Doctoral theses at NTNU, 2023:223

Henrikke Dybvik

Introducing fNIRS to multimodal in-situ experiments in design research

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Engineering
Department of Mechanical and Industrial
Engineering

NTNU
Norwegian University of
Science and Technology

Henrikke Dybvik

Introducing fNIRS to multimodal in-situ experiments in design research

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2023

Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Engineering
Department of Mechanical and Industrial Engineering

ISBN 978-82-326-7146-5 (printed ver.)
ISBN 978-82-326-7145-8 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:223

Printed by NTNU Grafisk senter

To myself in 2019.

You’re right, this isn’t real life. Keep going.

Thank you for not giving up.

Abstract

This PhD dissertation introduces fNIRS to multimodal, in-situ experiments in design research.

Design serves physical, intellectual, and emotional human needs. Design solves problems. Through design and development of products, technology, interactions, experiences, etc., the design field responds to new challenges—for a better world and the people in it. In pursuit of improving the outcome of designing, design research develops scientific knowledge about design—and how to improve it—by increasing our understanding of designers' processes and tools, and users' experience or interaction with designed artifacts.

Experiments are foundational for generating scientific knowledge and impactful guidance of design. Physiological sensors and neuroimaging enable better experimental studies of design’s impact on the human experience, by accounting for individual variation and human unpredictability, while providing objective information of human cognition. Triangulation of multiple sensor- and neuroimaging modalities is beneficial as it reduces bias and error, while increasing rigor and validity of results. Further, experiments conducted in the real world, i.e., in-situ, establish how humans design and appropriate designed artifacts in their intended contexts, avoiding experimental biases common in constrained laboratory environments, ultimately evidencing more generalizable results. Therefore, in our design research, we are working towards conducting in-situ multimodal experiments.

Electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) are historically the most frequently used neuroimaging modalities in design. However, they cannot be used in-situ, as both are extremely prone to motion artifacts. Functional near-infrared spectroscopy (fNIRS) is in comparison relatively robust to motion, while offering a good tradeoff between spatial- and temporal resolution. Thus, fNIRS opens a venue of research opportunities. fNIRS allow in-situ investigation of high-performing, established design teams working in industry/companies, their design processes, and team dynamics. We could evaluate different interface designs of e.g., a shore control center for remote ship operation, based on users’ cognitive load. fNIRS could quantify the effects of architecture on humans’ cognitive state and well-being.

Therefore, this dissertation aims to introduce fNIRS to multimodal experiments in design. Thereby, enabling multimodal in-situ experiments, advancing our scientific knowledge of design, to ultimately improve design, i.e., better responding to human needs.

First, we provide a foundation for multimodal experimental design research conducted in-situ. Thereafter, we introduce fNIRS, its theoretical and technical principles, design and analysis of experiments, fNIRS applied in-situ, and in design research. Lastly, we discuss current limitations, implications, and recommendations for future research.

Preface

This thesis has been submitted to the Norwegian University of Science and Technology (NTNU) for the degree of Philosophiae Doctor (PhD). The work has been conducted by Henrikke Dybvik at TrollLABS at the Department of Mechanical and Industrial Engineering (MTP), Faculty of Engineering Science (IV), NTNU. Professor Martin Steinert has supervised the doctoral work.

Throughout my time at TrollLABS, I have thus far written 18 academic articles (not counting those that died in the process), 15 of which are included as a part of this thesis. I, Henrikke Dybvik, am the first author of 10 of those articles.

Additionally, the doctoral work encompassed research stays abroad. A one-month stay at the Cognition and Computation in Design (Co-Design) Lab at the University of California, Berkeley, with the support of Assistant Professor Kosa Goucher-Lambert. A four months stay at Stanford University under the auspice of SCANCOR, during which research was conducted at the Center for Interdisciplinary Brain Sciences Research (CIBSR) with the support of Professor Allan L. Reiss. Numerous international conferences have been attended, both in-person (4) and online (6).

In this thesis, *we* is used, rather than *I,* to emphasize that research is a team sport. I could never have conducted all this research alone: I am grateful to all my co-authors for our collaborations, and thank all participants. Although we use the pronoun “we”, the opinions expressed in this doctoral dissertation are to be considered those of the author. Any errors are, of course, the sole responsibility of the author.

Henrikke Dybvik
Trondheim, February 2023

Better late than never

Acknowledgments

This thesis has been a long time coming. There are many people I want to thank.

First, foremost. My supervisor, Professor Martin Steinert. I would not have embarked on a PhD had it not been for you, your guidance, and the great environment you have created at TrollLABS. For inspiring, guiding, supporting, promoting, advocating, sponsoring, and mentoring me, thank you. Thank you for being patient with me when I am not patient with myself.

To all Steinerts, thank you for letting me hang out at Vikhammer for a good nine months. You are all fantastic. Andrea, thanks for your continued, kind support. Lucia, thanks lending me your room.

TrollLABS feels like a second home at this point. I want to thank all my fellow Trolls. Thanks to TrollLABS generation 1: Andreas Simskar Wulvik, Heikki Sjöman, Carlo Kriesi, Jørgen Erichsen, Jørgen Blindheim, Achim Gerstenberg, Kristoffer Slåttsveen. Andreas, thanks for introducing me to a world of research and teaching me about craftsmanship. Thanks to TrollLABS generation 2—my own generation: Marius Auflem, Håvard Vestad, Sampsa Kohtala, Torjus Steffensen, Evangelos Tyflopoulos. Thanks for all the support, encouragement, and discussions in the last years. It's a strange realization that, what was supposed to be 3 years of PhD, but became around 5, is finished. Time has flown by. Thanks to TrollLABS generation 3: Sindre Wold Eikevåg, Daniel Ege, Kim Christensen, Kjetil Baglo, Ole Nesheim. Thanks for breathing fresh air into the lab and continuing TrollLABS' legacy. Pasi, I'm not exactly sure which generation you're a part of (maybe 2.5?), but thank you for all the interesting, and derailing, conversations about architecture. All master students at TrollLABS, in particular those I have been so fortunate to co-supervise, thank you. Thank you Sindre and Marius for reading through and providing feedback on "the wall of text".

Big thank you and hugs to my delightful friends. Ingrid Lande, Marthe Ness, Elin Marie Peersen Ringvoll, and Karoline Hopland. I can always call any one of you, whether it's been six days or six months since last time. You answer and we pick up right where we left of. I am immensely grateful for that. For your encouragement, support, belief in me, and humor. Thank you.

Area 37. My dear nuclear family. Thank you. Thanks to my mom and dad for giving me a name that makes every new conversation in academia predictable ("How do I pronounce your name?") and for teaching me to believe that I could do anything I wanted to—I just had to work for it. My dear sister—the beacon of knowledge—Rebekka. Although you're supposed to be the little sister, it often feels like I am. Your advice throughout the years have been most valuable and meant a lot to me. You are smart and courageous. As one of the strongest people I know, you continue to inspire me every day, professionally and personally. For being a great sister and friend—thank you. Mom, Dad, Rebekka, I feel privileged and lucky to have had your unconditional and relentless support, encouragement, and love. I love you all very much.

I want to thank my grandparents for always encouraging, supporting, and stressing the importance of education. My grandfather, Bestefar Asbjørn, for being the original globetrotter. Your stories of sailing around the world as a young telegraphist to support your education have been exceptionally inspiring. As far back as I can remember, I've been unafraid and keen on exploring the world, perhaps thanks to you. My grandmother, bestemor Judith, for always giving me a hug.

My cousins, aunts, uncles, and the rest of my huge family—thank you. Kristina, you are a truly amazing woman, and I am so happy you're family.

There are a few "unmentionables" I want to thank. I won't mention names. I know you'll understand. For all your help, truly, thank you.

Dag, thank you for doing yoga and lockdown with me.

Everyone that has housed me at some point during this PhD—there are quite many of you: Elin, Ådne, and Hanne, Ragnhild and family, my mom and dad, Rebekka, Onkel Lars, and so the list goes on. Thank you so much.

fNIRS Analysis Club facebook group and various other discussion forums that have helped with scripts for data collection and analysis.

CIBSR, for a warm welcoming and making me feel like a part of the lab—thanks Steph, Safiyyah, Apurva, and Cassie. SCANCOR, for your ridiculously unfounded support and belief in me—thank you.

My best friend, partner, teammate. Sindre. I am immensely grateful to wake up next to you every day. I would never have thought I would be so lucky as to fall in love with my best friend and that they would fall in love with me too—but here we are. These last months have been some of the best ones in my life (which is saying something, given that I've spent parts of that writing this dissertation). For your support and enthusiasm, thank you very much my love.

Content

List of contributions

C1 Wulvik, A. S., Dybvik, H., & Steinert, M. (2019). Investigating the relationship between mental state (workload and affect) and physiology in a control room setting (ship bridge simulator). Cognition, Technology & Work. https://doi.org/10.1007/s10111-019-00553-8
Journal article.

C2 Dybvik, H., Wulvik, A., & Steinert, M. (2018). STEERING A SHIP - INVESTIGATING AFFECTIVE STATE AND WORKLOAD IN SHIP SIMULATIONS. Proceedings of the Design Society: DESIGN Conference, 2003–2014. https://doi.org/10.21278/idc.2018.0459
1st author. Conference article.

C3 Rørvik, S. B., Auflem, M., Dybvik, H., & Steinert, M. (2021). Perception by Palpation: Development and Testing of a Haptic Ferrogranular Jamming Surface. Frontiers in Robotics and AI, 8, 311. https://doi.org/10.3389/frobt.2021.745234
Journal article.

C4 Dybvik, H., Abelson, F.G., Aalto, P., Goucher-Lambert, K., Steinert, M. (2023). Inspirational Stimuli Attain Visual Allocation: Examining Design Ideation with Eye-Tracking. In: Gero, J.S. (eds) Design Computing and Cognition'22. DCC 2022. Springer, Cham. https://doi.org/10.1007/978-3-031-20418-0_28
1st author. Conference article.

C5 Dybvik, H., Abelson, F. G., Aalto, P., Goucher-Lambert, K., & Steinert, M. (2022). Inspirational Stimuli Improve Idea Fluency during Ideation: A Replication and Extension Study with Eye-Tracking. Proceedings of the Design Society, 2, 861–870. https://doi.org/10.1017/pds.2022.88
1st author. Conference article.

C6 Dybvik, H., Løland, M., Gerstenberg, A., Slåttsveen, K. B., & Steinert, M. (2021). A low-cost predictive display for teleoperation: Investigating effects on human performance and workload. International Journal of Human-Computer Studies, 145, 102536. https://doi.org/10.1016/j.ijhcs.2020.102536
1st author. Journal article.

C7 Veitch, E., Dybvik, H., Steinert, M., & Alsos, O. A. (2022). Collaborative Work with Highly Automated Marine Navigation Systems. Computer Supported Cooperative Work (CSCW). https://doi.org/10.1007/s10606-022-09450-7
Journal article

C8 Dybvik, H., Veitch, E., & Steinert, M. (2020). EXPLORING CHALLENGES WITH DESIGNING AND DEVELOPING SHORE CONTROL CENTERS (SCC) FOR AUTONOMOUS SHIPS. Proceedings of the Design Society: DESIGN Conference, 1, 847–856. https://doi.org/10/ggz7zb
1st author. Conference article.

C9 Dybvik, H., & Steinert, M. (Under review). Operationalized hypotheses build bridges between qualitative and quantitative design research. Journal of Mixed Methods Research.
1st author. Journal article.

C10 Hatlem, L. A., Chen, J., Dybvik, H., & Steinert, M. (2020). A Modular Research Platform – Proof-of-Concept of a Flexible Experiment Setup Developed for Rapid Testing of Simulators, UIs and Human Physiology Sensors. Procedia CIRP, 91, 407–414. https://doi.org/10.1016/j.procir.2020.02.193
Conference article.

C11 Erichsen, C. K., Dybvik, H., & Steinert, M. (2020). Integration of low-cost, dry-comb EEG-electrodes with a standard electrode cap for multimodal signal acquisition during human experiments. DS 101: Proceedings of NordDesign 2020, Lyngby, Denmark, 12th - 14th August 2020, 1–12. https://doi.org/10.35199/NORDDESIGN2020.19
Conference article.

C12 Dybvik, H., & Steinert, M. (2021). Real-World fNIRS Brain Activity Measurements during Ashtanga Vinyasa Yoga. Brain Sciences, 11(6), 742. https://doi.org/10.3390/brainsci11060742
1st author. Journal article.

C13 Dybvik, H., Kuster Erichsen, C., Steinert, M. (2021) 'Description of a Wearable Electroencephalography + Functional Near-Infrared Spectroscopy (EEG+fNIRS) for In-situ Experiments on Design Cognition', in Proceedings of the International Conference on Engineering Design (ICED21), Gothenburg, Sweden, 16-20 August 2021. https://doi.org/10.1017/pds.2021.94
1st author. Conference article.

C14 Dybvik, H., Erichsen, C. K., Steinert, M. (2021) 'Demonstrating the Feasibility of Multimodal Neuroimaging Data Capture with a Wearable Electoencephalography + Functional Near-Infrared Spectroscopy (EEG+fNIRS) in-situ', in Proceedings of the International Conference on Engineering Design (ICED21), Gothenburg, Sweden, 16-20 August 2021. https://doi.org/10.1017/pds.2021.90
1st author. Conference article.

