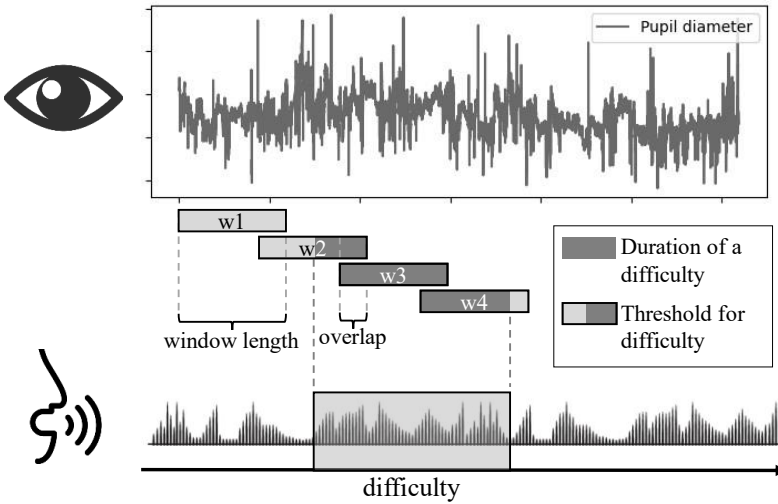


Figure 6.3: Schematic procedure for segmentation. (Zimmerer et al., 2025)
 This involves dividing the time series into individual overlapping windows and then assigning whether there is a difficulty in the respective windows. The threshold specifies the minimum proportion of difficulty in a window required for the window to be labeled as a difficulty. The overlap describes how much the individual windows overlap, and the duration indicates the total duration of a difficulty from start to finish.

After segmentation, the data was standardized. This is a fundamental requirement for classifiers to ensure good classification quality. This process ensures that the features are comparable and do not have large influences due to their scaling, which is especially important for distance-based algorithms. For each feature type and its derivatives, the mean, median, and maximum were calculated within the given time windows. To eliminate individual differences between the participants and the environmental conditions, the difference and the quotient of the individual data points with the mean value of the entire recording area were also used as additional features.

During feature selection, the most important features were determined that had the greatest influence on the predictive ability of the model. The aim of feature selection is to reduce model complexity, avoid overfitting and reduce calculation time by removing less relevant or redundant features. First, a variance analysis was

performed to identify and remove features with too little variance, as these are irrelevant for the classification. The feature spaces were then transformed into a normal distribution using a transformer, as some of the used classifiers require this for good functionality. The RobustScaler from scikit-learn (Pedregosa et al., 2011) was used here, as it is less susceptible to outliers and delivers more robust results. In a further step, the number of features used for training was reduced. For this purpose, an ANOVA analysis of variance was applied to the training data and a score was assigned to the features. The best features were selected based on this score.

As a further step, upsampling was carried out. In the present data set, the classes are distributed unequally, which can lead to the model favoring the over-represented class and poorly classifying the under-represented class. In the binary classification 'difficulty' - 'no difficulty', the difference between the two classes is still negligible. However, if individual difficulties are to be classified against the rest, as is the case with multibinary classifiers, problems can arise when training the classifiers. To solve this problem, the RandomUpsampler from imbalanced-learn (Pedregosa et al., 2011) was used. Here, individual data points of the less strongly represented class were duplicated to equalize the distribution and thus improve the classification quality. Upsampling strengthens the model capacity and increases the generalization capability, which leads to a more balanced and reliable classification.

6.1.2.2 Classification

This chapter describes the procedure for training the algorithm. First, a parameter study is performed. Then, the classification approaches used to train the algorithm are presented.

Parameter study

To determine the optimal combination of parameters for classification, a parameter study was carried out. Four variables were analyzed: the window length, the overlap of the windows, the threshold for classifying a window as a difficulty and the duration of a difficulty. The variables are described in more detail in Figure 6.3.

Three factor levels were considered for each of these variables:

- window length: 5 s, 10 s, 30 s;
- overlap: 0 %, 25 %, 50 %;
- threshold for difficulty: 50 %, 75 %, 100 %;
- duration of a difficulty: full duration, first 30 s, first occurrence only.

This resulted in a total of 81 combinations, each of which was tested to identify the ideal combination of parameters. The aim was to maximize the classification quality and ensure the most balanced and reliable classification possible.

Binary classification

The established way to generate a two-class problem from a multi-class problem is to combine all difficulties into one class and adopt a binary classification approach (see Figure 6.4). Eleven different classifiers were trained to determine the most appropriate one for the problem.

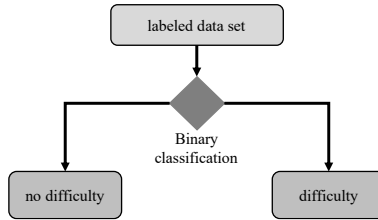


Figure 6.4: Functionality of a binary classifier (Zimmerer et al., 2025)

The selection consisted of the following classifiers:

- logistic regression,
- support vector machine,
- k-nearest-neighbors,
- Gaussian Naive Bayes,
- passive aggressive classifier,
- Decision Tree Classifier,
- Random Forest Classifier,
- ridge classifier,
- gradient boost classifier,
- perceptron classifier, and
- multi-layer perceptron classifier.

Multibinary classification

A different approach for solving multi-class problems in machine learning classification is multibinary classification. Here, each class is classified individually against all remaining classes, and the resulting binary classifiers are combined to form a multibinary classifier. This approach is called one-against-all (OAA). For each resulting binary classification problem, the best classifier and the best features are selected based on the training data.

Each classifier then classifies the test data set, and each data point that is identified as a difficulty by at least one of the classifiers is labeled accordingly. Thus, the score calculation is again applied to the two-class problem on which the binary classifier is based. As an additional result, the multibinary classifiers provide additional information about the specific difficulty for each segment labeled as such.

There are various approaches for combining the training data for the individual binary classifiers. The OAA algorithm presented by Jeatrakul and Wong (2012) is used here. Based on this OAA algorithm, a variation of this approach was developed, the cascading one-against-all (cOAA) algorithm.

In the OAA algorithm, the same training data set is always used and only the class assignment of each data point is changed. With the cOAA algorithm, on the other hand, a sequence is defined for the classification of the difficulties, starting with the difficulty that can be classified best. After a difficulty has been classified against the others, all data points belonging to this difficulty are removed from the training data set. As a result, the remaining data to be classified becomes smaller with each classification level, which should theoretically improve the classification quality of the subsequent binary classifiers.

A comparison of the two approaches is provided in Figure 6.5.

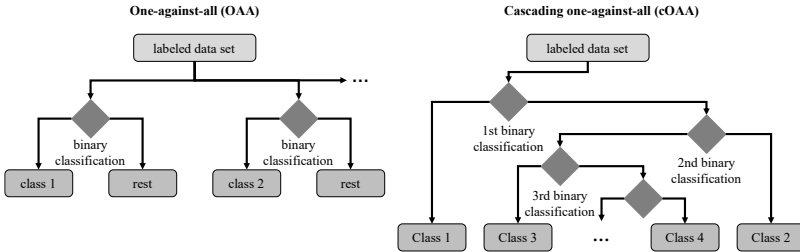


Figure 6.5: Graphical comparison of the multibinary approaches one-against-all (OAA) and cascading one-against-all (cOAA) (Zimmerer et al., 2025)

6.1.3 Scoring Metric

In this study, two metrics were used to evaluate the classifiers: the f1-score, which is a well-established metric, and the fn-score (false-negative-score), which is of particular interest for the use case. To determine the scores and standard deviations, a fivefold cross-validation was performed.

The f1-score, a harmonic mean of precision (how many of the predicted positives are correct) and recall (how many of the actual positives are correctly predicted), is widely used for unbalanced datasets and ranges from zero to one, with higher values indicating better performance of the classifier.

A high f1-score close to one means that the classifier is performing very well at balancing precision and recall. Figure 6.6 explains the calculation of the score. A more detailed explanation of the score can be found in Sasaki (2007).

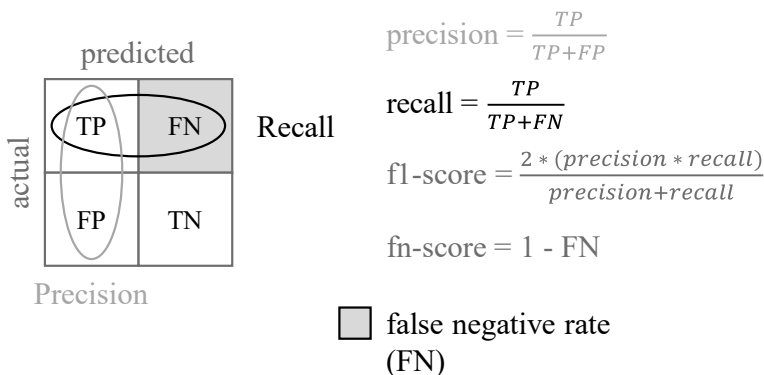


Figure 6.6: Calculation of the f1-score and fn-score (Zimmerer et al., 2025)

The fn-score is a measure of the false-negative rate. The false-negative rate indicates the percentage of data points that are true positives (segments with difficulties) but are incorrectly labeled as negatives (not difficulties). Since lower false-negative rates are better, the fn-score is calculated as the inverted fn-rate ($1 - \text{fn-rate}$) to make it more comparable with the f1-score. This means the fn-score also ranges from 0 to 1, with higher values indicating better performance. For example, an fn-score of 0.95 suggests the classifier correctly identifies 95 % of the segments with difficulties.