C15 Dybvik, H., Erichsen, C. K., Steinert, M. (Manuscript). Tetris' effect on cognitive load, performance, and systemic neurophysiology. To be submitted to Frontiers in Human Neuroscience.
1st author. Journal article.

C16 Aalto, P., Dybvik, H., & Steinert, M. (Under review). A stroll through a cathedral: FNIRS and space sequences in architecture. Frontiers in Neuroergonomics.
Journal abstract.

D1 The Ashtanga Vinyasa Yoga Data Set
Dybvik, H. (2023, February 23). The Ashtanga Vinyasa Yoga Data Set. https://doi.org/10.17605/OSF.IO/F8SQ3

D2 The Design Ideation Data Set
Abelson, Filip Gornitzka; Dybvik, Henrikke; Steinert, Martin, (2021) "Dataset for Design Ideation Study", https://doi.org/10.18710/PZQC4A, DataverseNO, V1, UNF:6:1NVkv2+s87SAf1qR8RRwvg== [fileUNF]
Abelson, F. G., Dybvik, H., & Steinert, M. (2021). Raw Data for Design Ideation Study. Norges teknisk-naturvitenskapelige universitet. https://doi.org/10.21400/7KQ02WJL

D3 The Palpation Data Set
Rørvik, Sigurd Bjarne; Auflem, Marius; Dybvik, Henrikke; Steinert, Martin, (2021), "Replication Data for: Perception by Palpation: Development and Testing of a Haptic Ferrogranular Jamming Surface", https://doi.org/10.18710/OCMXVP, DataverseNO, V1

D4 The Tetris Data Set
[Will be published along with manuscript C15]

Figures

Tables

Abbreviations and definitions

AR-IRLS	Autoregressive, iterative robust least-squares model
AUC	Area under curve
CBF	Cerebral blood flow
CBSI	Correlation-Based Signal Improvement
Cerebral cortex	The outer grey matter layer covering the entire human brain
Chromophore	The part of a molecule responsible for its color
CSF	Cerebrospinal fluid
CV	Coefficient of variation
CW	Continuous wavelength
DMN	Default Mode Network
ECG	Electrocardiography
EDA	Electrodermal Activity
EEG	Electroencephalography
ERP	Event-related potentials
FC	Functional Connectivity
fNIRS	Functional near-infrared spectroscopy
fMRI	Functional magnetic resonance imaging
FD	Frequency-domain
FDR	False discovery rate
FWER	Family-wise error rate
GSR	Galvanic skin response
GLM	General Linear Model
HbO_2	Oxygenated hemoglobin
HbR	Deoxygenated hemoglobin
HR	Hemodynamic response
HRF	Hemodynamic response function
Hyperscanning	Simultaneous functional neuroimaging of two or more brains
IBS	Interbrain synchrony
ICA	Independent component analysis
ISI	Inter-Stimulus Interval
MNI	Montreal Neurological Institute
MRI	Magnetic resonance imaging
NASA-TLX	NASA Task Load Index
NIR	Near-infrared
OLS	Ordinary Least Squares

PCA	Principal component analysis
PLSR	Partial least-squares regression
PSP	Peak spectral power
ROC	Receiver operator characteristic
ROI	Region of interest
SCI	Scalp coupling index
SD	Standard deviation
SNR	Signal-to-noise ratio
SS	Short separation (about fNIRS channel)
SQI	Signal quality index
TD	Time-domain
TDDR	Temporal Derivative Distribution Repair
Vasculature	The blood vessels or arrangement of blood vessels in an organ or body part

1 Introduction

This is a cumulative, article-based PhD dissertation according to NTNU's Regulations concerning the degree of Philosophiae Doctor (PhD).

This dissertation's contributions comprise 15 academic articles, one abstract, and 3 publicly available data sets. The articles include five published peer-reviewed journal articles, one journal article under peer-review, one journal manuscript to be submitted for peer-review, and eight peer-reviewed conference articles. All published articles appear in internationally recognized journals and conferences. Because the articles have been[1] peer-reviewed, they are an established part of the scientific literature. Therefore, in the following text, the academic contributions will appear as regular citations, highlighted in bold, e.g., **(Dybvik & Steinert, 2021)**.

The following text ("kappa") briefly overviews the research topic, the components of experimental design research, fNIRS, and the increasing use of fNIRS as a neuroscientific tool in design research.

1.1 Motivation

> McKim (1959) expressed that "human values and needs is of prime importance to the designer. Design is, after all, a response to human needs." (Auernhammer & Roth, 2021)

Design purposefully serve physical, intellectual, and emotional human needs (Auernhammer & Roth, 2021; McKim, 1959). Design may be understood and used from different perspectives (Auernhammer & Roth, n.d.). Often, it "is […] conceptualized as a process or set of activities incorporating various methods" (Auernhammer & Roth, 2021) that provide a comprehensive, creative, and humanistic design philosophy and practice (Auernhammer & Roth, 2021). Thereby, inventing "radical or transformative new solutions and designs that obsolete existing barriers and problems" (Leifer & Steinert, 2011), leading to entrepreneurship, innovation, and growth (Auernhammer & Roth, 2021). In essence, design solves problems, providing solutions to human needs (Auernhammer & Roth, 2021). Thus, design may be defined as the design and development of products, services, technology, interactions, and experiences in response to new challenges—for a better world and the people in it.

Design research seeks to develop scientific knowledge about design and how we might improve it. By enhancing our understanding of processes and tools employed by designers and users' experience or interaction with the designed solution or artifact, through scientific inquiry, ultimately, design research aims to improve the outcome of designing (Auernhammer & Roth, n.d.; Cash et al., 2016; Gero & Milovanovic, 2020).

Experiments provide a foundation for generating scientific knowledge and impactful guidance of design (Auernhammer & Roth, n.d.; Balters & Steinert, 2017; Blessing & Chakrabarti, 2009; Borgianni & Maccioni, 2020; Cairns & Cox, 2008; Cash et al., 2016; Hay et al., 2020). Experiments may integrate physiological sensors and neuroimaging to provide objective information about human cognition, which by accounting for individual variation and human unpredictability, better allow us to study the impact of design on

[1] or will be

the human experience (Balters et al., 2023; Balters & Steinert, 2017; Borgianni & Maccioni, 2020; Hu & Shepley, 2022; Steinert & Jablokow, 2013). Triangulation between multiple modalities in experiments increases the rigor and validity of results; at the same time, triangulation reduces bias and error, thereby allowing better interpretation and, thus, more accurate conclusions (Balters & Steinert, 2017; Blessing & Chakrabarti, 2009; Cash et al., 2016; Pinti et al., 2019; Steinert & Jablokow, 2013). Experiments conducted in the real world—i.e., *in-situ*—establish how humans design and appropriate designed artifacts in their intended contexts (Consolvo et al., 2007; Mayseless et al., 2019; Okamoto et al., 2004b). In-situ experiments avoid experimental biases common in constrained laboratory environments, evidencing more generalizable results (Balters & Steinert, 2017; Cash et al., 2016; Mayseless et al., 2019; Okamoto et al., 2004b; **Wulvik et al., 2019**). Therefore, in our design research, we are working towards conducting in-situ multimodal experiments.

Historically, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) are the most frequently used neuroimaging modalities in design (Balters et al., 2023; Balters & Steinert, 2017; Gero & Milovanovic, 2020). Although EEG and fMRI might seem complementary, suitable for triangulation—EEG has high temporal resolution, but low spatial resolution, while fMRI has low temporal resolution, but high spatial resolution—both are extremely prone to motion artifacts and can thus not be used in in-situ experiments (Balters et al., 2023; Gero & Milovanovic, 2020). In comparison, functional near-infrared spectroscopy (fNIRS) is relatively robust to motion, offering an acceptable tradeoff between spatial- and temporal resolution (Ferrari & Quaresima, 2012; Pinti et al., 2020; Quaresima & Ferrari, 2019). Thus, fNIRS opens a venue of research opportunities. fNIRS allow us to investigate, compare, and evaluate different interface designs of e.g., shore control centers for remote ship operation[2] based on users cognitive load in different ship operation scenarios[3]. We could investigate the design processes, and team dynamics of high-performing, established design teams working in industry—in-situ. FNRIS could measure the human response to architecture and the built environment, determining how facades, interior, object placements, lighting etc., impacts cognitive state and well-being. Therefore, this dissertation aims to introduce fNIRS to multimodal experiments in design research. We thus enable multimodal in-situ experiments that advance our scientific knowledge, to, ultimately, improve design and thus better respond to human needs.

1.2 Research objective

This dissertation aims to introduce fNIRS to in-situ multimodal experiments in design research.

1.3 Scope

It has been established that we are working towards multimodal in-situ experiments. To make this possible we need to, first, comprehend and outline the current foundation for experimental design research to understand its shortcomings and, thus, what is missing. This is described in Chapter 2. Second, we need to address those shortcomings by implementing fNIRS, described in Chapter 3. Although we are well on our way to genuine

[2] see **(Dybvik et al., 2020; Veitch et al., 2022)** for a discussion on designing shore control centers

[3] see **(Dybvik et al., 2018; Wulvik et al., 2019)** for a discussion of how measures of workload and affective state inferred from physiology measurements could aid designers

in-situ experiments, we are not fully there yet. Existing limitations are discussed in Chapter 4. Chapter 4 also discusses the contributions and future research.

1.4 List of contributions

This section presents this thesis' contributions. Table 1 lists the 15 academic publications and the abstract. Table 2 lists the 3 publicly available data sets.

NTNU follows a policy for open science (NTNU, 2020) in line with national goals and guidelines for open access publishing (Norwegian Ministry of Education and Research, 2017) and research data (Norwegian Ministry of Education and Research, 2018). Resting on principles of collaboration, transparency, verifiability, and accessibility, "open science adheres to the scientific ideals of knowledge as a public good, independence of research, universalism, and systematic critical appraisal of sources" (NTNU, 2020). In adherence to this, all academic contributions are published under open-access agreements. Associated research data are also made publicly available as far as possible.

1.4.1 Academic publications

Table 1 Contributions: Academic publications arranged by which chapter it pertains to.

No.	Citation
Chapter 2: Towards, and foundations for, in-situ multimodal experiments in design research	
C1	Wulvik, A. S., Dybvik, H., & Steinert, M. (2019). Investigating the relationship between mental state (workload and affect) and physiology in a control room setting (ship bridge simulator). Cognition, Technology & Work. https://doi.org/10.1007/s10111-019-00553-8
	Journal article.
C2	Dybvik, H., Wulvik, A., & Steinert, M. (2018). STEERING A SHIP - INVESTIGATING AFFECTIVE STATE AND WORKLOAD IN SHIP SIMULATIONS. Proceedings of the Design Society: DESIGN Conference, 2003–2014. https://doi.org/10.21278/idc.2018.0459
	1st author. Conference article.
C3	Rørvik, S. B., Auflem, M., Dybvik, H., & Steinert, M. (2021). Perception by Palpation: Development and Testing of a Haptic Ferrogranular Jamming Surface. Frontiers in Robotics and AI, 8, 311. https://doi.org/10.3389/frobt.2021.745234
	Journal article.
C4	Dybvik, H., Abelson, F.G., Aalto, P., Goucher-Lambert, K., Steinert, M. (2023). Inspirational Stimuli Attain Visual Allocation: Examining Design Ideation with Eye-Tracking. In: Gero, J.S. (eds) Design Computing and Cognition'22. DCC 2022. Springer, Cham. https://doi.org/10.1007/978-3-031-20418-0_28
	1st author. Conference article.
C5	Dybvik, H., Abelson, F. G., Aalto, P., Goucher-Lambert, K., & Steinert, M. (2022). Inspirational Stimuli Improve Idea Fluency during Ideation: A Replication and Extension Study with Eye-Tracking. Proceedings of the Design Society, 2, 861–870. https://doi.org/10.1017/pds.2022.88
	1st author. Conference article.
C6	Dybvik, H., Løland, M., Gerstenberg, A., Slåttsveen, K. B., & Steinert, M. (2021). A low-cost predictive display for teleoperation: Investigating effects on human performance and workload. International Journal of Human-Computer Studies, 145, 102536. https://doi.org/10.1016/j.ijhcs.2020.102536
	1st author. Journal article.
C7	Veitch, E., Dybvik, H., Steinert, M., & Alsos, O. A. (2022). Collaborative Work with Highly Automated Marine Navigation Systems. Computer Supported Cooperative Work (CSCW). https://doi.org/10.1007/s10606-022-09450-7