In this use case, the fn-score is important as the algorithm is used for the automated evaluation of eye tracking studies. It is therefore crucial that as few segments with difficulties as possible are not detected and included in the manual evaluation. However, incorrectly classified difficulties can be easily identified and corrected by the study director.

6.2 Results

In the following chapter, the results are presented. In chapter 6.2.1, the results of the parameter study are presented first, followed by the results on the binary (chapter 6.2.2) and multibinary (chapter 6.2.3) classification. Finally, the key findings are summarized in chapter 6.2.4.

6.2.1 Parameter Study

The parameter study was initially carried out with the binary classifier. This showed that all parameter combinations that included a shorter duration of the difficulty achieved significantly poorer results. It also turned out that the threshold for a difficulty only plays a minor role in the classification quality. Based on these findings, only the window length and the overlap of the windows were used as variable parameters for the more computationally intensive OAA and cOAA approaches to save computing capacity. A value of 75 % was set for the difficulty threshold, and there was no limit to the duration of a difficulty in all further calculations.

The three classification approaches are compared for each parameter combination of window length and overlap. The f1-score for the respective classifier is shown in light, the false-negative-score (fn-score) in dark grey. The lines at the top of the bars represent the standard deviation.

The binary classifier showed that the f1-score improved the shorter the window was (Figure 6.7, left side) and the less overlap the windows had (Figure 6.7, top). The fn-score remained largely constant across all parameter combinations, although it was also affected by a high standard deviation.

For the OAA classifier, the influence of window overlap was significantly greater than that of window length. The best results were obtained with large overlap of 50 % (Figure 6.7, bottom), and the classification quality decreased with decreasing overlap. The fn-score was higher than for the binary classifier, and again the standard deviation was relatively high. There was no clear pattern in the dependence of the fn-score on the parameters.

The cOAA achieved the best results in the f1-score. A clear trend was that the f1-scores increased with greater overlap (Figure 6.7, bottom), with significantly better results being achieved with a window length of 5 s (Figure 6.7, left) than with longer windows. The fn-score was also found to be significantly lower with shorter windows. Overall, the fn-scores were higher and the standard deviations of both scores were lower than for the other classifiers.

The results of the parameter study are presented in Figure 6.7.

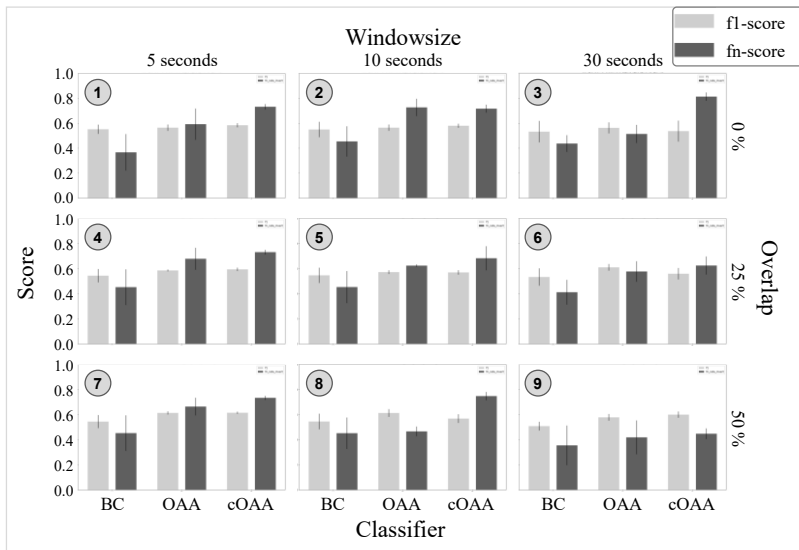


Figure 6.7: Results of the parameter study. For the binary classification the best results are achieved using a window length of 5 seconds and an overlap of 0 % (1, top left). The multibinary approaches in turn achieve the best results with a window length of 5 seconds and an overlap of 50 % (7, bottom left). (adapted from Zimmerer et al., 2025)

These results emphasize the importance of the choice of window length and overlap for the performance of the classifiers. Especially the cascading OAA approach showed promising results in terms of accuracy and reliability of the classification.

After selecting the most suitable parameters for each case, the individual classification approaches are presented in detail below.

6.2.2 Binary Classification

For the BC, all types of difficulties were summarized in the category 'difficulty'. This results in only the two classes 'difficulty' and 'no difficulty'. As can be seen from the parameter study, the best classification results are obtained with a window size of 5 s and an overlap of 0 %. The feature selection using ANOVA showed that the features mean pupil diameter and median pupil diameter have the highest scores and are therefore used for classification.

Table 6.1 shows the results of the cross-validation that was carried out to determine the scores and standard deviations.

The f1-score lies between 0.419 and 0.552, with a mean standard deviation of 0.05. The Support Vector Machine classifier achieved the best results with an f1-score of 0.552 (SD = 0.037), followed by the Ridge classifier, which achieved comparable results with a score of 0.544 (SD = 0.062).

For the fn-score, there were significant differences between the classifiers. The fn-score lies between 0.176 and 0.704, with an average standard deviation of 0.15. The perceptron classifier achieved the best value here with 0.7038, but also showed a very high standard deviation of 0.26.

Table 6.1: Comparison of the tested classifiers and their achieved scores. The highest values for each scoring metric are highlighted. (Zimmerer et al., 2025)

Classifier	f1-score	f1-score-std	fn-score	fn-score-std
Support Vector Machine	0.552	0.0371	0.3667	0.1467
Ridge Classifier	0.544	0.0624	0.5554	0.1565
Gaussian Naive Bayes	0.5437	0.0514	0.4504	0.1482
Logistic Regression	0.5434	0.0618	0.5562	0.158
Gradient Boosting Classifier	0.5424	0.0296	0.429	0.0748
MLP Classifier	0.5398	0.0291	0.4177	0.141
k Nearest Neighbors	0.5179	0.014	0.4888	0.0743
Random Forest Classifier	0.5036	0.0242	0.4162	0.0844
Decision Tree Classifier	0.5026	0.0227	0.439	0.0867
Perceptron Classifier	0.4292	0.1011	0.7038	0.26
Passive Aggressive Classifier	0.4185	0.0951	0.1755	0.3511

6.2.3 Multibinary Classification

For multibinary classification, a parameter combination of a window size of 5 s and an overlap of 50 % proved to be optimal. In contrast to BC, an individual selection of features and classifiers was made for each class of difficulty.

First, the best classifiers were determined individually for each difficulty. Figure 6.8 illustrates the f1-scores of the best classifiers for each specific difficulty in the multibinary classification. The different difficulties are listed on the x-axis: ‘assembly’, ‘Concept selection’, ‘embodiment design’, ‘function unclear’, ‘general overload’, ‘manufacturing process’, ‘no concept idea’, and ‘visualization’. The y-axis represents the f1-score, which serves as a measure of the classification quality.

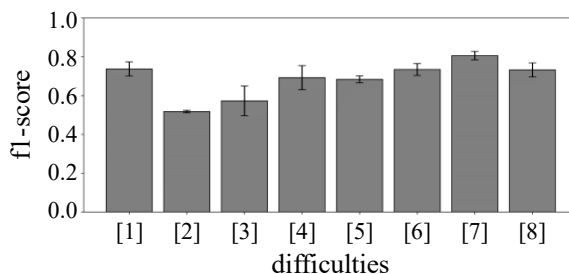


Figure 6.8: Highest possible classification result for each class of difficulty using the most appropriate classifiers ([1]: visualization; [2]: no concept idea; [3]: embodiment design; [4]: general overload; [5]: function unclear; [6]: manufacturing process; [7]: concept selection; [8]: assembly) (Zimmerer et al., 2025)

The bars in the figure represent the average f1-scores of the best classifiers for each difficulty, while the error bars at the top of the bars indicate the standard deviation of the f1-scores. This allows an assessment of the variability and reliability of the classification results.

The overall classifier achieves a combined f1-score of 0.6137 and an fn-score of 0.6646. The highest classification score for a single difficulty was achieved for the difficulty ‘concept selection’, with an f1-score of 0.81. The difficulties ‘visualization’, ‘manufacturing process’, and ‘assembly’ are in a medium range with f1-scores between 0.73 and 0.74, while ‘general overload’ and ‘function unclear’ achieve an f1-score of slightly below 0.7. The difficulties ‘embodiment design’ and ‘no concept idea’ have the lowest scores of 0.57 and 0.52 respectively and thus the lowest classification quality of all difficulties.

A low standard deviation, as for the difficulties '*no concept idea*', '*embodiment design*', '*function unclear*', and '*manufacturing process*', indicates consistent classification performance. A medium standard deviation can be observed for the difficulty '*general overload*', which indicates a somewhat greater variability in the results. The highest standard deviations can be seen for the difficulties '*visualization*', '*concept selection*', and '*assembly*', which indicates a higher variability of the classification results despite high f1-scores.