	Journal article
C8	Dybvik, H., Veitch, E., & Steinert, M. (2020). EXPLORING CHALLENGES WITH DESIGNING AND DEVELOPING SHORE CONTROL CENTERS (SCC) FOR AUTONOMOUS SHIPS. Proceedings of the Design Society: DESIGN Conference, 1, 847–856. https://doi.org/10/ggz7zb
	1st author. Conference article.
C9	Dybvik, H., & Steinert, M. (Under review). Operationalized hypotheses build bridges between qualitative and quantitative design research. Journal of Mixed Methods Research.
	1st author. Journal article.
C10	Hatlem, L. A., Chen, J., Dybvik, H., & Steinert, M. (2020). A Modular Research Platform – Proof-of-Concept of a Flexible Experiment Setup Developed for Rapid Testing of Simulators, UIs and Human Physiology Sensors. Procedia CIRP, 91, 407–414. https://doi.org/10.1016/j.procir.2020.02.193
	Conference article.
C11	Erichsen, C. K., Dybvik, H., & Steinert, M. (2020). Integration of low-cost, dry-comb EEG-electrodes with a standard electrode cap for multimodal signal acquisition during human experiments. DS 101: Proceedings of NordDesign 2020, Lyngby, Denmark, 12th - 14th August 2020, 1–12. https://doi.org/10.35199/NORDDESIGN2020.19
	Conference article.
Chapter 3: fNIRS	
C12	Dybvik, H., & Steinert, M. (2021). Real-World fNIRS Brain Activity Measurements during Ashtanga Vinyasa Yoga. Brain Sciences, 11(6), 742. https://doi.org/10.3390/brainsci11060742
	1st author. Journal article.
C13	Dybvik, H., Kuster Erichsen, C., Steinert, M. (2021) 'Description of a Wearable Electroencephalography + Functional Near-Infrared Spectroscopy (EEG+fNIRS) for In-situ Experiments on Design Cognition', in Proceedings of the International Conference on Engineering Design (ICED21), Gothenburg, Sweden, 16-20 August 2021. https://doi.org/10.1017/pds.2021.94
	1st author. Conference article.
C14	Dybvik, H., Erichsen, C. K., Steinert, M. (2021) 'Demonstrating the Feasibility of Multimodal Neuroimaging Data Capture with a Wearable Electoencephalography + Functional Near-Infrared Spectroscopy (EEG+fNIRS) in-situ', in Proceedings of the International Conference on Engineering Design (ICED21), Gothenburg, Sweden, 16-20 August 2021. https://doi.org/10.1017/pds.2021.90
	1st author. Conference article.
C15	Dybvik, H., Erichsen, C. K., Steinert, M. (Manuscript) Tetris' effect on cognitive load, performance, and systemic neurophysiology. To be submitted to Frontiers in Human Neuroscience.
	1st author. Journal article.
C16	Aalto, P., Dybvik, H., & Steinert, M. (Under review). A stroll through a cathedral: FNIRS and space sequences in architecture. Frontiers in Neuroergonomics.
	Journal abstract.

1.4.2 Publicly available datasets

Table 2 Contributions: Publicly available data sets

Data set	Citation
1	**The Ashtanga Vinyasa Yoga Data Set**
	Dybvik, H. (2023, February 23). The Ashtanga Vinyasa Yoga Data Set. https://doi.org/10.17605/OSF.IO/F8SQ3
	The Design Ideation Data Set

2	Abelson, Filip Gornitzka; Dybvik, Henrikke; Steinert, Martin, (2021) "Dataset for Design Ideation Study", https://doi.org/10.18710/PZQC4A, DataverseNO, V1, UNF:6:1NVkv2+s87SAf1qR8RRwvg== [fileUNF]
	Abelson, F. G., Dybvik, H., & Steinert, M. (2021). Raw Data for Design Ideation Study. Norges teknisk-naturvitenskapelige universitet. https://doi.org/10.21400/7KQ02WJL
3	**The Palpation Data Set**
	Rørvik, Sigurd Bjarne; Auflem, Marius; Dybvik, Henrikke; Steinert, Martin, (2021), "Replication Data for: Perception by Palpation: Development and Testing of a Haptic Ferrogranular Jamming Surface", https://doi.org/10.18710/OCMXVP, DataverseNO, V1
4	**The Tetris Data Set**
	[Will be published along with manuscript C15]

> “The scientific method, as far as it is a method, is nothing more than doing one’s damnedest with one’s mind, no holds barred”
>
> Percy Williams Bridgman[4]

[4] In *Reflections of a Physicist*, 1955, pg. 535 New York: Philosophical Library.

2 Towards, and foundations for, in-situ multimodal experiments in design research

This chapter first explicates why multimodal, in-situ experiments are important for design research. Thereafter, we provide the foundation for experiments in design research by outlining the constituting components. Lastly, we highlight the limitations of present-day experimental design research.

2.1 Why are multimodal, in-situ experiments important for design research?

2.1.1 Why are experiments important in design research?

Experiments serve the fundamental goal of improving design processes or artifacts in engineering and product design, and are thus important and useful for the design field. They may be used for product design and evaluation: inform designers about desirable product characteristics, to evaluate user interfaces, processes, and interaction styles; to better understand humans' usage, experiences, perceptions, and how they process interactive technology of increasing complexity. Experiments are needed to truly understand the depth and underlying mechanisms of human cognition in interaction with new and existing objects, tools and technical systems—i.e., whichever designed artifact (Balters & Steinert, 2017; Blessing & Chakrabarti, 2009; Borgianni & Maccioni, 2020; Cairns & Cox, 2008). Moreover, experiments may be used for design practice: to understand the designers' processes and practices, infer what the better practice may be, and change it based on experimental outcome (Cash et al., 2016; Hay et al., 2020). Experiments are one pillar in the theory-building and -testing cycle in design research, providing both the foundation for generating scientific knowledge and impactful guidance of design (Cash et al., 2016). In summary, experiments are a crucial and integral part of design at the level of both the designer and user of a(ny) designed artifact. Broadly, experiments may be divided into two based on whether they involve a) the designer or design process or b) product development and evaluation.

2.1.2 Why multimodal physiological and neurological sensors?

2.1.2.1 Why physiological and neurological sensors in design

Design practice and research is human-centered, but humans are not easily modeled due to individuality and unpredictability (Balters & Steinert, 2017; Borgianni & Maccioni, 2020). The underlying reasons for human (re)actions are difficult to subjectively distill (also for the individual human in question) (Cash et al., 2016; Goucher-Lambert & McComb, 2019), and any subjective interpretation by researchers is not easily replicable. On the other hand, using generic models of users/designers, where they respond with stable, rational behavior in a predictable manner, is simply inaccurate (Balters & Steinert, 2017; **Wulvik et al., 2019**). Design research is therefore increasingly requested to include and establish objective measures to explain behavioral patterns (Balters &

Steinert, 2017; Borgianni & Maccioni, 2020; Cash et al., 2016). This is particularly important whenever unexpected or unpredictable reactions occur and in-situ. Physiology sensors and neuroimaging techniques objectively measure certain biological changes from which it is possible to infer psychological states, e.g., mental state and human cognition (Bordens & Abbott, 2016). Physiological and neurological recordings may thus provide objective information of human cognition, better explaining human behavioral patterns—whether those are fully conscious or not (Borgianni & Maccioni, 2020). Thus, experiments in design integrate multiple physiological and neurological measures at an increasing rate, allowing better studies of the impact of design on humans (Balters & Steinert, 2017; Goucher-Lambert et al., 2019; Hu & Shepley, 2022; Steinert & Jablokow, 2013). This applies whether we are researching designers' cognition, cognitive state, or emotions/affect—or product usability and evaluation, user-product interaction, or user experience (Borgianni & Maccioni, 2020). Moreover, it is recognized that this information (including behavioral, neurological, and physiological data) can be used in creation of tools that optimally support the designer in their activities (Balters et al., 2023).

2.1.2.2 Multimodality

Physiology sensors and neuroimaging techniques are best used in combination with other measurement modalities, such as systemic measurements (physiology sensors such as electrocardiography (ECG) and galvanic skin response (GSR)), behavioral measurements (eye tracking, motion capture, video recordings), and subjective (self-report) measures (Pinti et al., 2019, 2020; Xu et al., 2019) for several reasons. Multimodality provides a more complete understanding, by providing a 360 view of neurodynamics and its coordination with other bodily changes, allowing insight at the interface of emotion and cognition (Balters et al., 2023; Pinti et al., 2020); it allows better interpretation and ultimately formulate more accurate conclusions (Pinti et al., 2019) because it enables data and method triangulation. Triangulation between multiple modalities (methods) in experiments increases rigor, increase validity of results, reduce bias and error, and accommodates individualism (i.e., individual behavior, physiology and psychology) (Balters & Steinert, 2017; Blessing & Chakrabarti, 2009; Cash et al., 2016; Steinert & Jablokow, 2013). Triangulation of results, including behavioral, neurological, and physiological data is one defined goal of design thinking research (Balters et al., 2023; Gero & Milovanovic, 2020), as it can increase understanding of the mechanisms underpinning design performance. Further, multiple modalities provide complementary information, as one sensor's advantages could compensate for another sensor's limitations (Hu & Shepley, 2022; Li et al., 2022). The combination of EEG and fNIRS is one such example, where spatially precise information from fNIRS complements the temporally precise information from EEG, while simultaneously validating recorded brain activity (Balters et al., 2023; **Erichsen et al., 2020**; Li et al., 2022). Moreover, multimodal signal analysis exhibit improved performance in signal classifications (Al-Shargie et al., 2016; Fazli et al., 2012; Lee et al., 2015) and human reactivity prediction (Cisler et al., 2019) compared to one modality only.

The human is often called the "black box" in design (Balters et al., 2023; Balters & Steinert, 2017; Cash et al., 2016; Gero & Milovanovic, 2020; Steinert & Jablokow, 2013). Physiology measures, neuroimaging techniques, together with other behavioral measures and modalities enables us to open that black box, providing an insight into human cognition (Balters & Steinert, 2017; Cash et al., 2016; Gero & Milovanovic, 2020; Steinert & Jablokow, 2013). Multimodal experiments may thus benefit design both in an early design process and at the evaluation stage (Balters & Steinert, 2017; Cairns & Cox, 2008; **Wulvik et al., 2019**).

2.1.3 Why in-situ?

To identify and understand the effect of any given task, it is best to measure it directly. Experiments conducted in the actual setting—*in-situ*—have the highest ecological validity. In-situ experiments may establish how humans design and appropriate designed artifacts, technological solutions, or interfaces in their intended contexts. High ecological validity is sought-after in design, as it relates to the extent to which results can be applied to real-world situations outside research (Cash et al., 2016). The formal, tightly controlled laboratory setting may change participants' behavior and approaches to any given task and induce Hawthorne effects, John Henry effects, observer-expectancy effects, or other experimental biases. Furthermore, research has demonstrated that tasks (or activities) when practiced in everyday lives produce different results than a mock version of the same task performed in a more constrained laboratory environment (Cairns & Cox, 2008; Okamoto et al., 2004b)—i.e., results obtained in a laboratory does not replicate nor generalize to the real world. In-situ studies accommodate real-world variability and unpredictability (Consolvo et al., 2007), making them suitable for design research (Balters & Steinert, 2017; Hay et al., 2020; Mayseless et al., 2019). As such, we are working towards conducting multimodal design research experiments completely in-situ.

2.2 What are the components of experiments in design?

> The purest implementation of the scientific method is the experiment (Cash et al., 2016, p. 50).
>
> (…) an experiment is a procedure carried out to verify, refute, or establish the validity of a hypothesis, or a set of related hypotheses. By means of a systematic manipulation of the factors determining a phenomenon, experiments provide insight into both the input factors (influences) and the output factors (implications), as well as the input–output correlations and cause-and-effect relations (Cash et al., 2016, p. 12)

As explained previously (section 2.1.1) experiments in design research may be broadly separated in two categories: those that concern a) the designer and their processes, and b) product development/design and evaluation. Despite this two-buckets categorization, there are several aspects of design that ought to be considered in experimental design research, see Figure 1. We may e.g., seek to experimentally investigate the interaction between designers and their tools or processes (e.g., sketches (Cash & Maier, 2021; Nguyen & Zeng, 2014), ideation techniques such as TRIZ and brainstorming (Shealy et al., 2020b)), several designers and their tools/process (social interaction) (Mayseless et al., 2019), or users' interaction with designed artifacts **(Dybvik et al., 2021c; Rørvik et al., 2021)**.

> Due to the logistics of experiments, it is not possible to empirically explore with human participants all possible influences on design search processes; design researchers must carefully consider the research goals they wish to explore prior to conducting a study. These limitations are an inherent part of the scientific process and also push experiments towards pragmatically investigating phenomena that are measurable since all scientific experiments must be conducted within the confines of time and resources available. It is particularly important for design researchers to differentiate between the knowledge they hope to gain from scientific studies, and the knowledge that is feasible to gain from scientific studies (Cash et al., 2016, p. 202).

Therefore, to attain a well-designed experiment, we provide in the following sections, the basis for experimental design research. We provide an overview of the most fundamental elements of experiments in design research, which are hypothesis creation, participants and sampling, types of variables and operationalization, the assignment of participants to experimental conditions (also known as design of experiments), and analysis.

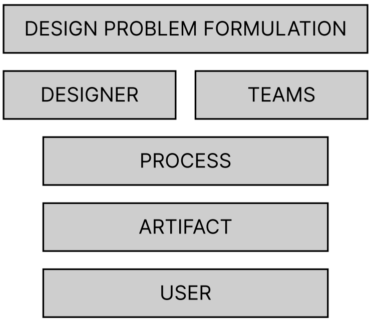


Figure 1 Systems view of design. Design could (simplistically) be considered to consist of these aspects: a designer or team of designers applies design knowledge (internal and external) and a process, to a design problem, following a suitable process to obtain design solutions [Figure based on (Cash et al., 2016, p. 14)].

2.2.1 Hypothesis

One or more hypotheses are an integral part of experiments in design. The hypothesis relates closely to the research purpose (Cash et al., 2016), expressing the proposed theory in a specific way. Specifically, hypotheses specify an expectation about how a particular phenomenon is, works, impacts, or relates to some other phenomena. It's written as a tentative, falsifiable statement that expresses the expected theory (often as a causal relationship) by means of operationalized variables in a matter that allows experimental testing (**Dybvik & Steinert, Under review**; Popper, 2002; Salkind, 2010). Design propositions may be used to create hypotheses, as hypotheses, in essence, are falsifiable propositions (**Dybvik & Steinert, Under review**; Kelle, 1997). The hypothesis may be completely new, or concern (i.e., aiming to test) existing theory (Cash et al., 2016).

Design research hypotheses are usually created in a theory-building mode (Cash, 2018). We will briefly review some of the many approaches to creating hypotheses. Hypotheses may be derived from existing theories that cannot fully explain the studied phenomenon or they may be based on prior observation of some phenomena (Cash et al., 2016). Theory-building mode most often encompass qualitative research, and qualitative approaches more generally may therefore be used to create experimental hypotheses **(Dybvik & Steinert, Under review)**. One could e.g., use a qualitative approach that encompass interviews with experts in the field and current and potential users of the designed artifact (**Dybvik et al., 2020; Veitch et al., 2022)**, and combine that with insights from field observations **(Veitch et al., 2022)**. Hypotheses may also be formulated based on experimental results, as new questions or discoveries might arise during data interpretation. If experimental results are not fully explainable with the collected data, or if participants behave in an unexpected manner during the experiment, it warrants a new hypothesis and experiment with a new investigational focus. E.g., in one experiment that involved the design of a predictive display, participants were not explicitly informed that there would be a predictive display nor how it worked. Rather, participants were given 30 seconds to intuitively learn to use the predictive display **(Dybvik et al., 2021c)**. As a result, some participants immediately understood how the predictive display worked. They typically performed better than participants who did not understand that there had been a predictive display until the experiment was over. Thus, we hypothesize that participants would have performed better if they had been properly instructed about the predictive display's functionality. This new hypothesis could be tested in a new experiment.

2.2.2 Participants and sampling

Experimental design research entails human participants. Research involving human participants require consideration of potential ethical issues, and participation must be voluntary, include informed consent, and the right to withdraw at any time (Cash et al., 2016; Kirk, 2013). Depending on the research purpose and hypotheses participants may be either designers or users of an artifact. The participants represent a sample of the greater population of interest to us, i.e., the population we seek to infer causal relationships about, or explanations of (Cash, 2018). Participants are allocated to experimental conditions, but this is discussed in 2.2.4.

The size and profile of the sample are crucial contributors to external validity and generalizability. Larger sample sizes are generally preferred because statistical power increase with sample size, and because a certain proportion of the population is required for generalizable results. Smaller "test" experiments—so-called pilot experiments—may be conducted with smaller sample sizes, as it is not practical nor physically possible to explore all possible influences on the design process. Pilot experiments are thus a good way to explore effects before conducting a full-scale study. A modular, flexible experiment setup may be helpful in piloting **(Hatlem et al., 2020)**.

To attain population validity, participants should be an unbiased, representative sample of the population, ideally obtained through random sampling or stratified random sampling (among other sampling techniques). However, in practice, researchers are often constrained in one or more ways to prevent truly random samples. Convenience and practicality often trump genuine random sampling. Different biases may influence whether participants are a true representation of the population sought to test. Students are for example often used in experiments because of general availability (Druckman & Kam, 2011). This issue is debated (Cash et al., 2022; Druckman & Kam, 2011) and important also for design research (Cash et al., 2022). Student participants might work well for e.g., educational research purposes, but might not generalize to design practitioners working in industry.

For example, in one experiment **(Rørvik et al., 2021)** we recruited healthy engineering students to test a prototype of a palpation training device under development. The device's intended purpose was for medical students to learn palpation skills or for medical personnel to maintain said skills. In this context, we considered it appropriate to recruit healthy students untrained in the task as the population of users of the final concept would include untrained students, and this was the first evaluation of whether the concept prototype would work as a training device. However, engineering students are not a representative sample of medical students, rather, they were selected due to availability. At the time, the alternative would likely be no testing of the concept. We also recruited students in another experiment **(Dybvik et al., 2018; Wulvik et al., 2019)** that investigated mental state and physiology in different scenarios related to large ship navigation. Our goal was to inform designers of how they might compare interface design for remote ship operation based on their influence on mental state and physiology. Students were unfamiliar with the task which was to operate a large ship in a simulator. Thus, we do not know how well the results would relate to trained professionals participating in the same experiment. We expected directionality of responses to be the same, although the magnitude would differ, but a new experiment would be needed to test this.

2.2.3 Variables

An experiment is characterized by researchers manipulating one or more independent variables (i.e., the cause), controlling possible confounding or nuisance variables, before observing or measuring one or more dependent variables (i.e., the effect) (Cash et al., 2016; Kirk, 2013). The variables are obviously dependent on the research question, the aim, and the hypothesis. In design research:

> Variables may relate to designers (skill levels, creativity), artifacts (complexity, adaptability, modularity), methods (efficiency, effectiveness), design teams (composite personality, skill profile), design ideas, and so on (Cash et al., 2016, p. 17).

In design, we are more often than not studying complex constructs that are not easily measurable (Cash et al., 2016). The construct we want to measure might not have a known objective, physical property that can be counted, or an established sensor that measures it. If so, we must operationalize the variable, i.e., temporarily define which variables to use to measure the construct in place of the "true" variable (Cash et al., 2016; **Dybvik & Steinert, Under review**). Operationalization is relevant for both dependent and independent variables. Fortunately, many well-researched constructs in design have established operational definitions and measures (Cash et al., 2016; Gero & Milovanovic, 2020; Surma-aho & Hölttä-Otto, 2022). In absence of established measures, one may create a new proxy **(Dybvik & Steinert, Under review)**.

2.2.3.1 Independent variables

Generally, the independent variable(s) are any suspected causal event that is under investigation (Kirk, 2013): the cause of the effect. Independent variables are also known as stimuli, treatment, predictors, explanatory variable, regressor, co-variate, factor, and feature. If the independent variable consists of several levels (i.e., versions), a small pilot experiment may be carried out to identify the most beneficial levels and determine the number of required participants (Kirk, 2013).

If for example, our hypothesis concerns the design process and how different tools/aids/techniques affect concept generation (Cash et al., 2016; Gonçalves & Cash, 2021; Goucher-Lambert et al., 2019; Shealy et al., 2020b), the independent variable would naturally be one or more tools, or several levels (i.e., degrees) of the tool. In one study **(Dybvik et al., 2022, 2023)** we investigated whether inspirational word stimuli would aid design concept generation. The independent variable was operationalized as five worlds presented on a screen. Different words were presented in three different conditions, two of which were inspirational stimuli at two different degrees/levels (near and far from the design problem space). These were compared to a control condition that had no inspirational stimuli, but instead reused words from the problem statement **(Dybvik et al., 2022, 2023)**.

2.2.3.2 Dependent variables

Generally, the dependent variable(s) are the measurement(s) used to assess the effects, if any, of the independent variable (Kirk, 2013). Dependent variables are also known as outcome, regressand, response, and target. The choice of dependent variables may be theory-based, although in practice it is often also based on practical considerations and availability. Sensitivity, reliability, distribution, and practicality should be considered, as behavioral research—which design research is—involves a sizable investment in time and material resources. Thus, the dependent variable should be reliable and maximally sensitive to the phenomenon under investigation (Kirk, 2013).

Several dependent variables were used to measure the effect in the design ideation experiment **(Dybvik et al., 2022, 2023)**. The number of generated ideas, and the novelty and quality of those ideas assessed whether inspirational stimuli had an effect on design ideation fluency (how many concepts were generated) and the perceived nature of those concepts. The relevancy and usefulness of inspirational stimuli was also subjectively evaluated to assess whether participants perceived the stimuli differently. Finally, we used two behavioral measures: eye-tracking to investigate how the inspirational stimuli affect visual allocation and gaze patterns, a think-aloud protocol to record what the generated ideas were.

In the experiment evaluating the palpation training device we collected several dependent variables, including; participants' subjective evaluation of hardness, as it was important for the concept to be able to create different hardnesses; a drawing of the identified irregularity, to determine whether different shapes could be accurately determined by participants; and we asked whether participants became better at locating the irregularity during the experiment, to assess the device's educational potential **(Rørvik et al., 2021)**.

2.2.4 Experimental design

> Considering the environment of conducting experiments, we can talk about (i) field experiments, (ii) laboratory experiments, and (iii) mixed-placed experiments (Cash et al., 2016, p. 210).

Experimental design—or design of experiments—is a plan for assigning participants to experimental conditions and the associated statistical analysis (Kirk, 2013). Analysis will be considered in section 2.2.5. There are many types of experimental designs in which participants may be assigned to different conditions, groups, treatments, etc., depending on the research question and hypotheses. A specific, planned assignment of participants is necessary to control for potentially confounding, or nuisance, variables, limiting bias, ensuring that the intended hypothesis is tested, and the research question answered. All considerations in design of experiments are essentially concerned with identifying the effect of the independent variable while minimizing the effect of (i.e., controlling for) confounds and nuisance variables.

There are four general approaches to control for nuisance variables. The first three are experimental ways of control and are: 1) hold nuisance variables constant for all participants, often called repeated-measures (or within-subjects) design where all participants are exposed to the same experimental conditions; 2) random assignment of participants to conditions, splitting participants into two or more groups, often called independent-measures (or between-subjects) design where different participants are exposed to different conditions; 3) include the nuisance variable as a factor in the experimental design. Some combination of the three approaches may also be used. The fourth is an analytical approach to control, involving 4) statistical control, i.e., by using regression techniques to statistically remove the effects of the nuisance variable.

Relatedly, important concepts for any experiment design, when deciding how to control for nuisance variables, are randomization (assigning participants to conditions/groups at random), counterbalancing (systematic variation of order of conditions or experimental units, and blocking. Blocking is the non-random arrangement of experimental units (e.g., conditions, groups, participants) into groups (blocks) that are similar to one another (Kirk, 2013). A blocking factor in design research could be whether participants are students or practitioners in industry. A generalization of blocking includes Latin Square Design, which, if e.g., given three conditions, these are ordered into three groups such that each group experience each condition only once, but in a different order. Participants are then randomly assigned to the three groups, and usually also counterbalanced so that each group has approximately the same number of participants. All three concepts limit bias by avoiding order and sequence effects, thereby enhancing internal validity.

A control condition is usually required for valid inference making of whether an effect exists. A control condition is a condition (or group/experimental unit) that does not involve exposure to the cause (*APA Dictionary of Psychology*, n.d.), to which all experimental conditions are usually compared (Popper, 2002). Acquiring a baseline, i.e., a measurement prior to exposure of experimental conditions is one conventional way to include a control condition.

The design ideation experiment featured repeated measures design **(Dybvik et al., 2022, 2023)** with three conditions (control, near-inspiration, far-inspiration) as we were interested in determining existence of effect (whether there was an effect of inspirational stimuli, warranting the control condition) and investigating the dose-response relationship (whether different exposures of inspirational stimuli yields differentiating effects and investigating the most beneficial level of exposure, warranting near- and far conditions). The ship operation experiment used only two conditions **(Dybvik et al., 2018; Wulvik et al., 2019)** as we were interested in potential differences in workload and stress[5] between two regularly occurring ship navigation scenarios, but not concerned with establishing existence of workload and stress (thus, no control condition). The predictive display experiment **(Dybvik et al., 2021c)** had three conditions whose sequence was determined according to a 3x3 Latin Square Design to avoid potential order or learning effects.

Despite excellent experimental design that limits as much bias as possible, real studies with human participants will ultimately have limitations. E.g., the design ideation study was limited by the central fixation bias[6] **(Dybvik et al., 2023)**. This might have been avoided or reduced by randomizing the position of the stimuli and problem statement, but this was not possible as we employed existing stimuli, and as replication was a secondary aim of the study. Trade-offs, such as this one, are inherent considerations to design of experiments.

[5] as measured physiologically and subjectively

[6] The "marked tendency to fixate the center of the screen when viewing scenes on computer monitors" (Tatler, 2007).

2.2.5 Analysis

The consideration and selection of analysis approach should be conducted jointly with the experimental design and the hypothesis. Fundamentally, the analysis type must be appropriate for answering the hypothesis and the type of research questions that is asked, as it is vital to apply appropriate statistical and mathematical rules.

There are several types of statistics. Descriptive statistics describe the data in a manner that provides information about what the collected data looks like, such as its distribution, range, most frequently occurring values, described through statistical mean, standard deviation, range, median, etc. Descriptive statistics are useful for gaining a first impression of the data and represent one outcome of the experiment. Inferential statistics (i.e., classical statistics) seek to investigate and potentially establish (infer) relationships between variables by statistical testing of hypotheses, and generalize beyond the realms of the experiment. Inferential statistics establish differences, associations, and relationships (i.e., correlations) between two or more variables (Cash et al., 2016). Within statistical inferential techniques there are a range of possible approaches, thus it is impractical and not appropriate to cover all within the scope of this thesis. Instead, we provide a brief overview. Different statistical approaches are appropriate depending on the number of dependent variables. Strictly speaking, univariate techniques are appropriate in cases with one dependent variable, multivariate techniques in cases with more than one dependent variable. Nevertheless, mass univariate statistical testing—i.e., collecting multiple dependent variables, but analyzing each one with univariate tests as if it was the only variable—is omnipresent. There is a multitude of statistical tests (t-tests, ANOVA variations, etc.), however, most may be considered special cases of regression. We refer to Field (2018) and Kirk (2013) for comprehensive overviews.