For each difficulty, the classifiers and features with the highest accuracy were used. Table 6.2 lists the features and classifiers selected in each case. The Decision Tree Classifier and the Random Forest Classifier were each used three times. The Gradient Boosting Classifier and the K-Nearest-neighbors Classifier were each used once. There are only a few overlaps among the features used: 'Pupil diameter (mean)', 'Fixation point (1st derivation, median)', and 'Gaze direction (2nd derivation, max)' were each used twice. Interestingly, there are no identical combinations of features for the different difficulties.

Table 6.2: Used Feature OAA (Zimmerer et al., 2025)

difficulty	classifier	used features
assembly	Random Forest Classifier	Gaze direction (1 st derivation, max) Pupil position (1 st derivation, max)
Concept selection	Decision Tree Classifier	Gaze direction (2 nd derivation, max), Fixation point (2 nd derivation, median)
embodiment design	Decision Tree Classifier	Pupil diameter (mean), Pupil diameter (2 nd derivation, median)
function unclear	K-Nearest Neighbors Classifier	Blink rate, Fixation point (2 nd derivation)
general overload	Random Forest Classifier	Fixation point (1 st derivation, max), Pupil diameter (mean)
manufacturing process	Random Forest Classifier	Gaze point (1 st derivation, max), Gaze direction (2 nd derivation, max)
no concept idea	Gradient Boosting Classifier	Pupil diameter (mean), Pupil diameter (median)
visualization	Decision Tree Classifier	Fixation point (2 nd derivation, max), Fixation point (1 st derivation, median)

Looking at the features used, it can be seen that pupil-based features were used six times. Fixation-based features were used five times, while gazing direction was used as a feature four times. In addition, blinking was used once as a feature. It is noteworthy that saccade-based features were not used.

To further increase accuracy, a cascading OAA approach was also applied. The order of the individual classifications was determined in such a way that the best possible overall result is achieved.

Table 6.3 presents the changes in f1-scores between the OAA and the cOAA approaches for various difficulties. As shown in the table, the cascading approach led to an improvement in the f1-scores for most of the difficulties.

Table 6.3: Change in the individual classification results due to the cascading one-against-all approach (Zimmerer et al., 2025)

difficulty	f1-Score OAA	f1-Score cOAA
concept selection	0.8059	0.8059
visualization	0.7373	0.7436
general overload	0.6925	0.7588
assembly	0.7325	0.7642
manufacturing process	0.7345	0.7366
function unclear	0.6837	0.6985
no concept idea	0.5176	0.5475
embodiment design	0.5728	0.5586

The '*concept selection*' difficulty maintained its f1-score of 0.8059, indicating no change. This lack of change can be attributed to the fact that this difficulty is classified first, with no data points yet removed, mirroring the situation in the OAA approach.

Improvements can be observed in all remaining categories except the '*embodiment design*' class. The greatest improvement can be observed in the '*general overload*' class. Here the score rises by 0.06 to 0.76. There was also a significant improvement of 0.03 in the '*assembly*' class. However, there was a slight deterioration observed in the '*embodiment design*' class, where the f1-score decreased from 0.5728 to 0.5586.

The final result of the classification is an overall f1-score of 0.615 (SD = 0.001). The false-negative score is even higher with 0.734 (SD = 0.016)

There were some changes in some of the classifiers due to the new approach:

In the '*visualization*', '*embodiment design*', and '*concept selection*' classes, an improvement in prediction was achieved by changing the classifier used. All three classes used a Random Forest Classifier in the one-against-all approach.

While '*visualization*' and '*concept selection*' now achieve better results with a Decision Tree Classifier, the classifier of the '*embodiment design*' class switches to a support vector machine. In addition to the classifiers, there were also changes in the features used in two difficulty classes. In the '*embodiment design*' class, 'Pupil diameter (1st derivation, median)' was replaced by 'Pupil diameter (median)', while in the '*general overload*' class 'Pupil diameter (mean)' was replaced by 'Fixation point (2nd derivation, max)'.

6.2.4 Summary of Results

The Binary Classifier achieved an f1-score of 0.5520 (SD = 0.0367), and a false-negative rate of 0.3667 (SD = 0.1467). The one-against-all (OAA) approach improved the performance, yielding an f1-score of 0.6137 (SD = 0.0126), and a clearly better false-negative score of 0.6646 (SD = 0.0703). The cOAA approach further enhanced the results, achieving the highest f1-score of 0.6150 (SD = 0.0089), and the best false-negative rate of 0.7342 (SD = 0.0162). A summary is shown in Table 6.4. Even if the cOAA does not differ considerably from the OAA regarding the absolute f1 score, the clear reduction of the standard deviation and the increased false-negative score should be noted.

Table 6.4: Final result of the three classification approaches measured by f1-score and fn-score as well as the associated standard deviations. (Zimmerer et al., 2025)

classifier	f1-score	f1-score-std	fn-score	fn-score-std
Binary Classifier	0.5520	0.0367	0.3667	0.1467
OAA	0.6137	0.0126	0.6646	0.0703
cOAA	0.6150	0.0089	0.7342	0.0162

6.3 Discussion

In this chapter, the research question 2 “*How can difficulties in concept synthesis be detected automatically based on eye tracking data?*” could be answered. The cOAA algorithm answered this question by demonstrating a high accuracy of 62 % and a false-negative rate of 27 %, utilizing eye tracking data to identify critical situations during design tasks. In the following, the implications are discussed, and the limitations and possible next steps are analyzed.

6.3.1 Detection Difficulties via Eye Tracking

The integration of eye tracking data with traditional research methods represents a significant step forward in the field of design research. Compared to the accuracies reported in the state of research (see chapter 2.3.4), the approach of combining eye tracking with retrospective think aloud shows notable improvements, especially for complex and extended tasks. Existing research often focuses on short, self-contained tasks like arithmetic problems (Borys et al., 2017; Nourbakhsh et al., 2013), surveyed by questionnaires achieving high accuracy but limited to these narrow scopes. In contrast, the RTA approach employed in this study is designed for longer time intervals and more comprehensive tasks, making them suitable for capturing data over entire design sessions.

The continuous collection of difficulty data through RTA allowed relatively high prediction accuracies even for extended and complex tasks such as concept synthesis. A major advantage of RTA is that it does not impose any additional cognitive load on participants during the data collection process. This non-intrusive research method ensures that participants’ natural workflow is not disrupted, resulting in more accurate and authentic data. In addition, RTA measures subjective load as actually perceived by the participant, eliminating the need for interpretation by an external evaluator.

In addition, combining eye tracking data with established qualitative survey methods such as interviews and protocol analysis provides a more nuanced understanding of the design process (Ruckpaul et al., 2017). The results showed that different features and combinations of features were effective in detecting various difficulties. In particular, the integration of context-specific varying eye tracking metrics helped to achieve higher classification accuracy and a lower false-negative rate, i.e. less occurring difficulties not identified by the algorithm (see chapter 6.1.3), compared to the state of research, demonstrating its effectiveness in providing accurate and reliable data on design processes.

Especially the false-negative rate achieved by the cOAA approach supports the suitability of eye tracking data for identifying difficulties in design tasks. Lower false-negative rates have the great advantage that the algorithm does not leave any relevant situations undetected and therefore provides more reliable data for subsequent analysis. This is of crucial importance in design research, especially with regard to the evaluation effort, as the precise identification of design challenges is essential for this. By reducing the time and resources needed for analysis, the proposed approach also enables studies with larger participant groups, making it possible to move beyond small-scale investigations.

6.3.2 Tailored Feature Selection

These findings have important implications for research into the measurement of cognitive load using eye tracking. In particular, the fact that different types of difficulties can be detected with different eye tracking features is a highly valuable finding.

Fixations are the relevant features for identification in several difficulty categories, especially in '*visualization*', '*function unclear*', and '*concept selection*'. This implies that extended fixations in these cases are associated with increased cognitive load and can therefore serve as an indicator of difficulties.

In contrast, the results show a close relationship between pupil diameter and difficulties such as '*embodiment design*' and '*no concept idea*'. As is already known from the state of research, a dilated pupil can indicate increased cognitive load (F. Chen et al., 2016; Marquart et al., 2015) and thus phases in the design process in which complexity is high or information processing is intensified. Cognitive processes such as thinking and imagination therefore appear to play a greater role in the difficulties presented here, in which pupil diameter plays a role.

Difficulties relating to '*general overload*' are between these two cases. One fixation and one pupil-based feature are relevant here. This may primarily be since many different types of difficulties are combined here and therefore no clear correlations to the relevant features could be identified during clustering.

For classes of difficulties relating to the '*manufacturing process*' or '*assembly*', features on the gaze path are dominant. In these types of difficulties, sequences, and interrelationships between different design features or elements of the structure often play an important role. This fact could explain the strong dependence on gaze behavior in these cases.

From this, it can be deduced that the present results emphasize the need for a differentiated consideration of eye tracking metrics depending on the context, in this case the type of difficulty. For the development of algorithms for recording cognitive load using eye tracking, it is therefore crucial when compiling training data not to focus purely on a one-dimensional measurement of cognitive load, but also to always consider the specific situation at hand.

Nevertheless, further research is needed to validate the accuracy and reliability of the eye parameters in different design contexts and to further improve the integration of eye tracking into the design research process. This study provides an important contribution to the study of cognitive processes in design and lays the foundation for future work on the use of eye tracking to capture difficulties in the design process.

6.3.3 Algorithm for Automated Evaluation

The classification results of the three approaches – Binary Classifier, one-against-all (OAA), and cOAA – demonstrate significant differences in their performance. The Binary Classifier achieves an f1-score of 0.5520 with a standard deviation of 0.0367. In contrast, the OAA approach shows an improvement with an f1-score of 0.6137 and a lower standard deviation of 0.0126. The cOAA approach achieves the best results with an f1-score of 0.6150 and the lowest standard deviation of 0.0089, indicating a more consistent performance.

The same picture emerges when looking at the false-negative rate. The binary classifier is far behind with an fn-score of 0.367, which corresponds to a false-negative rate of 63.3 %. In a comparison of the multibinary classifiers, the OAA algorithm with an fn-score of 0.665 is also noticeably less accurate than the cOAA with an fn-score of 0.734. The standard deviation of 0.016 for the cOAA is also significantly lower than for the OAA ($SD = 0.07$). This means that the cOAA does not recognize around a quarter of the difficulties mentioned by the participants.

In conclusion, the cOAA approach, with its cascading strategy, offers a considerable improvement over both the Binary Classifier and the OAA approach, making it a more effective method for accurately and consistently classifying difficulties in design tasks. This advancement highlights the importance of using specialized classifiers and features tailored to specific types of difficulties, thereby enhancing the overall classification performance and reliability in design research studies. This facilitates larger-scale studies and statistically more reliable findings, making eye tracking a powerful research method in design research for capturing detailed and objective data on cognitive processes. With this approach, the efficiency of data

collection and evaluation will be increased, and design processes thus better analyzed and interpreted.

The exact extent of the advantage of the approach developed has yet to be fully investigated. In particular, further work is needed to increase the accuracy of the classification models. While the accuracy of 73 % of detected difficulties (based on the fn-score) achieved in this study is a very good step in the right direction and shows the potential and feasibility of this project, this value needs to be further increased to improve the practical applications and the reliability of the results.

6.3.4 Implications for the Field of Design Research

The central problem, which was introduced in chapter 2.1, is the difficulty of achieving statistically significant results, as design studies are often only carried out with a small number of participants due to the enormous amount of evaluation required (Cash et al., 2022; Dinar et al., 2015).

The present work shows that the developed machine learning algorithms and the cascading classification approach have the potential to significantly reduce the required evaluation effort. Based on the example that one hour of study material requires about 25 additional hours of work in data processing and (Ahmed et al., 2003), even small improvements in the evaluation effort can provide benefits.

This may enable more extensive studies in future and thus improve the statistical robustness of the results. Thus, the use of these research methods could significantly increase the overall validity and reliability of design studies. A further benefit of this approach is that the automated evaluation of eye tracking data can reduce the influence of subjective influences by the evaluators of studies. The objectivity of the evaluation is thus considerably increased by the use of eye tracking data (Ruckpaul, Fürstenhöfer, & Matthiesen, 2015).

6.4 Interim Conclusion

This chapter answered the second research question:

RQ2: “How can difficulties in concept synthesis be detected automatically based on eye tracking data?”

The aim was to build on the difficulties identified in RQ1 to develop an algorithm that can automatically detect these difficulties from eye tracking data.

To answer the research question, various classification approaches based on machine learning methods were evaluated. The focus was on the extent to which the different types of difficulties from RQ1 can be measured in the eye tracking data and which methods enable robust and valid detection. Through the systematic evaluation of the approaches, both the strengths and limitations of each approach were identified. Overall, the results show that the automated detection of difficulties in concept synthesis using eye tracking is possible and that machine learning approaches represent a promising way to complement established research methods in empirical design research. Analysis of the results led to three key findings:

- Different types of difficulties can be identified using different gaze parameters.
- Combining eye tracking, machine learning, and qualitative research methods has great potential for the evaluation of empirical studies and can significantly accelerate evaluation processes and make them more reliable.
- Automated data acquisition, collection, and evaluation offer a scalable alternative to classic, resource-intensive design research methods as the algorithm can automatically identify the most relevant situations.

The study showed that not all difficulties can be identified using the same characteristics of the eye tracking data. While longer fixations are particularly relevant for difficulties such as '*visualization*', '*function unclear*', and '*concept selection*', pupil diameter indicates cognitively demanding processes such as '*embodiment design*' or '*no concept idea*'. In contrast, sequences with many saccades in the gaze path are decisive for difficulties relating to the '*manufacturing process*' or '*assembly*'. This differentiation illustrates that the type of difficulty is closely linked to different cognitive processes and therefore requires a targeted selection of eye tracking parameters.

The combination of eye tracking with established qualitative research methods such as retrospective think aloud (RTA) proved particularly valuable. While eye tracking provides objective, continuous data on difficulties, RTA enables context-sensitive validation of this data from the subjective perspective of the participants. The combination of both approaches leads to higher classification accuracy, reduces misclassifications, and allows even complex and lengthy design tasks to be evaluated reliably. This creates methodological added value that goes beyond the possibilities of purely quantitative or purely qualitative methods.

The ability to automatically evaluate eye tracking data addresses a key bottleneck in empirical design research: the high manual effort required for protocol analysis,

interviews, or video evaluation. The classification approaches developed open up the possibility of conducting studies with larger numbers of participants without the evaluation exceeding the available resources. This can significantly increase the robustness and generalizability of research results in design research.

In summary, the results of this study make an important contribution to the development of new research methods in design research. Although the algorithms tested do not yet represent a definitive solution, they show that automated processes have the potential to fundamentally change the methodological standard in the field. The ability to continuously measure cognitive load with a high degree of objectivity has given rise to an innovative approach to investigating design processes that combines both efficiency and accuracy. The next step will be to build on this foundation and investigate how the developed algorithm can be integrated into broader methodological approaches to enable the analysis of design processes on a larger scale.

7 Summary and Outlook

The following chapter summarizes the key findings of this thesis. It then provides an outlook on possible follow-up research topics.

7.1 Summary

This thesis addresses a central challenge in design research: established research methods such as interviews, protocol analyses, or the evaluation of eye tracking videos require a considerable amount of time and effort. This high level of effort results in a number of limitations: the number of participants that can be included in studies is limited, making it difficult to obtain statistically reliable results. In addition, the evaluation depends heavily on the interpretation of the researchers, which limits objectivity. As a result, studies on the design process with sample sizes large enough to yield statistically significant and reliable results are currently almost impossible to carry out.

In order to overcome these limitations, there is an urgent need in design research for new research methods and tools that reduce resource-intensive data preparation and analysis while delivering reliable, objective results. These methods must be able to

- consider the need for support individually from the engineering designer's point of view,
- enable time-continuous recording during the process,
- to be able to automatically process a large amount of data in a scalable manner, and
- deliver reliable results.

Considering this, the aim of this thesis is to develop an algorithm that enables an automated evaluation of support needs in empirical participant studies based on eye tracking data. To this end, self-reported difficulties by participants are used as ground truth. These difficulties are interpreted as indicators of support needs.

The overarching question is derived from this objective:

How can eye tracking data be used to objectively and automatically measure support needs in the form of difficulties in concept synthesis as part of the design process?

To answer the overarching research question, two research questions were examined:

RQ1: *"What difficulties do engineering designers encounter in concept synthesis that can be used as an indicator of support needs?"*

RQ2: *"How can difficulties in concept synthesis be detected automatically based on eye tracking data?"*

To answer the research questions, a study design was developed in which participants had to solve conceptual design tasks in the field of sheet metal design. The self-reported difficulties encountered in solving the tasks were used as a basis for answering the research questions.

To examine the **first research question**, an exploratory approach was applied using retrospective think-aloud (RTA) sessions conducted after task completion to capture perceived difficulties. In this setting, participants reviewed video recordings of their own approach and commented on moments during which they had experienced difficulties. This procedure allowed difficulties to be captured directly from the participants' perspective without interrupting them while working on the tasks.

The reported difficulties were systematically analyzed and clustered in order to identify recurring patterns across participants. This process resulted in eight distinct categories of difficulties: *'assembly'*, *'concept selection'*, *'embodiment design'*, *'function unclear'*, *'general overload'*, *'manufacturing process'*, *'no concept idea'*, and *'visualization'*. The clustering provides a structured way to capture the heterogeneity of difficulties while enabling systematic comparison across different participants and tasks. These identified difficulty categories are directly relevant for the conceptual design phase, as they highlight the specific cognitive and procedural challenges designers face when generating and developing concepts. Understanding these difficulties provides a foundation for targeted methodological support during concept synthesis.

The categorization offers several methodological benefits. First, it systematizes a wide range of individual, subjective reports into a consistent set of categories that can be used for further analyses. Second, it establishes a reproducible basis for detecting support needs in concept synthesis. Furthermore, the clusters serve as a

necessary ground truth for subsequent algorithmic development, forming the link between qualitative insights from RTA and quantitative data from eye tracking. Finally, this approach also provides the opportunity to extend the analysis to other phases of the design process. By transferring the categorization framework to these phases, difficulties specific to each phase can be identified in an analogous manner, allowing the subsequent analytical steps to be reproduced accordingly.

The **second research question** builds directly on the results of RQ1. The aim was to develop and evaluate an algorithm that can identify the categorized difficulties by using eye tracking data. For this purpose, various machine learning approaches were systematically tested and validated.