Computational approaches use more advanced mathematical models than the regression-based statistics described above to inform conclusions of relationships or patterns in the data, including e.g., machine learning models. Computational tools in design research include; latent semantic analysis, which is based on singular value decomposition from linear algebra (Cash et al., 2016; Dong, 2005; Dong et al., 2013); simulations, which are modeling strategies of e.g., artifact behavior with various inputs with real or generated data, or e.g., social simulations that inductively study dynamic interactions in creative teams such as an agent-based simulation of the process and performance of engineering teams (McComb et al., 2017b); and (Hidden) Markov Models, which are probabilistic, statistical models of a sequence (chain) of events. They have been used to compare design activities across domains (Cash et al., 2016); e.g., modeling the design process, describing the probability of one event leading to another (McComb et al., 2017a, 2017b); and e.g., inferring cognitive states (patterns of brain activity) during design ideation (Goucher-Lambert & McComb, 2019).

In the design ideation experiment we used one-way repeated measures ANOVA **(Dybvik et al., 2023, 2022)** to determine whether there were differences in ideation fluency, while a Friedman's test investigated differences in subjective ratings and fixation distribution between inspirational and non-inspirational stimuli. Their respective post-hoc tests included pairwise comparisons with a Bonferroni correction and a Wilcoxon signed-rank test with a Bonferroni correction. Standard null-hypothesis significance testing was accompanied by effect sizes. As an exploratory analysis we generated individual scan-paths, inspected them visually and qualitatively assessed whether they might suggest individual search patterns and aid hypothesis generation. In another experiment **(Wulvik**

et al., 2019) we used correlation tests and multivariate analysis (i.e., principal component analysis (PCA) and partial least-squares regression (PLSR)) to investigate relationship between different physiology variables (derived from ECG and EDA), and subjective variables. This revealed which EDA and ECG variables had the strongest correlation to subjective workload and stress, and we were able to show how the variables change depending on the experimental condition. We also used PCA in another experiment **(Dybvik et al., 2021c)** for pattern exploration, revealing correlations between most of the subjective NASA-TLX workload measures—except for negative correlation between subjectively evaluated performance and the other subjective workload measures.

2.3 What are the current limitations?

Black box experiments are the go-to in design research, often including e.g., think-aloud protocols or behavioral measures. Think-aloud protocols have inherent limitations. In retrospective think-aloud protocols, participants may be unable to remember everything, while a think-aloud protocol conducted during the experiment could influence the process itself. Though providing a record of participants' actions, behavioral measures are filtered by participants themselves, not necessarily reflecting the internal processes prior to participants exhibiting the behavior. In both cases, researchers must subjectively infer the brain's internal workings (Cash et al., 2016; Goucher-Lambert & McComb, 2019).

EEG and fMRI studies measure brain activity directly, providing objective information about underlying cognitive processes. The drawback is that EEG and fMRI are highly susceptible to motion artifacts, which require posing restraints on participant movements and conducting the experiment in a laboratory environment—accompanied by experimental biases. As a result, neither modality can be used in-situ (Balters et al., 2023; Gero & Milovanovic, 2020). EEG and fMRI are on opposite sides of the spectra regarding spatial and temporal resolution, but there are no multimodal studies (Balters et al., 2023).

Design studies with high ecological validity or in-situ would require relatively good temporal and spatial determination of brain activity—we need to know the where and when of brain activity—but no established "existing" method allows technically that. This is crucial for design experiments that, e.g., investigate industry design processes, infrastructure and architecture's effect on design (think tanks, "creative" spaces), and product evaluation of interfaces. We expect that results from such studies would differ depending on whether they were conducted in-situ or in a laboratory—or even that it would be too difficult to sufficiently recreate certain environments in a laboratory, such that the study would not be conducted at all.

fNIRS covers the limitations outlined here. fNIRS provides a direct, objective measure of brain activity that avoids participants' self-filtering. Because it is robust to motion artifacts, fNIRS can be used in highly ecologically valid settings and in-situ. fNIRS offers a great spatial and temporal resolution tradeoff and allows multimodal integration. Chapter 3 further describes fNIRS, its advantages, and its suitability for design research.

"If you torture the data long enough, it will confess".

Ronald H. Coase[7]

[7] Attributed to a talk Coase gave at the University of Virginia in the early 1960s. Paraphrased versions of this saying later appeared in various prints.

3 fNIRS

This chapter introduces the technical principles of fNIRS as well as its advantages and limitations. Considerations when designing experiments with fNIRS, fNIRS analysis, and examples of in-situ experiments with fNIRS are thereafter provided. Lastly, fNIRS in design research is discussed.

3.1 The technical principles of fNIRS

Functional near-infrared spectroscopy—fNIRS—is an optical neuroimaging technique that allows investigation of cortical hemodynamic activity (i.e., superficial brain activity), by using near-infrared light. The word *functional* reflects that we are interested in obtaining measurements of the changes in hemodynamic activity associated with specific tasks (or functions). fNIRS is based on neurovascular coupling theory and optical spectroscopy, by which it quantifies cerebral tissue oxygenation (Phillips et al., 2016).

3.1.1 Neurovascular coupling

Neurovascular coupling is the mechanism that relates hemodynamic activity to neural activity.

> Neurovascular coupling reflects the close temporal and regional linkage between neural activity and cerebral blood flow (Phillips et al., 2016).

Brain activity (i.e., neural activity) is highly energy-demanding. Energy is consumed within the brain by oxidative metabolism of glucose, i.e., consummation of oxygen and glucose. On average, the brain consumes six oxygen molecules per glucose molecule. Oxygen must be delivered to the brain via the blood stream, through the transport molecule hemoglobin[8], because oxygen has low solubility in water. Hemoglobin may be saturated with oxygen molecules (oxygenated hemoglobin) or desaturated with oxygen molecules (deoxygenated hemoglobin). Oxygen is transported to brain tissue via diffusion. Increases in neural activity are directly followed by increased oxygen demand, because brain tissue cannot store any meaningful amount of oxygen. Increases in neural activation are quickly followed by the physiological blood flow response. Cerebral blood flow (CBF) increases within a few seconds after neural activation onset, followed by a rapid decline after stimulus offset. The increase in CFB is paralleled by changes in oxygenation and blood volume[9,10], and an increase in cerebral metabolic rate of oxygen ($CMRO_2$). CBF increases much more than $CMRO_2$[11]: ratios of relative CBF increases to relative $CMRO_2$ increases are estimated to be around 2-4 during a wide range of neuronal activity. This discrepancy between large CBF and small $CMRO_2$ leads to hyperoxygenation

[8] In fact, the blood stream's oxygen content can increase from 150 nmoL/mL to 9,000 nmoL/mL.

[9] Waste products of brain energy metabolism (lactate, CO_2, or heat) have a role in neurovascular coupling, but experimental data and calculations suggest that the CBF does not primarily serve to remove these waste products (Leithner & Royl, 2014) (CBF response does not depend on deoxygenation of hemoglobin).

[10] CBF may relevantly increase the oxygen gradient from capillary to tissue, and thus oxygen availability to the brain.

[11] It's not known exactly why there is this discrepancy between CBF and $CMRO_2$ responses, but several hypotheses have been proposed.

(an oversupply of oxygenation which necessarily comes in the form of oxygenated hemoglobin) of the activated brain region and a decrease in deoxygenated hemoglobin. This is called the hemodynamic response and it provides the basis for blood-oxygen-level-dependent fMRI (Leithner & Royl, 2014) and oxy- and deoxy-hemoglobin concentrations in fNIRS.

To summarize, an increase in neural activity increases the neuronal tissue's energetic demand, which causes an increase in oxygen metabolism and CBF. The intensified CBF triggered by neural activity increases blood flow to active brain regions (Phillips et al., 2016), in which the local oxygen supply becomes greater than oxygen consumption. Therefore, within active brain regions, we expect higher concentrations of oxygenated hemoglobin (and total hemoglobin) and simultaneously decreased concentrations of deoxygenated hemoglobin (Herold et al., 2018; Zohdi et al., 2021).

3.1.2 Optical spectroscopy of biological tissue

Brain activity is associated with a number of physiological events, some of which are associated with changes in the optical properties of brain tissue (which is a consequence of neurovascular coupling). Thus, these changes can be assessed with optical techniques. (Ferrari & Quaresima, 2012).

Optical spectroscopy studies the interaction of light with matter and may be used to measure certain characteristics of molecular structures. fNIRS exploits the principles of near-infrared spectroscopy (Ferrari & Quaresima, 2012), meaning that light in the near-infrared (NIR) spectrum (650–900/1000 nm) is used (Ferrari & Quaresima, 2012; Quaresima & Ferrari, 2019; Santosa et al., 2020). Biological tissue (i.e., human tissue) exhibits relative transparency to light in the NIR spectra (Phillips et al., 2016). When NIR light passes through biological tissue, light particles (photons) may either be absorbed (photon hits molecule, is destroyed, transferring energy) or scattered (photon movements may include refraction, reflection, or diffraction). Scattering is about 100 times more probable than absorption in human tissue (Quaresima & Ferrari, 2019), thus allowing NIR light to penetrate several centimeters of human tissue along a diffuse path (Santosa et al., 2017). The dominant *absorbers* of NIR light are oxygenated and deoxygenated hemoglobin. The absorption properties of oxygenated and deoxygenated hemoglobin molecules differ greatly in NIR spectrum (Figure 2). For example, oxygenated hemoglobin (HbO_2) has lower absorption of 760 nm wavelength light than deoxygenated hemoglobin (HbR), while at 850 it has higher absorption.

NIR light of different wavelengths are emitted from a light source placed on the scalp. Light penetrates the outer layers of the head (scalp, temporal muscle, skull, frontal sinus, cerebrospinal fluid, and dura (Quaresima & Ferrari, 2019)) and into the first few centimeters of cortical neuronal tissue. The NIR light's optical properties change inside the neuronal tissue due to HbO_2 and HbR's different absorption properties; this change is recorded by a light detector, also placed on the scalp. The recorded time course of (raw) light voltages may be converted to optical density, which are then used to estimate oxygenated and deoxygenated hemoglobin concentrations by means of the modified Beer-Lambert Law (Delpy et al., 1988). In brief, the Beer-Lambert Law models the way light is transported through a homogenous, non-scattering tissue. When light scatters, the distance traveled by photons changes (it becomes more random), it takes longer time to exit the tissue, increasing the probability of absorption, and introducing a certain amount of loss. The modified Beer-lambert Law accounts for this by changing the optical pathlength, accounting for the increased pathlength, and adding a term that accounts for light loss. It's conventional to assume constant scattering, which means the pathlength and light loss can be considered constant. Several works have estimated and tabulated pathlength and light loss constants for different ages, tissues, and wavelengths (Gemignani, 2019).

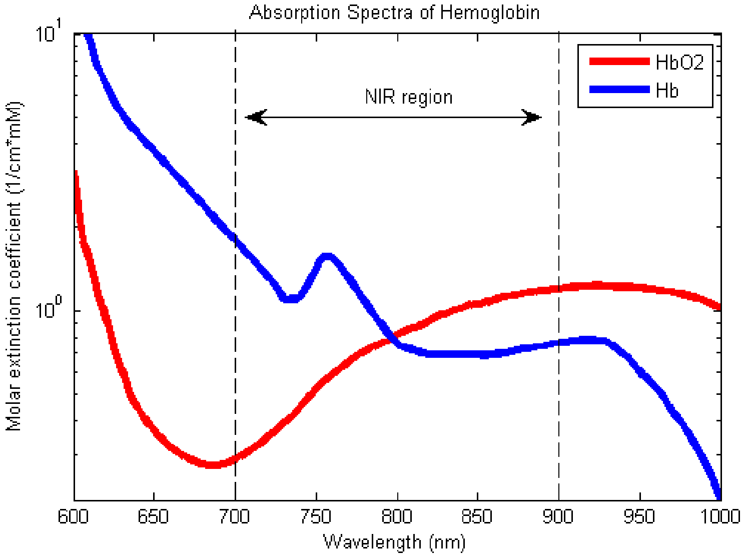


Figure 2 Absorption spectra of oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb) for near-infrared wavelengths. [12]

3.1.3 fNIRS hardware and terminology

An fNIRS system is composed of light sources and detectors, collectively named optodes. As described above, NIR-light leaves the sources, placed perpendicularly on the scalp, travels in all directions, photons are scattered and absorbed by hemoglobin, before some photons reach the detectors. The average photon path reaching a detector is "banana-shaped," and hence it is often referred to as the "photon banana" (see Figure 3). The depth of the measured tissue depends on the optode spacing. Greater spacing results in greater measurement depths. A 3 cm source-detector distance is usual for adults (2 cm for infants), enabling approximately a 1.5 cm measurement depth. Each source-detector pair forms a "channel." A channel is defined as the mid-point between a source-detector pair[13].