By combining eye tracking data with the statements on self-assessed difficulties by the participants, three different machine learning approaches were applied to the data set: a binary and two multibinary approaches (one against all (OAA) and cascading one against all (cOAA)). The difficulties identified using eye tracking data were then compared with the participants' self-reported difficulties and the accuracy of the prediction was measured. With cOAA, an accuracy of up to 62 % was achieved. The false negative rate of only 26.6 % is also particularly relevant, meaning that few difficulties cannot be detected by the cOAA algorithm.

Importantly, the analyses revealed that different categories of difficulties are associated with different eye tracking parameters. For example, extended fixations are most relevant for challenges in '*visualization*', '*function unclear*', and '*concept selection*'. Meanwhile, pupil diameter proved significant for cognitively demanding difficulties such as '*embodiment design*' or '*no concept idea*'. In contrast, saccadic patterns in gaze paths were decisive for difficulties related to '*manufacturing process*' or '*assembly*'. The differentiated detection of difficulties in conceptual design tasks demonstrates how automated research methods can support designers during this phase, enabling identification of individual support needs.

These findings highlight that a differentiated feature selection is necessary to capture the various cognitive processes underlying different types of difficulties. At the same time, they demonstrate the potential of combining eye tracking data and machine learning to substantially reduce manual evaluation efforts. The automated analysis enables studies with larger sample sizes and more statistically reliable results, thereby representing a significant step toward more efficient and insightful research methods in design cognition.

Taken together, the results of this thesis provide a coherent answer to the overarching research question. By identifying and systematizing difficulties in concept synthesis (RQ1), a structured foundation was created that allows the

support needs of engineering designers to be captured in a reproducible and comparable manner. Building on this, the automated detection of these difficulties using eye tracking data and machine learning through the developed cOAA algorithm (RQ2) demonstrates that support needs can be measured objectively and at scale. In this way, the thesis shows how eye tracking data can be used to automate the assessment of difficulties in design processes, thereby reducing evaluation effort while at the same time increasing objectivity and statistical robustness. These contributions mark an important step toward the development of new research methods in design research that enable individualized, time-continuous, and scalable assessment of support needs.

7.2 Outlook

The results of this thesis open up several perspectives for future research in design cognition and the development of new research methods. A first step concerns the further development of the cOAA algorithm itself. While this thesis focused on difficulties in concept synthesis, the procedure can be extended to additional activities in the design process. This would allow a more comprehensive coverage of cognitive challenges across different stages of design and provide a broader empirical basis for analyzing support needs. Moreover, this creates a foundation for the development and improvement of design methods themselves, as their application can now be systematically examined and empirically evaluated under real design conditions.

A second step relates to the transfer of the approach to professional practice. The present study relied on student participants, who offer the advantage of homogeneity and availability. However, in order to examine the external validity of the cOAA algorithm, it is essential to replicate the study with professional engineering designers and to test the cOAA algorithm under realistic working conditions. This will provide insights into the robustness and applicability of the approach in everyday design contexts.

In addition, the integration of further physiological signals represents a promising opportunity for further improvement the automated detection of difficulties. Signals such as electromyography (EMG), electrocardiography (ECG), or electrodermal activity (EDA) capture additional dimensions of cognitive and emotional load that cannot be derived from gaze data alone. Combining eye tracking with these physiological signals may increase the accuracy and reliability of classification and provide a more holistic view of cognitive processes in design.

At the same time, new technological developments create opportunities for scaling and broadening the application of such methods. Webcam-based eye tracking lowers technical barriers and enables data collection with larger and more diverse participant groups. This creates the possibility to conduct more extensive studies while maintaining objectivity and reducing resource requirements. Moreover, it opens up new possibilities for studying designers directly in their real working environments, providing insights under authentic conditions. This not only enables large-scale and continuous data collection but also paves the way for data-driven evaluation and further development of design methods in practice.

Finally, immersive digital environments such as virtual learning environments or augmented workspaces offer a vision for embedding these methods directly into design practice. These environments not only support the unobtrusive recording of eye tracking data and other physiological signals but also allow complex and realistic design tasks to be studied under controlled conditions. In the long term, such technologies may enable automated, real-time assessment of support needs during the design process and thereby contribute to the development of adaptive design environments.

References

- Ahmed, S. (2007). Empirical research in engineering practice. *Journal of Engineering Design*, 6(3), 359–380.
<https://doi.org/10.1504/JDR.2007.016389>
- Ahmed, S., Wallace, K. M., & Blessing, L. T. (2003). Understanding the differences between how novice and experienced designers approach design tasks. *Research in Engineering Design*, 14(1), 1–11.
<https://doi.org/10.1007/s00163-002-0023-z>
- Albers, A. (2010). Five Hypotheses about Engineering Processes and their Consequences. In I. Horvath, F. Mandorli, & Z. Rusak (Eds.), *Proceedings of the TMCE 2010*, 12.-16. April, Ancona, Italy (pp. 343–356).
- Alpaydın, E. (2020). *Introduction to machine learning* (Fourth edition). *Adaptive computation and machine learning*. The MIT Press.
- Andreasen, M. M. (2011). 45 Years with design methodology. *Journal of Engineering Design*, 22(5), 293–332.
<https://doi.org/10.1080/09544828.2010.538040>
- Andreasen, M. M., Hansen, C. T., & Cash, P. J. (2015). *Conceptual Design: Interpretations, Mindset and Models*. Springer International Publishing.
<https://doi.org/10.1007/978-3-319-19839-2>
- Badke-Schaub, P., Daalhuizen, J., & Roozenburg, N. (2011). Towards a Designer-Centered Methodology: Descriptive Considerations and Prescriptive Reflections. In H. Birkhofer (Ed.), *The Future of Design Methodology* (pp. 181–197). Springer London.
- Balters, S., & Steinert, M. (2017). Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices. *Journal of Intelligent Manufacturing*, 28(7), 1585–1607. <https://doi.org/10.1007/s10845-015-1145-2>

- Bender, B [Bernd] (2003). Task design and task analysis for empirical studies into design activity. *Journal of Engineering Design*, 14(4), 399–408.
<https://doi.org/10.1080/09544820310001606894>
- Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011). Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behaviour*, 14(3), 199–208.
<https://doi.org/10.1016/j.trf.2010.12.001>
- Bi, Y., & Reid, T. (2017). Evaluating Students' Understanding of Statics Concepts Using Eye Gaze Data. *International Journal of Engineering Education*, 33(1(A)), 225–235. https://www.ijee.ie/latestissues/Vol33-1A/21_ijee3392ns.pdf
- Birkhofer, H., Klobdanz, H., Sauer, T., & Berger, B. (2002). Why methods don't work and how to get them to Work. In R. Rohatynski (Ed.), *DS 29: Proceedings of the 3rd International Seminar and Workshop on Engineering Design in Integrated Product Development (EDIProD 2002)*, 10.-12.10.2002, Zielona Góra, Poland (pp. 29–36).
- Bishop, C. M. (2006). *Pattern recognition and machine learning. Computer science*. Springer.
- Björklund, T. A. (2013). Initial mental representations of design problems: Differences between experts and novices. *Design Studies*, 34(2), 135–160. <https://doi.org/10.1016/j.destud.2012.08.005>
- Blessing, L. T., & Chakrabarti, A. (2009). *DRM, a Design Research Methodology*. Springer. <https://doi.org/10.1007/978-1-84882-587-1>
- Borys, M., Plechawska-Wójcik, M., Wawrzyk, M., & Wesołowska, K. (2017). Classifying Cognitive Workload Using Eye Activity and EEG Features in Arithmetic Tasks. In R. Damasevicius & V. Mikasys (Eds.), *Communications in Computer and Information Science: Vol. 756, Information and software technologies: 23rd International Conference, ICIST 2017, Druskininkai, Lithuania, October 12–14, 2017 : Proceedings* (pp. 90–105). Springer. https://doi.org/10.1007/978-3-319-67642-5_8

- Bruggen, A. (2015). An empirical investigation of the relationship between workload and performance. *Management Decision*, 53(10), 2377–2389. <https://doi.org/10.1108/MD-02-2015-0063>
- Buettner, R. (2013). Cognitive Workload of Humans Using Artificial Intelligence Systems: Towards Objective Measurement Applying Eye-Tracking Technology. In M. Thimm & I. J. Timm (Eds.), *SpringerLink Bücher: Vol. 8077, Ki 2013: Advances in Artificial Intelligence: 36th Annual German Conference on AI, Koblenz, Germany, September 16-20, 2013. Proceedings* (pp. 37–48). Springer. https://doi.org/10.1007/978-3-642-40942-4_4
- Cash, P. J., Elias, E., Dekoninck, E., & Culley, S. (2012). Methodological insights from a rigorous small scale design experiment. *Design Studies*, 33(2), 208–235. <https://doi.org/10.1016/j.destud.2011.07.008>
- Cash, P. J., Isaksson, O., Maier, A., & Summers, J. D. (2022). Sampling in design research: Eight key considerations. *Design Studies*, 78, 101077. <https://doi.org/10.1016/j.destud.2021.101077>
- Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, 74, 221–232. <https://doi.org/10.1016/j.apergo.2018.08.028>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust Multimodal Cognitive Load Measurement*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-31700-7>
- Chen, J., Zhang, Q., Cheng, L., Gao, X., & Ding, L. (2019). A Cognitive Load Assessment Method Considering Individual Differences in Eye Movement Data. In L. Xie, Z. Ding, & Y. Zhang (Eds.), *2019 IEEE 15th International Conference on Control and Automation (ICCA)*, 16.-19.07.2019, Edinburgh, United Kingdom (pp. 295–300). <https://doi.org/10.1109/ICCA.2019.8899595>

- Chen, S., Epps, J., Ruiz, N., & Chen, F. (2011). Eye activity as a measure of human mental effort in HCI. In P. Pu, M. Pazzani, E. André, & D. Riecken (Eds.), *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11)*, 13.-16.02.2011, Palo Alto, CA, USA (pp. 315–318). <https://doi.org/10.1145/1943403.1943454>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037//0033-2909.112.1.155>
- Cramer-Petersen, C. L., Christensen, B. T., & Ahmed-Kristensen, S. (2019). Empirically analysing design reasoning patterns: Abductive-deductive reasoning patterns dominate design idea generation. *Design Studies*, 60, 39–70. <https://doi.org/10.1016/j.destud.2018.10.001>
- Cross, N. (2004). Expertise in design: an overview. *Design Studies*, 25, 427–441. <https://doi.org/10.1016/j.destud.2004.06.002>
- Cross, N. (2005). *Engineering design methods: Strategies for product design* (3rd ed.). Wiley.
- Daalhuizen, J. (2014). *Method Usage in Design: How methods function as mental tools for designers* [Doctoral dissertation, Delft University of Technology]. <https://repository.tudelft.nl/record/uuid:4ac01165-c6b5-4057-a2fe-3418907f251e>
- Daalhuizen, J., Badke-Schaub, P., & Batill, S. M. (2009). Dealing with uncertainty in design practice: issues for designer-centered methodology.
- de Winter, J. C. F. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: a measurement perspective. *Cognition, Technology & Work*, 16(3), 289–297. <https://doi.org/10.1007/s10111-014-0275-1>
- Dias, R. D., Ngo-Howard, M. C., Boskovski, M. T., Zenati, M. A., & Yule, S. J. (2018). Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. *The British Journal of Surgery*, 105(5), 491–501. <https://doi.org/10.1002/bjs.10795>
- Dieter, G. E., & Schmidt, L. C. (2013). *Engineering design* (5th ed.). McGraw-Hill.

- Dinar, M., Shah, J. J., Cagan, J., Leifer, L., Linsey, J., Smith, S., & Hernandez, N. V. (2015). Empirical Studies of Designer Thinking: Past, Present, and Future. *Journal of Mechanical Design*, 137(2), 247. <https://doi.org/10.1115/1.4029025>
- Dinges, D. F., & Grace, R. (1998). PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance. *US Department of Transportation, Federal Highway Administration, Publication Number FHWA-MCRT-98-006*. Advance online publication. <https://doi.org/10.21949/1502740>
- Doellken, M., Zimmerer, C., & Matthiesen, S. (2020). Challenges Faced by Design Engineers when Considering Manufacturing in Design – An Interview Study. In D. Marjanović, M. Štorga, S. Škec, & T. Martinec (Eds.), *DS 102: Proceedings of the DESIGN 2020 16th International Design Conference*, 26.-29.10.2020, Online (pp. 837–846). <https://doi.org/10.1017/dsd.2020.302>
- Dorst, K. (2008). Design research: a revolution-waiting-to-happen. *Design Studies*, 29(1), 4–11. <https://doi.org/10.1016/j.destud.2007.12.001>
- Dorst, K., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. *Design Studies*, 22(5), 425–437. [https://doi.org/10.1016/S0142-694X\(01\)00009-6](https://doi.org/10.1016/S0142-694X(01)00009-6)
- Ehrlenspiel, K., & Meerkamm, H. (2013). *Integrierte Produktentwicklung: Denkabläufe, Methodeneinsatz, Zusammenarbeit* (5th ed.). Carl Hanser Verlag. <https://doi.org/10.3139/9783446436275>
- Ehrlenspiel, K., & Meerkamm, H. (2017). *Integrierte Produktentwicklung: Denkabläufe, Methodeneinsatz, Zusammenarbeit* (6th ed.). Carl Hanser Verlag. <https://doi.org/10.3139/9783446449084>
- Ericsson, K. A., & Lehmann, A. C. (1996). Expert and Exceptional Performance: Evidence of Maximal Adaptation to Task Constraints. *Annual Review of Psychology*, 47, 273–305. <https://doi.org/10.1146/annurev.psych.47.1.273>
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal reports as data* (3rd ed.). The MIT Press.

- Escudero-Mancebo, D., Fernández-Villalobos, N., Martín-Llorente, Ó., & Martínez-Monés, A. (2023). Research methods in engineering design: a synthesis of recent studies using a systematic literature review. *Research in Engineering Design*, 34(2), 221–256. <https://doi.org/10.1007/s00163-022-00406-y>
- Eysenck, M. W., & Keane, M. T. (2020). *Cognitive psychology: A student's handbook* (Eighth edition). *A Psychology Press book*. Psychology Press Taylor & Francis Group.
- Gericke, K., Eckert, C., & Stacey, M. (2017). What do we need to say about a design method? In A. Maier, S. Škec, H. Kim, M. Kokkolaras, J. Oehmen, G. Fadel, F. Salustri, & M. van der Loos (Eds.), *21th International Conference on Engineering Design (ICED 2015)*, 21.-25.08.2017, Vancouver, Canada (pp. 101–110).
- Gero, J. S., & Milovanovic, J. (2020). A framework for studying design thinking through measuring designers' minds, bodies and brains. *Design Science*, 6, e19. <https://doi.org/10.1017/dsj.2020.15>
- Grier, R. A. (2015). How High is High? A Meta-Analysis of NASA-TLX Global Workload Scores. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1), 1727–1731. <https://doi.org/10.1177/1541931215591373>
- Grohs, J., Shealy, T., Maczka, D., Hu, M., Panneton, R., & Yang, X. (2017). Evaluating the Potential of fNIRS Neuroimaging to Study Engineering Problem Solving and Design. *2017 ASEE Annual Conference & Exposition Proceedings*, 28305. <https://doi.org/10.18260/1-2--28305>
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- Halverson, T., Estepp, J., Christensen, J., & Monnin, J. (2012). Classifying Workload with Eye Movements in a Complex Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1), 168–172. <https://doi.org/10.1177/1071181312561012>

- Hancock, G. M., Longo, L., Young, M. S., & Hancock, P. A. (2021). Mental Workload. In G. Salvendy & W. Karwowski (Eds.), *Handbook of Human Factors and Ergonomics* (5th Edition, pp. 203–226). Wiley.
- 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. <https://doi.org/10.1177/154193120605000909>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 52, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd Edition). Springer. <https://doi.org/10.1007/b94608>
- Hay, L., Cash, P. J., & McKilligan, S. (2020). The future of design cognition analysis. *Design Science*, 6, e20. <https://doi.org/10.1017/dsj.2020.20>
- Hay, L., Duffy, A. H. B., McTeague, C., Pidgeon, L. M., Vuletic, T., & Greal, M. (2017). A systematic review of protocol studies on conceptual design cognition: Design as search and exploration. *Design Science*, 3, e10. <https://doi.org/10.1017/dsj.2017.11>
- He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, A. L., & Christ, R. E. (1992). Comparison of Four Subjective Workload Rating Scales. *Human Factors*, 34(4), 429–439. <https://doi.org/10.1177/001872089203400405>
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford Univ. Press.

- Howard, T. J., Culley, S. J., & Dekoninck, E. (2008). Describing the creative design process by the integration of engineering design and cognitive psychology literature. *Design Studies*, 29(2), 160–180. <https://doi.org/10.1016/j.destud.2008.01.001>
- Hsu, W., & Liu, B. (2000). Conceptual design: issues and challenges. *Computer-Aided Design*, 32(14), 849–850. [https://doi.org/10.1016/S0010-4485\(00\)00074-9](https://doi.org/10.1016/S0010-4485(00)00074-9)
- Hussy, W., Schreier, M., & Echterhoff, G. (2013). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor* (2nd Edition). Springer. <https://doi.org/10.1007/978-3-642-34362-9>
- Jain, A. K., Duin, P. W., & Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4–37. <https://doi.org/10.1109/34.824819>
- Jeatrakul, P., & Wong, K. W. (2012). Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 10.-15.06.2012, Brisbane, QLD, Australia (pp. 1–8). <https://doi.org/10.1109/IJCNN.2012.6252450>
- Jin, Y., & Benami, O. (2010). Creative patterns and stimulation in conceptual design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 24(2), 191–209. <https://doi.org/10.1017/S0890060410000053>
- Jonassen, D., Strobel, J., & Lee, C. B. (2006). Everyday Problem Solving in Engineering: Lessons for Engineering Educators. *Journal of Engineering Education*, 95(2), 139–151. <https://doi.org/10.1002/j.2168-9830.2006.tb00885.x>
- Klingner, J. M. (2010). *Measuring Cognitive Load During Visual Tasks by Combining Pupillometry and Eye Tracking* [Doctoral dissertation, Stanford University]. <https://graphics.stanford.edu/papers/klingner-dissertation/klingner-dissertation.pdf>