[12] From Curtin, Adrian. (2012). *Absorption spectra of oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb) for Near-infrared wavelengths (NIR)* [Figure] Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Oxy_and_Deoxy_Hemoglobin_Near-Infrared_absorption_spectra.png). In the public domain.

[13] fNIRS-channels are thus unlike EEG-channels, which are the number of actual electrodes.

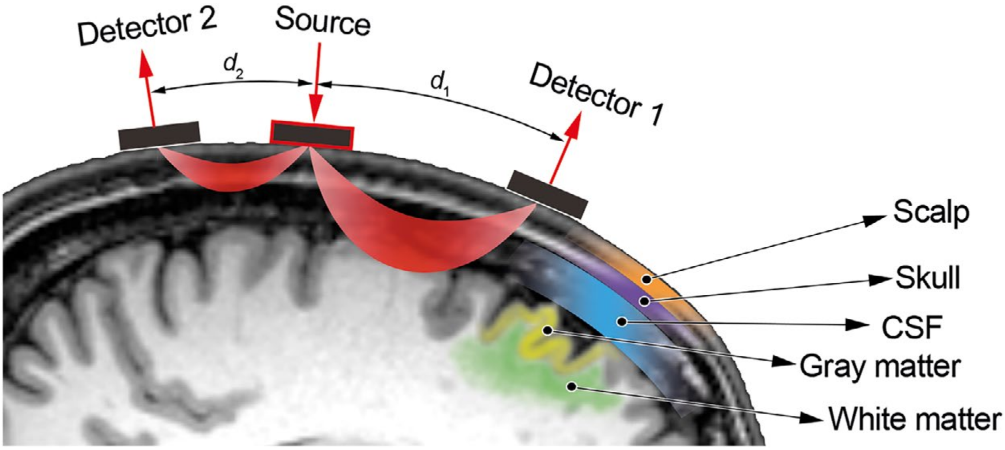


Figure 3 The NIR light's photon banana path is illustrated in red. The light's penetration depth is proportional to the source–detector distance (about one half of the source-detector distance, d$_1$: deeper (long) channel; d$_2$: superficial (short) channel). [Figure reproduced from Pinti et al. (2020)].

Naturally, the number of channels depends on the number of sources and detectors of the fNIRS system, but also on the location of the sources and detectors. Optodes are arranged spatially into a montage (or probe), and optodes may be arranged differently to increase spatial resolution or differentiate depths. ***Short-channels*** take advantage of this; they are shorter distance source-detector measurements (ideally 8 mm separation in adults, 2 mm for infants), yielding shorter depth measurements that obtain a measure of scalp blood flow. Scalp blood flow is a part of the regular channels[14] measurements, i.e., contaminating the signal, potentially significantly impacting data and resulting conclusions. By obtaining short-channel measurements of the scalp blood flow component close to the long channels (Gagnon et al., 2012a), it could be accounted for during analysis, e.g., by regressing out this component. More than one short-channel accounts for heterogenous scalp blood flow dynamics (Santosa et al., 2020; Wyser et al., 2020). fNIRS studies are encouraged to collect short-channel data whenever possible (Wyser et al., 2020).

3.1.4 Different fNIRS techniques

Based on different types of illumination, there are three different types of fNIRS techniques. 1) Continuous wavelength (CW) fNIRS continuously illuminates tissue with a constant amplitude; thus, measuring the light's overall attenuation while traveling through tissue. CW-NIRS cannot differentiate scattering and absorption, therefore only quantifying concentration changes relative to a baseline. 2) Frequency-domain (FD) fNIRS illuminates tissue with intensity-modulated light (in radio frequency range ~100MHz), which allows it to measure light attenuation, phase shift of emergent light, and the decay of modulation depth (ratio of AC to DC component). The two latter are affected differently by absorption and scattering, and thus it is possible to quantify absolute concentration levels with FD-NIRS. 3) Time-domain (TD) fNIRS illuminates tissue with short light pulses (picoseconds), detecting the light pulse's shape as it exits tissue. The properties of the recorded photon distribution (area under curve, time of maximum, time of width) allow assessment of the tissue's absorption and scattering, and

[14] Regular distance channels are sometimes referred to as long channels.

thus quantification of absolute concentration levels (Ferrari & Quaresima, 2012; Gemignani, 2019; Quaresima & Ferrari, 2019).

The most common of the three, and the focus of this thesis, is CW-NIRS, which measures relative changes in oxygenated and deoxygenated hemoglobin. FD-and TD-NIRS systems can measure absolute HbO_2 and HbR concentrations, but increase in cost and technical complexity, compared to CW-NIRS, which is lower cost and easy to transport as it can be made portable (Ferrari & Quaresima, 2012; Quaresima & Ferrari, 2019). Most fNIRS research uses CW systems as absolute values are not critical for most neuroscientific applications (Gemignani, 2019).

3.1.5 History and applications

fNIRS is a relatively young neuroimaging technique. Therefore, it is worth noting that many of the considerations for experimental design, signal processing, and data analysis in fNIRS research stem from the fMRI and EEG communities, because many early fNIRS researchers came from these fields. Optical spectroscopy of biological tissue would not be possible without the field of physics, nor would the sophisticated signal processing techniques used today. The excerpt below illustrates these points:

> In general, time-series models (correlation or Granger causality), which examine the relatedness of two slow hemodynamic signals over time, have historically been more popular in the sFC analysis of fMRI data [41,42]. By contrast, the FD metrics of coherence and phase locking have been more widely used in electrophysiological recordings, such as electroencephalography and magnetoencephalography, where the frequency-specific values can provide insight into connections at specific neural oscillatory bands (e.g., the so-called alpha or beta rhythms)[7]. In fNIRS research, which has a higher acquisition rate than fMRI but still measures the slow hemodynamic signals, both TD (e.g., Refs. [43–47]) and FD (e.g., Refs. [22, 48–50]) approaches have been previously used (Santosa et al., 2017).

fNIRS is applicable to a wide range of research areas, including, but not limited to: BCI and neurofeedback, cognitive and developmental disorders, movement, exercise and balance, infant monitoring, speech and language development, stroke and rehabilitation, traumatic brain injury, visual impairment and stimulation, social interaction and hyperscanning (multiple participants) (Davies et al., 2015; Herold et al., 2018; *Published fNIRS Research with NIRx Systems*, n.d.; Quaresima & Ferrari, 2019; Rupawala et al., 2018).

Hyperscanning is the simultaneous functional neuroimaging of two or more brains. Hyperscanning allows investigation of interbrain neuronal synchrony, which only occurs in social interactions (Babiloni & Astolfi, 2014; Quaresima & Ferrari, 2019). This is nearly impossible in fMRI, has been done with EEG, but fNIRS hyperscanning allows for more natural contexts and a wider array of interactions as it is more robust to motion artifacts (Babiloni & Astolfi, 2014; Quaresima & Ferrari, 2019), and possess other advantages (described in section 3.2.1). Hyperscanning has investigated cooperation (Balters et al., 2022; Cui et al., 2012; Li et al., 2021; Mayseless et al., 2019), various games (Babiloni & Astolfi, 2014; Liu et al., 2016), student-teacher interactions, leader-follower dynamics, and more (Quaresima & Ferrari, 2019).

3.2 Advantages and limitations

This section outlines advantages and limitations inherent to fNIRS and compares it with other neuroimaging modalities. Thereafter, we discuss practical challenges associated with collecting fNIRS data, the different sources of noise in fNIRS data, and discuss fNIRS in combination with other sensor measurements.

3.2.1 Advantages

fNIRS is safe, noninvasive, and silent (Pinti et al., 2020; Quaresima & Ferrari, 2019; Scholkmann & Vollenweider, 2022). The NIR light does not induce any sensorial sensations in participants (Quaresima & Ferrari, 2019). fNIRS has a relatively high temporal resolution (sampling rates up to 100 Hz exist, but are typically between 1 – 10 Hz), and a relatively good spatial resolution (approximately 1-3/4 cm). The technical principle is often considered a good compromise between spatial and temporal resolution. Starting costs and maintenance costs are relatively low, and it is relatively quick to set up a system. fNIRS systems are portable as they are relatively lightweight and compact, and can further be made wearable and wireless (Quaresima & Ferrari, 2019).

fNIRS is relatively robust to motion artifacts, it is e.g., much less affected by sweat and muscle activity than EEG (Al-Shargie et al., 2016; Balters & Steinert, 2017; Pinti et al., 2020). Thus, fNIRS may be used during movement, enabling a wide range of different stimuli, tasks, and environments. In addition, lightweight and wearable fNIRS systems may be transported to participants in the field, rendering more realistic, real-world environments and tasks possible. Specifically, fNIRS enables in-situ studies.

fNIRS can be used on difficult (i.e., low compliance) participants of all ages (including preterm newborns), due to the technical principle, and being more comfortable than alternatives (fMRI/EEG), rendering possible a greater range of participant populations. Moreover, other electrical or magnetic monitoring systems may be used at the same time as fNIRS (e.g., pacemaker, hearing aid, cochlear implants) (Pinti et al., 2020; Quaresima & Ferrari, 2019). fNIRS may record multiple participants simultaneously, i.e., perform hyperscanning, and thus be used in social interaction studies (Quaresima & Ferrari, 2019). Continuous measurements over longer time periods (several hours) are allowed (Pinti et al., 2020; Quaresima & Ferrari, 2019; Scholkmann & Vollenweider, 2022). fNIRS is suited for multimodal integration with a range of modalities (Al-Shargie et al., 2016; Pinti et al., 2020; Scholkmann & Vollenweider, 2022; Solovey et al., 2009), including EEG, fMRI, transcranial magnetic stimulation, transcranial direct current stimulation, electrooculogram, electromyography, ECG, eye-tracking, thermal cameras, accelerometer, etc.

3.2.2 Limitations

fNIRS does not obtain anatomical information about brain structure (Pinti et al., 2020; Quaresima & Ferrari, 2019). The disadvantage is that it may be difficult to identify which cerebral regions were active, complicating meta-analyses of results between different studies and techniques. Establishing exact spatial origin of cortical hemodynamic response requires individual three-dimensional MRI. Although this is the best solution thus far, in terms of precision, it is expensive both in terms of cost and time. Other means of linking brain anatomy to fNIRS optode locations are discussed in section 3.4.1.2.

The penetration depth of NIR-light depends on many factors: source-detector separation, source power, detector sensitivity, optical properties of the skin/skull layers, degree of white matter myelination, etc. (Quaresima & Ferrari, 2019). Typical depth sensitivity is around 1.5 – 2 cm for most CW-fNIRS systems (Pinti et al., 2020; Quaresima & Ferrari, 2019), which limits investigations to cortical hemodynamics. Thus, deeper lying brain regions (e.g., the amygdala) cannot be investigated (Herold et al., 2018; Quaresima & Ferrari, 2019; Yücel et al., 2021).

A ~5 second delay from stimuli onset to peak response is inherent in fNIRS signals due to the nature of the hemodynamic response (Pinti et al., 2020). The fNIRS signal is comprised of several components, i.e., measures of neurological activity must be separated from systemic physiology and vascular changes (Scholkmann & Vollenweider, 2022; Tachtsidis & Scholkmann, 2016). Environmental or ambient light and electrical noise may further contaminate the data (Solovey et al., 2009). Exposure to differently colored light (e.g., red vs blue) additionally affect systemic physiology (Zohdi et al., 2021). Section 3.2.5 provides further details on noise in fNIRS data.

CW-fNIRS cannot measure optical path length, thus parameter estimations of optical path length must be used and considered when interpreting fNIRS data (Quaresima & Ferrari, 2019). There is not yet a standardized procedure for signal processing, data analysis and statistical procedures (Pinti et al., 2020; Quaresima & Ferrari, 2019; Scholkmann & Vollenweider, 2022; Yücel et al., 2021).

The cost of an fNIRS system is dependent on the number of optodes. Approximately 64 optodes are required for full-coverage of an average adult head using 3 cm source-detector separation (Scholkmann & Vollenweider, 2022). While whole-head montages allow complete investigation of cortical hemodynamics, lower resolution systems are generally more lightweight and thus easier to use in-situ, which is particularly relevant to design research.

3.2.3 Comparison and multimodal integration with other neuroimaging techniques

All neuroimaging tools, fMRI, fNIRS and EEG, have methodological advantages and disadvantages (see Table 3) that must be traded off with regards to intended research purpose (Herold et al., 2018). Compared to other neuroimaging tools, fNIRS is very tolerant to movement artifacts, superior to both fMRI and EEG, allowing investigation of brain activity in more realistic settings incompatible with e.g., fMRI (Quaresima & Ferrari, 2019). Compared to EEG, fNIRS has higher functional resolution (1.5 to 3 cm), but lower temporal resolution. Compared to fMRI, fNIRS has higher temporal resolution, but lower spatial resolution (Pinti et al., 2020). fNIRS costs less than fMRI (typically ~50,000 USD or more for a good-quality fNIRS system).