- Kokotovich, V. (2008). Problem analysis and thinking tools: an empirical study of non-hierarchical mind mapping. *Design Studies*, 29(1), 49–69. <https://doi.org/10.1016/j.destud.2007.09.001>
- Kroll, E., & Weisbrod, G. (2020). Testing and evaluating the applicability and effectiveness of the new idea-configuration-evaluation (ICE) method of conceptual design. *Research in Engineering Design*, 31(1), 103–122. <https://doi.org/10.1007/s00163-019-00324-6>
- Lavrsen, J. C., & Daalhuizen, J. (2024). Balancing cognitive load in design work: A conceptual and narrative review. In C. M. Gray, E. Ciliotta Chehade, P. Hekkert, L. Forlano, P. Ciuccarelli, & P. Lloyd (Eds.), *DRS2024: Boston*, 23.-28.06.2024, Boston, MA, USA (pp. 1–17). <https://doi.org/10.21606/drs.2024.928>
- Lawrie, E., Hay, L., & Wodehouse, A. (2023). A Classification of Methods and Constructs in Design Cognition Research. In J. S. Gero (Ed.), *Design Computing and Cognition'22* (pp. 97–114). Springer International Publishing. https://doi.org/10.1007/978-3-031-20418-0_7
- Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection: A data perspective. *ACM Computing Surveys*, 50(6), 1–45. <https://doi.org/10.1145/3136625>
- Lohmeyer, Q., & Meboldt, M. (2016). The Integration of Quantitative Biometric Measures and Experimental Design Research. In P. J. Cash, T. Stanković, & M. Štorga (Eds.), *Experimental design research: Approaches, perspectives, applications* (pp. 97–112). Springer International Publishing. https://doi.org/10.1007/978-3-319-33781-4_6
- Longo, L. (2018). On the Reliability, Validity and Sensitivity of Three Mental Workload Assessment Techniques for the Evaluation of Instructional Designs: A Case Study in a Third-level Course. In B. M. McLaren, R. Reilly, S. Zvacek, & J. Uhomobhi (Eds.), *CSEDU 2018 - Proceedings of the 10th International Conference on Computer Supported Education*, 15.03.2018 - 17.03.2018, Funchal, Madeira, Portugal (pp. 166–178). <https://doi.org/10.5220/0006801801660178>
- Marquart, G., Cabrall, C., & de Winter, J. C. F. (2015). Review of Eye-related Measures of Drivers' Mental Workload. *Procedia Manufacturing*, 3, 2854–2861. <https://doi.org/10.1016/j.promfg.2015.07.783>

- Matthiesen, S. (2021). Gestaltung – Prozess und Methoden. In B. Bender & K. Gericke (Eds.), *Pahl/Beitz Konstruktionslehre: Methoden und Anwendung erfolgreicher Produktentwicklung* (9. Auflage, pp. 397–465). Springer Vieweg. https://doi.org/10.1007/978-3-662-57303-7_13
- Matthiesen, S., Nelius, T., & Eisenmann, M. (2022). An Approach to Develop Designer-Centred Methods: Illustrated by an Example on How to Overcome Cognitive Bias in Product Development. In D. Krause & E. Heyden (Eds.) (Vol. 15, pp. 183–200). Springer. https://doi.org/10.1007/978-3-030-78368-6_10
- Meisenbacher, S., Turowski, M., Phipps, K., Rätz, M., Müller, D., Hagenmeyer, V., & Mikut, R. (2022). Review of automated time series forecasting pipelines. *WIREs Data Mining and Knowledge Discovery*, 12(6), Article e1475. <https://doi.org/10.1002/widm.1475>
- Nelius, T., Doellken, M., Zimmerer, C., & Matthiesen, S. (2020). The impact of confirmation bias on reasoning and visual attention during analysis in engineering design: An eye tracking study. *Design Studies*, 71, 100963. <https://doi.org/10.1016/j.destud.2020.100963>
- Nelson, J., & Menold, J. (2020). Opening the black box: Developing metrics to assess the cognitive processes of prototyping. *Design Studies*, 70, 100964. <https://doi.org/10.1016/j.destud.2020.100964>
- Nikulin, C., Lopez, G., Piñonez, E., Gonzalez, L., & Zapata, P. (2019). NASA-TLX for predictability and measurability of instructional design models: case study in design methods. *Educational Technology Research and Development*, 67(2), 467–493. <https://doi.org/10.1007/s11423-019-09657-4>
- Nolte, H., & McComb, C. (2021). The cognitive experience of engineering design: an examination of first-year student stress across principal activities of the engineering design process. *Design Science*, 7, e3. <https://doi.org/10.1017/dsj.2020.32>
- Nourbakhsh, N., Wang, Y., & Chen, F. (2013). GSR and Blink Features for Cognitive Load Classification. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2013* (pp. 159–166). Springer Berlin Heidelberg.

- Olsen, A., & Matos, R. (2012, March 28). Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. In S. N. Spencer (Ed.), *ETRA '12: Proceedings of the Symposium on Eye Tracking Research and Applications*, 28.-30.03.2012, Santa Barbara, CA, USA (pp. 317–320).
<https://doi.org/10.1145/2168556.2168625>
- Paas, F. G. W. C., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1), 63–71.
https://doi.org/10.1207/S15326985EP3801_8
- Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6(4), 351–371. <https://doi.org/10.1007/BF02213420>
- Paas, F. G. W. C., van Merriënboer, J. J. G., & Adam, J. J. (1994). Measurement of Cognitive Load in Instructional Research. *Perceptual and Motor Skills*, 79(1), 419–430. <https://doi.org/10.2466/pms.1994.79.1.419>.
- Pahl, G., Beitz, W., Blessing, L. T., Feldhusen, J., Grote, K.-H., & Wallace, K. M. (2007). *Engineering Design: A Systematic Approach* (3rd Edition). Springer London. <https://doi.org/10.1007/978-1-84628-319-2>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://doi.org/10.5555/1953048.2078195>
- Pedrotti, M., Mirzaei, M. A., Tedesco, A., Chardonnet, J.-R., Mérienne, F., Benedetto, S., & Baccino, T. (2014). Automatic Stress Classification With Pupil Diameter Analysis. *International Journal of Human-Computer Interaction*, 30(3), 220–236.
<https://doi.org/10.1080/10447318.2013.848320>

- Petkar, H., Yadav, R., Nguyen, T. A., Zeng, Y., & Dande, S. (2009, August 30). A Pilot Study to Assess Designer's Mental Stress Using Eye Gaze System and Electroencephalogram. In *Proceedings of the ASME 2009 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2009*, 30.08-02.09.2009, San Diego, California, USA (pp. 899–909). American Society of Mechanical Engineers.
- Reid, G. B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. *Advances in Psychology*, 52, 185–218. [https://doi.org/10.1016/S0166-4115\(08\)62387-0](https://doi.org/10.1016/S0166-4115(08)62387-0)
- Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3, 1371--1382. <https://doi.org/10.5555/944919.944978>
- Ruckpaul, A., Fürstenhöfer, T., & Matthiesen, S. (2015). Combination of Eye Tracking and Think-Aloud Methods in Engineering Design Research. In J. S. Gero & S. Hanna (Eds.), *Design Computing and Cognition '14* (pp. 81–97). Springer International Publishing. <https://doi.org/10.1007/978-3-319-14956-1>
- Ruckpaul, A., Kriltz, A., & Matthiesen, S. (2017, October 14). Using Eye Tracking to Understand the Engineering Designers' Behaviour in Synthesis Driven Analyzing Processes - Experiences in Study Design. In M. Meboldt, S. Matthiesen, P. Badke-Schaub, & Q. Lohmeyer (Chairs), *Analyzing Cognitive Processes During Design: Proceedings of the HBiD 2014*, Ascona, Switzerland.
- Ruckpaul, A., Nelius, T., & Matthiesen, S. (2015, July 27). Differences in Analysis and Interpretation of Technical Systems by Expert and Novice Engineering Designers. In C. Weber, S. Husung, M. Cantamessa, D. Cascini, D. Marjanović, & M. Bordegoni (Eds.), *Proceedings of the 20th International Conference on Engineering Design (ICED 15)*, 27.-30.07.2015, Mailand, Italy (pp. 339–348).
- Sasaki, Y. (2007). The truth of the F-measure. *Teach Tutor Mater*, 1(5), 1–5. https://nicolasshu.com/assets/pdf/Sasaki_2007_The%20Truth%20of%20the%20F-measure.pdf

- Shah, J. J., Smith, S., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design Studies*, 24(2), 111–134.
[https://doi.org/10.1016/S0142-694X\(02\)00034-0](https://doi.org/10.1016/S0142-694X(02)00034-0)
- Shealy, T., Grohs, J., Hu, M., Maczka, D., & Panneton, R. (2017, June 24). Investigating Design Cognition during Brainstorming Tasks with Freshmen and Senior Engineering Students using Functional Near Infrared Spectroscopy. In *2017 ASEE Annual Conference & Exposition*, 24.-28.6.2017, Columbus, OH, USA (p. 28584). American Society for Engineering Education. <https://doi.org/10.18260/1-2--28584>
- Sun, G., & Yao, S. (2012). Investigating the relation between cognitive load and creativity in the conceptual design process. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1), 308–312.
- Sun, L., Xiang, W., Chai, C., Yang, Z., & Zhang, K. (2014). Designers' perception during sketching: An examination of Creative Segment theory using eye movements. *Design Studies*, 35(6), 593–613.
<https://doi.org/10.1016/j.destud.2014.04.004>
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory* (1st ed.). *Explorations in the Learning Sciences, Instructional Systems and Performance Technologies: Vol. 1*. Springer. <https://doi.org/10.1007/978-1-4419-8126-4>
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, 31, 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Tan, J. J. Y., Otto, K. N., & Wood, K. L. (2017). Relative impact of early versus late design decisions in systems development. *Design Science*, 3, e12.
<https://doi.org/10.1017/dsj.2017.13>
- Ullman, D. G. (2010). *The mechanical design process* (4th Edition). *McGraw-Hill series in mechanical engineering*. McGraw-Hill Higher Education.
- van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. Academic Press.

- Verein Deutscher Ingenieure e.V. (Juni 1997). *Konstruktionsmethodik - Methodisches Entwickeln von Lösungsprinzipien Blatt 1* (VDI-Richtlinie VDI 2222 Blatt 1). Beuth Verlag GmbH.
- Vidulich, M. A., & Tsang, P. S. (1987). Absolute Magnitude Estimation and Relative Judgement Approaches to Subjective Workload Assessment. *Proceedings of the Human Factors Society Annual Meeting*, 31(9), 1057–1061. <https://doi.org/10.1177/154193128703100930>
- Villanueva, I., Campbell, B. D., Raikes, A. C., Jones, S. H., & Putney, L. G. (2018). A Multimodal Exploration of Engineering Students' Emotions and Electrodermal Activity in Design Activities. *Journal of Engineering Education*, 107(3), 414–441. <https://doi.org/10.1002/jee.20225>
- Weisbrod, G., & Kroll, E. (2018). Idea-configuration-evaluation (ICE): development and demonstration of a new prescriptive model of the conceptual engineering design process based on parameter analysis and C–K theory. *Research in Engineering Design*, 29, 203–225. <https://doi.org/10.1007/s00163-017-0263-6>
- Wierwille, W. W., & Casali, J. G. (1983). A Validated Rating Scale for Global Mental Workload Measurement Applications. *Proceedings of the Human Factors Society Annual Meeting*, 27(2), 129–133. <https://doi.org/10.1177/154193128302700203>
- Wierwille, W. W., Wreggit, S. S., & Knipling, R. R. (1994). Development of improved algorithms for on-line detection of driver drowsiness. In *SAE International P: Vol. 283. Leading change: The transportation electronic revolution; proceedings of the 1994 International Congress on Transportation Electronics* (pp. 331–340). Society of Automotive Engineers.
- Wittel, H., Muhs, D., Jannasch, D., & Voßiek, J. (2013). *Roloff/Matek Maschinenelemente: Normung, Berechnung, Gestaltung* (21st Edition). Springer Vieweg. <https://doi.org/10.1007/978-3-658-02327-0>
- Wu, C., Cha, J., Sulek, J., Zhou, T., Sundaram, C. P., Wachs, J., & Yu, D. (2020). Eye-Tracking Metrics Predict Perceived Workload in Robotic Surgical Skills Training. *Human Factors*, 62(8), 1365–1386. <https://doi.org/10.1177/0018720819874544>

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5), 459–482. <https://doi.org/10.1002/cne.920180503>

Zagermann, J., Pfeil, U., & Reiterer, H. (2018, April 21). Studying Eye Movements as a Basis for Measuring Cognitive Load. In R. Mandryk & M. Hancock (Eds.), *CHI EA '18: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 21.-26.04.2018, Montreal, QC, Canada (pp. 1–6). <https://doi.org/10.1145/3170427.3188628>

Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O. N. M., & Atkins, M. S. (2012). Workload assessment of surgeons: Correlation between NASA TLX and blinks. *Surgical Endoscopy*, 26, 2746–2750. <https://doi.org/10.1007/s00464-012-2268-6>

Zimmerer, C., & Matthiesen, S. (under revision). Challenges in conceptual design: A comparison between engineering design students and experienced designers.

Zimmerer, C., & Matthiesen, S. (2021). Study on the impact of cognitive load on performance in engineering design. In O. Isaksson (Ed.), *Proceedings of the International Conference on Engineering Design (ICED21)*, 16.-20-08.2021, Gothenburg, Sweden (2761-2770). <https://doi.org/10.1017/pds.2021.537>

Zimmerer, C., Nelius, T., & Matthiesen, S. (2023). Using Eye Tracking to Measure Cognitive Load of Designers in Situ. In J. S. Gero (Ed.), *Design Computing and Cognition'22* (pp. 481–495). Springer International Publishing. https://doi.org/10.1007/978-3-031-20418-0_29

Zimmerer, C., Wittig, C., & Matthiesen, S. (2025). Towards an automated evaluation of design processes—an algorithm to predict critical situations during concept synthesis. *Research in Engineering Design*, 36(3). <https://doi.org/10.1007/s00163-025-00444-2>

Publications Co-authored by the Author of this Thesis

- Doellken, M., Zimmerer, C. & Matthiesen, S. (2020). Challenges faced by design engineers when considering manufacturing in design - An interview study. *Proceedings of the Design Society: DESIGN Conference*, 1, 837–846. <https://doi.org/10.1017/dsd.2020.302>
- Matthiesen, S., Zimmerer, C. & Doellken, M. (2022). Entwicklung und Validierung von Gestaltungsrichtlinien: Gestaltungsrichtlinien durch erfolgreiche Vorgehensweisen intuitiver gestalten um Herstellungskosten zu minimieren. *Konstruktion*, 74(3), 53.
- Matthiesen, S., Zimmerer, C., Eisenmann, M., Pähler, L. & Grauberger, P. (2024). Der Prozess Des Wissensaufbaus In Der Produktentwicklung. In S. Matthiesen & P. Grauberger (Hrsg.), *Konstruktionswissen für Ingenieure* (S. 57–84). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-68986-8_2
- Matthiesen, S., Zimmerer, C., Pähler, L. & Grauberger, P. (2024). Wissen in der Produktentwicklung – Grundlagen. In S. Matthiesen & P. Grauberger (Hrsg.), *Konstruktionswissen für Ingenieure* (S. 1–56). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-68986-8_1
- Nelius, T., Doellken, M., Zimmerer, C. & Matthiesen, S. (2020). The impact of confirmation bias on reasoning and visual attention during analysis in engineering design: An eye tracking study. *Design Studies*, 71. <https://doi.org/10.1016/j.destud.2020.100963>
- Peters, J., Zimmerer, C., Gwosch, T., Herbst, F., Hartmann, C., Chadda, R. et al. (2022). Test-driven Development to Overcome Challenges in the Design of Sensor-integrating Machine Elements. In D. Krause, K. Paetzold & S. Wartack (Hrsg.), *DS 119: Proceedings of the 33rd Symposium Design for X (DFX2022)* (S. 1–10). The Design Society. <https://doi.org/10.35199/dfx2022.12>

- Steinhauser, N., Zimmerer, C., Grauberger, P., Neliu, T. & Matthiesen, S. (2023). FUNCTIONAL ANALYSIS IN PHYSICAL AND VIRTUAL REALITY (VR) ENVIRONMENTS – A COMPARATIVE STUDY. *Proceedings of the Design Society*, 3, 2015–2024. <https://doi.org/10.1017/pds.2023.202>
- Wittig, C., Zimmerer, C., Hemmerich, J. & Matthiesen, S. (2024). Requirements for utilizing student projects as validation environment for design methods and collecting research data. In DS 133: *Proceedings of the 35th Symposium Design for X (DFX2024)* (S. 125–134). The Design Society. <https://doi.org/10.35199/dfx2024.13>
- Zimmerer, C. & Matthiesen, S. (2021). Study on the impact of cognitive load on performance in engineering design. *Proceedings of the Design Society*, 1, 2761–2770. <https://doi.org/10.1017/pds.2021.537>
- Zimmerer, C., Neliu, T. & Matthiesen, S. (2021). Investigation on the Aha-Experience as an Indicator of Correct Solutions in Functional Analysis in Engineering Design. *Proceedings of the Design Society*, 1, 1947–1956. <https://doi.org/10.1017/pds.2021.456>
- Zimmerer, C., Neliu, T. & Matthiesen, S. (2023). Using Eye Tracking to Measure Cognitive Load of Designers in Situ. In J. S. Gero (Hrsg.), *Design Computing and Cognition'22* (S. 481–495). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-20418-0_29
- Zimmerer, C., Wittig, C., Grauberger, P. & Matthiesen, S. (2025). Mit empirischem Testing zur Innovation. Konstruktionswissen als Schlüssel zu erfolgreichen Produkten. *Konstruktion*, 77(06), 61.
- Zimmerer, C., Wittig, C. & Matthiesen, S. (2025). Towards an automated evaluation of design processes - an algorithm to predict critical situations during concept synthesis. *Research in Engineering Design*, 36(3). <https://doi.org/10.1007/s00163-025-00444-2>

Use of Tools for assisted Copy Editing

The following AI-based tools were used for copy editing during the creation of this work:

- DeepL
- Grammarly
- ChatGPT

These tools were used solely to improve wording, grammar, and formatting. This includes adjustments to enhance readability and style, as well as to check spelling, and punctuation. There was no use of these tools for generative purposes at any time.