In contrast to fMRI, EEG, and fNIRS are both portable and do not immobilize participants. Although fMRI remains the gold standard in neuroscience, EEG and fNIRS are both suitable substitutes (Cui et al., 2011; Gagnon et al., 2012b; Herold et al., 2018; Jacko, 2012; Wijeakumar et al., 2017). The combination of EEG and fNIRS has been proposed as an excellent alternative for multimodal integration because fNIRS' high spatial resolution compensates for EEG's low spatial resolution, and EEG's high temporal resolution compensates for fNIRS' lower temporal resolution (Ahn et al., 2016; Ahn & Jun, 2017; Hassib et al., 2017; Li et al., 2022; Lukanov et al., 2016; Pinti et al., 2020).

Moreover, as EEG measures electrical brain activity and fNIRS measure metabolic response, they validate identified brain activity (Li et al., 2022).

Multimodal integration of fNIRS and EEG is relatively new (Ahn & Jun, 2017; Yeung & Chu, 2022), which means the number of available sensor systems is limited. Options might require two computers (one for each modality), custom-built systems (Ahn et al., 2016), or procurement of separate systems recording data independently, requiring data synchronization post-experiment (Al-Shargie et al., 2016; Fazli et al., 2012). This could become expensive. We have proposed one low-cost alternative to multimodal integration of fNIRS and EEG **(Dybvik et al., 2021b; Erichsen et al., 2020)** and demonstrated its usage in-situ **(Dybvik et al., 2021a)**.

Table 3 Comparison of fNIRS with other neuroimaging modalities (Al-Shargie et al., 2016; Herold et al., 2018; Lee & Park, 2018; Pinti et al., 2020)

Feature	**fMRI**	**EEG**	**fNIRS**
Signal	BOLD (HbR)	Electromagnetic	HbR, HbO_2
Spatial resolution	0.3 mm voxels	5–9 cm	1–3/4 cm
Penetration depth	Whole head	Brain cortex	Brain cortex
Temporal sampling rates	1–3 Hz	>1000 Hz	Up to 160Hz
Range of possible tasks	Limited	Large	Enormous
Robustness to motion	Limited	Somewhat robust	Very good
Range of possible participants	Restricted	Everyone	Everyone
Sounds	Very noisy	Silent	Silent
Portability	No	Yes	Yes
Cost	High	Low	Low

3.2.4 Practical challenges in fNIRS measurements

Several practical challenges relate to fNIRS headgear. A stable contact between optodes and scalp is critical for obtaining high signal quality (Quaresima & Ferrari, 2019). Light penetration and measurement depth depend on the thickness of the scalp, skull, and cerebrospinal fluid, hair density and color, and, in some cases, hair due. NIR light is attenuated by layered and darker-colored hair, thus is usually more difficult to obtain high-quality data from participants with darker and thicker hair. It is usually easiest to obtain high-quality data from pre-frontal and temporal regions as they have thinner scalps and lack hair. Figure 4 depicts fNIRS headgear with a montage covering the prefrontal cortex. To obtain high-quality data, it is important to keep the optodes connection to the scalp, ensuring that optodes remain orthogonal to the head’s surface, also during movement, since essentially any changes to the photon banana path causes noise in the signal. Having strong enough optodes and good cable management helps in stabilizing optodes and cap. Optodes must easily penetrate through the hair to obtain scalp contact, for which it helps to have correct optode tips and grommets, while simultaneously being comfortable for participants, and remain in place during recording.

Specialized probe sets may be necessary for unique participants. It could be uncomfortable to wear the cap if proper care isn’t taken when outfitting participants.

Light may impact measurements. Environmental light, such as sunlight and room lighting (especially fluorescent light with overlapping wavelengths), and infrared light-based trackers (e.g., motion-trackers or eye-trackers), may thus pose problems. An over-cap (black shower cap or similar) may be used to block ambient light. Size and fit of cap and over-cap are important: poorly fitting headgear may cause optodes to wiggle during recording, changing the photon banana path, greatly affecting the signal. Caps must fit participants comfortably and thus have correct sizes. Participant head-shape may affect cap fit.

There is evidence for a different hemodynamic response and systemic changes occurring when exposed to differently colored light compared to regular lighting conditions (Metz et al., 2017; Zohdi et al., 2021). Therefore, lighting conditions must be taken into consideration during experimental design, and be consistent during the experiment to ensure comparability of results (Yücel et al., 2021).

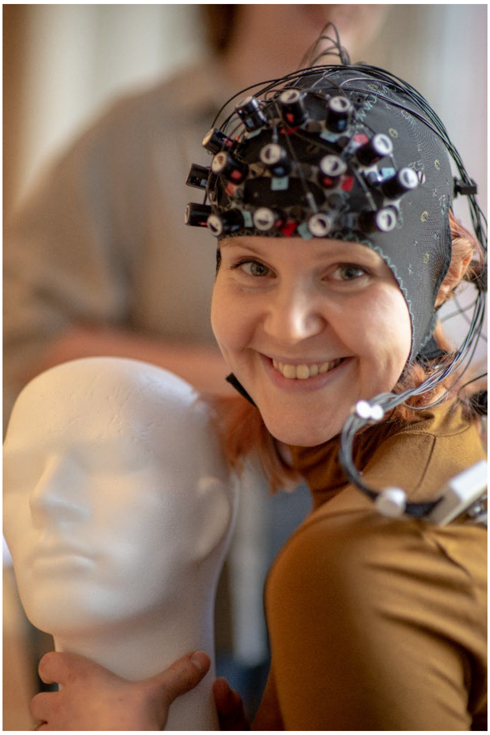


Figure 4 fNIRS headgear. Sources (marked in red) and detectors (marked in blue) are arranged into a montage placed on the prefrontal cortex. [Photo taken by Pasi Aalto]

The number of optodes used must be large enough to cover the region of interest. The resolution of an fNIRS system may be increased by adding optodes. However, additional optodes increase size and weight of the headgear, which could induce greater movement artifacts, and increase setup time and participant discomfort.

Detector sensitivity and source strength impact the measurement. Higher detector sensitivity and greater source strength result in a better signal, which allows for slightly greater spacing and measurement depths. Cable length (for fiber optics) impacts the signal; every 1 m of cable reduces the signal by approximately 5%, which is critical to account for with multimodal MEG and fMRI applications.

In addition to careful sensor application following a protocol, consistent experimental procedure (Cairns & Cox, 2008), signal optimization, and signal quality control are needed to obtain high-quality data **(Dybvik et al., 2021b)**, ensure equal experience across participants, together limiting confounding variables (Cairns & Cox, 2008). Data synchronization is crucial when recording multiple modalities. In one article **(Dybvik et al., 2021b)**, we provided an exhaustive description of a EEG+fNIRS sensor application procedure, data collection, and signal quality control to ensure collection of high-quality data, in a design research context. Further, as multimodal data synchronization protocols are not always disclosed, in the interest of best practices for data synchronization, we fully disclosed our data synchronization protocol and encouraged other researchers to do the same **(Dybvik et al., 2021b)**.

3.2.5 Noise in fNIRS measurements necessitates recording other sensor data

> The noise in fNIRS signal may originate either from the measurement system (e.g., due to light source instability, electronic noise, and shot noise), which we call merely "noise," or of physiological origin or head/body motion which we call "confounding signals" (Yücel et al., 2021, p. 11)

Apart from measurement noise and motion artifacts, changes in systemic physiology contaminate fNIRS measurements and it is worth discussing in greater detail. The fNIRS signal is comprised of neuronal activity, but also cerebral and extracerebral changes in systemic physiology which may or may not be task triggered (Tachtsidis & Scholkmann, 2016), see Figure 5. These sources of physiological noise operate on different frequency bands. E.g., changes in respiration (~5 sec period, 0.2-0.3 Hz), heart rate (~1 second period, ~1 Hz), blood pressure (<0.05 Hz), partial pressure of carbon dioxide in the arterial blood ($PaCO_2$), cerebral blood flow/volume, skin blood flow/volume (Quaresima & Ferrari, 2019; Tachtsidis & Scholkmann, 2016), Mayer waves/low frequency oscillations (~10 sec period, 0.08-0.12 Hz) (Huppert, 2016; Obrig et al., 2000; Paranawithana et al., 2022; Scholkmann et al., 2014; Yücel et al., 2016), and very low frequency oscillations (VLF) (<0.1 Hz) (Scholkmann et al., 2014).

Therefore, in particular for experimental tasks inducing larger systemic vascular changes or changes in breathing pattern, it is desirable to record the changes of these parameters (heart rate, blood pressure, skull blood flow, partial end-tidal carbon diaoxide, etc.) by means of other physiology sensors (Quaresima & Ferrari, 2019; Yücel et al., 2021; Zohdi et al., 2021). Extra-cerebral systemic changes can be measured with short-channels (Quaresima & Ferrari, 2019). It is moreover recognized that the combination of systemic physiology measures and fNIRS could provide new insight on the interaction and coordination between brain activity and systemic physiology (Pinti et al., 2020; Scholkmann & Vollenweider, 2022; Yücel et al., 2021).

> It is well known that the combination of the data from different techniques allows a description of human brain activity with a combination of spatial and temporal precision and contrast mechanisms that are impossible to achieve using any single imaging modality. (Ferrari & Quaresima, 2012)

In one experiment, we collected multiple physiology sensors and neuroimaging modalities. While tasking participants with a Tetris gameplay, we recorded ECG, GSR, EEG, and fNIRS, allowing novel investigations of interactions between brain activity and systemic physiology **(Dybvik et al., n.d.)**.

A lack of financial resources could limit multimodal data collection. We developed a wearable multimodal neuroimaging sensor setup with EEG+fNIRS at a lower financial threshold compared to most setups **(Dybvik et al., 2021b)**.

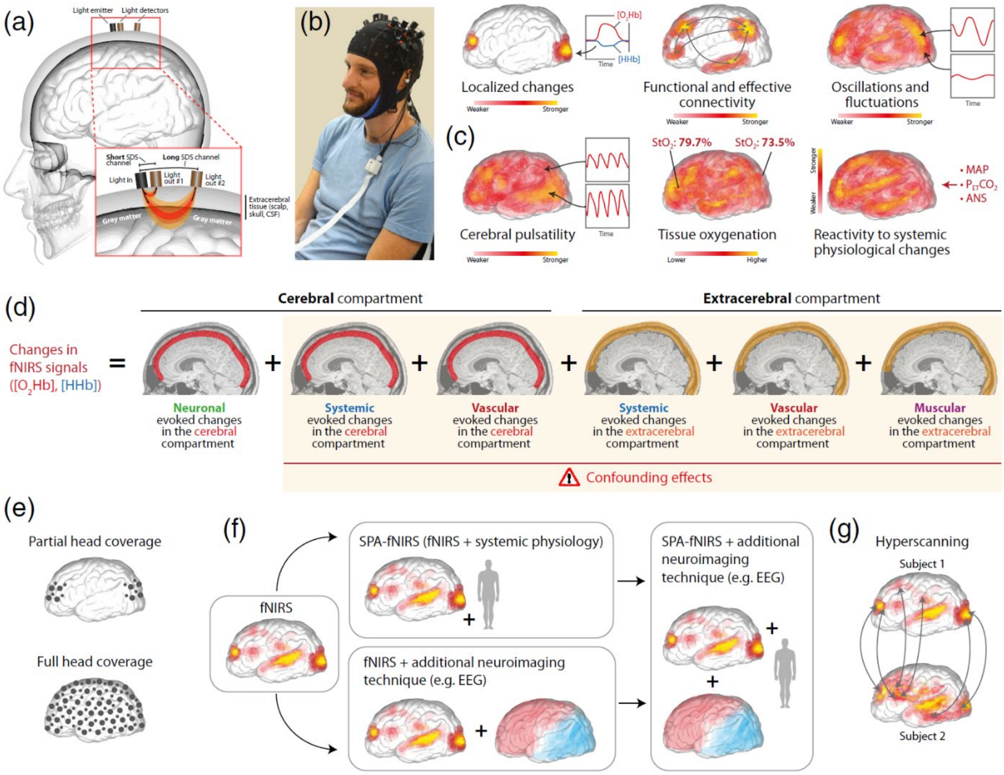

(a)
Light emitter
Light detectors
Short SDS channel
Long SDS channel
Light in
Light out #1
Light out #2
Extracerebral tissue (scalp, skull, CSF)
Gray matter
Gray matter
(b)
(c)
[O$_2$Hb]
[HHb]
Time
Localized changes
Weaker
Stronger
Functional and effective connectivity
Weaker
Stronger
Oscillations and fluctuations
Weaker
Stronger
Time
Cerebral pulsatility
Weaker
Stronger
Time
StO$_2$: 79.7%
StO$_2$: 73.5%
Tissue oxygenation
Lower
Higher
Stronger
Weaker
• MAP
• P$_{ET}$CO$_2$
• ANS
Reactivity to systemic physiological changes
(d)
Cerebral compartment
Extracerebral compartment
Changes in fNIRS signals ([O$_2$Hb], [HHb])
=
Neuronal evoked changes in the cerebral compartment
+
Systemic evoked changes in the cerebral compartment
+
Vascular evoked changes in the cerebral compartment
+
Systemic evoked changes in the extracerebral compartment
+
Vascular evoked changes in the extracerebral compartment
+
Muscular evoked changes in the extracerebral compartment
Confounding effects
(e)
Partial head coverage
Full head coverage
(f)
fNIRS
SPA-fNIRS (fNIRS + systemic physiology)
+
fNIRS + additional neuroimaging technique (e.g. EEG)
+
SPA-fNIRS + additional neuroimaging technique (e.g. EEG)
+
+
(g)
Hyperscanning
Subject 1
Subject 2

Figure 5 Visualization of key fNIRS aspects. (a) A two-channel fNIRS measurement using a long and a short channel. (b) Typical fNIRS headgear, here covering the right and left motor cortices. (c) The six main aspects that can be determined with optical neuroimaging employing fNIRS and NIRS-based oximetry. (d) The six main components of an fNIRS signal. (e) Example of two different spatial positionings for montages coverages spatial positioning of the light sources and detectors on the head. (f) Visualization of current trajectories of fNIRS development. (g) fNIRS hyperscanning [Figure and parts of figure text reproduced from Scholkmann & Vollenweider (2022)].

3.3 fNIRS measures brain regions of interest

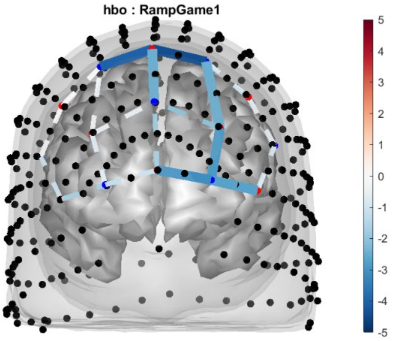
Anatomically, the human brain may be divided into functional and structural regions (or areas) at several resolutions. The cerebral cortex may broadly be divided into four: the frontal lobe, parietal lobe, temporal lobe, and occipital lobe. The Broadman areas have seminally been defined and numbered based on their cytoarchitecture (i.e., cellular composition and organization), and are frequently used today. fMRI research has localized and attributed various cognitive processes to distinct brain regions. The prefrontal cortex, a subregion of the frontal lobe, is associated with executive functions, including attention and awareness, decision-making, planning and selecting complex cognitive behavior, cognitive flexibility (task-switching), working memory, etc. (Fishburn et al., 2014; Funahashi, 2017; Miller et al., 2002; Purves et al., 2012; Vanderhasselt et al., 2009). Attention and working memory are also mediated by the parietal association cortex located in the parietal lobe. The visual cortex, a subregion of the occipital lobe, is associated with processing visual sensory input. The motor cortex, a subregion of the frontal lobe, is associated with the planning, initiation, and execution of movement (Purves et al., 2012). Assuming these general associations also hold outside fMRI, we may use the mapping of functional regions provided in anatomical atlases to define or select brain regions of interest a priori. Section 3.4.1.2 provides further detail on this.

For example, in the context of engineering and design cognition, Hu & Shealy (2019) have synthesized cognitive functions associated with decision-making in sub-regions of the prefrontal cortex. Moreover, Milovanovic et al. (2020) provide a non-exhaustive overview of functions in design and creative thinking associated with the prefrontal cortex.

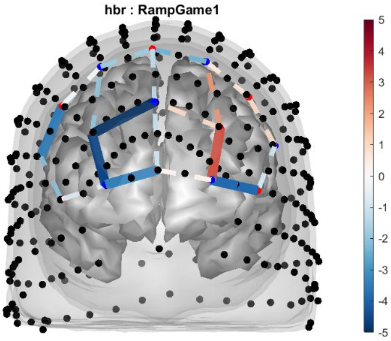
Cognitive load is one example of a frequently measured construct (Fishburn et al., 2014; Meidenbauer et al., 2021). It has been shown that fNIRS is sensitive to load-dependent activation changes in the prefrontal cortex due to different working memory loads (Fishburn et al., 2014). Cognitive load-dependent changes in the frontal and parietal cortical region is now a well-studied effect in both fNIRS and fMRI (Cui et al., 2011; Fishburn et al., 2014; Meidenbauer et al., 2021). The n-back is one common working memory task where load is easily manipulated (Cui et al., 2011; Fishburn et al., 2014; Hamann & Carstengerdes, 2022; Herff et al., 2014; Meidenbauer et al., 2021; Rahman et al., 2021). In our studies, cognitive load has also been the primary construct of interest **(Dybvik et al., n.d., 2021a, 2018; Dybvik & Steinert, 2021; Wulvik et al., 2019)**.

In our Tetris experiment we tasked participants with three 4-minute Tetris gameplays **(Dybvik et al., n.d.)**. One constant, low-difficulty game (EasyGame), one constant, high-difficulty game (HardGame), and one game where the difficulty level ramped up from very low at the beginning to very hard at the end (RampGame). Figure 6 displays main effects of RampGame at the group-level. Here, we observe an increase in HbR reduction from the 1st throughout 3rd minute, followed by a decreased HbR reduction. HbO_2 follows a similar pattern: starting with overall HbO_2 reduction, the number of significant channels decreases with increasing game difficulty, accompanied by the introduction of significant HbO_2 increases, up to the 3rd minute. Thereafter, a decrease in significant channels follows. Taken together, this could signify increased cognitive load (hemodynamic activation) in participants up to a certain threshold (between 3rd and 4th minute), after which participants exhibit lower activation, possibly due to cognitive overload.

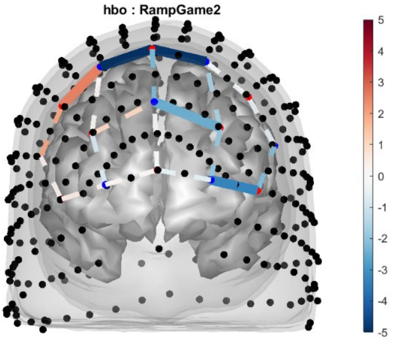
hbo : RampGame1



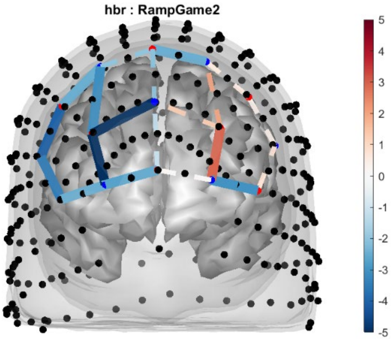
hbr : RampGame1



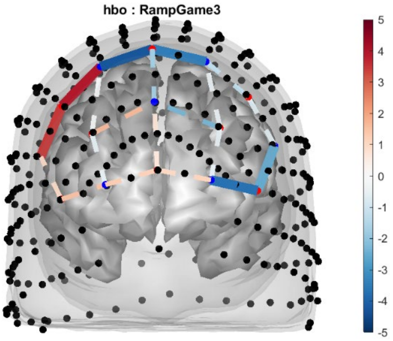
hbo : RampGame2



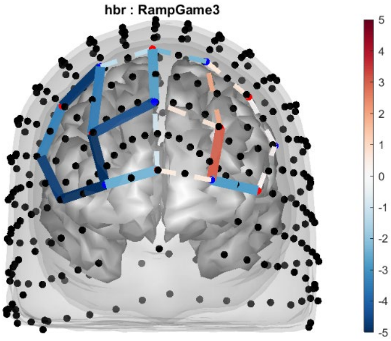
hbr : RampGame2



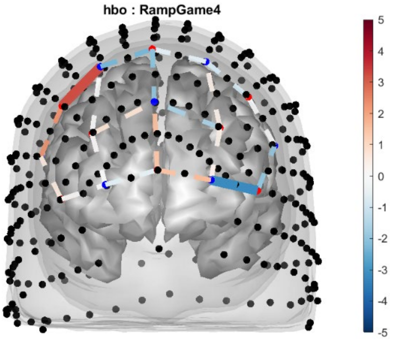
hbo : RampGame3



hbr : RampGame3



hbo : RampGame4



hbr : RampGame4

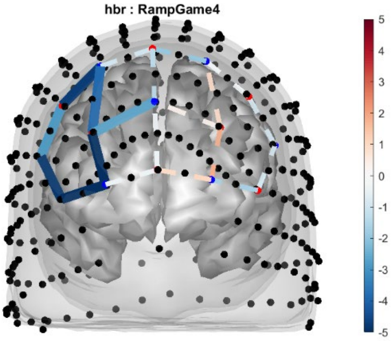


Figure 6 Main effects of Ramp Tetris game relative to baseline. The t-statistic is plotted onto the Colin27 atlas using NIRS Brain AnalyzIR toolbox (Santosa et al., 2018). Left) HbO_2 for minutes 1 (RampGame1) through 4 (RampGame4). Right) HbR for minutes 1 through 4. The color bar represents the t-statistic scaled to [-5,5] with full red/blue lines indicating statistically significant increase/decrease at an FDR-corrected p-value q<0.05.

We moreover found that there appears to be a temporal effect of workload, independent of difficulty, by comparing the 1st to 4th (i.e., last) minute of the gameplays with constant difficulty, i.e., EasyGame and HardGame **(Dybvik et al., n.d.)**. Figure 7 illustrates that the 4th minute was more cognitively demanding than the 1st minute in EasyGame. Figure 8 illustrates that the 4th minute was more cognitively demanding than the 1st minute in HardGame. As a sanity check, we compared the 1st minutes of EasyGame and HardGame to verify that EasyGame was less cognitively demanding than HardGame (see Figure 9).

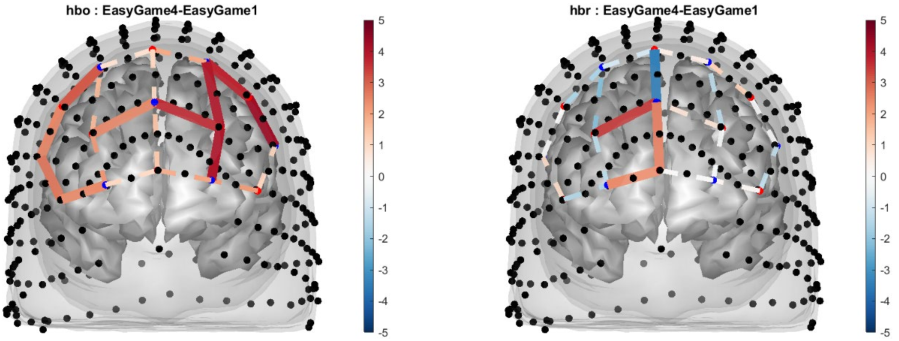


Figure 7 Contrast comparing the first and last minute of the Easy Tetris game. The t-statistic is plotted onto the Colin27 atlas using NIRS Brain AnalyzIR toolbox (Santosa et al., 2018). The last minute was more cognitively demanding than the first minute as evidenced by increased HbO_2 in the last minute compared to the first. HbR is depicted to the left. The color bar represents the t-statistic scaled to [-5,5] with full red/blue lines indicating statistically significant increase/decrease at an FDR-corrected p-value q<0.05.

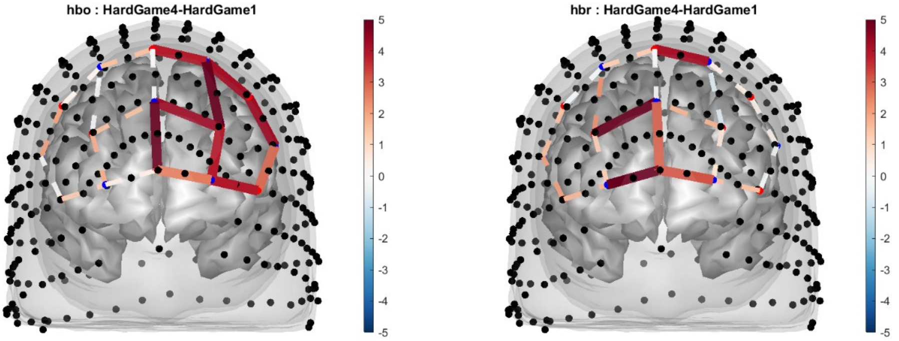


Figure 8 Contrast comparing the first and last minute of Hard Tetris game. The t-statistic is plotted onto the Colin27 atlas using NIRS Brain AnalyzIR toolbox (Santosa et al., 2018). The last minute was more cognitively demanding than the first as evidenced by increased HbO_2 (left side). HbR is depicted on the right. The color bar

represents the t-statistic scaled to [-5,5] with full red/blue lines indicating statistically significant increase/decrease at an FDR-corrected p-value q<0.05.

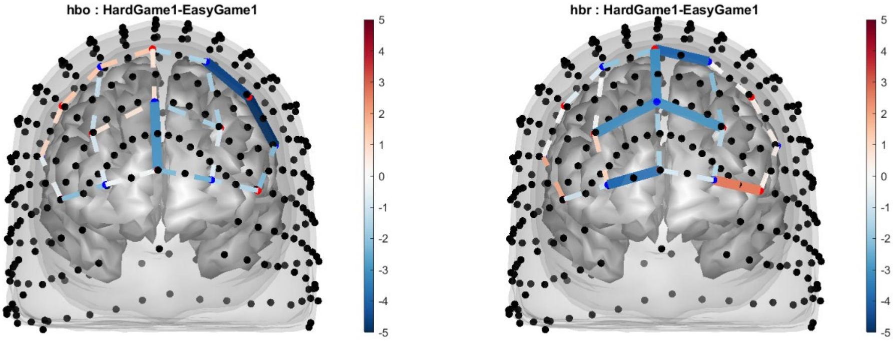
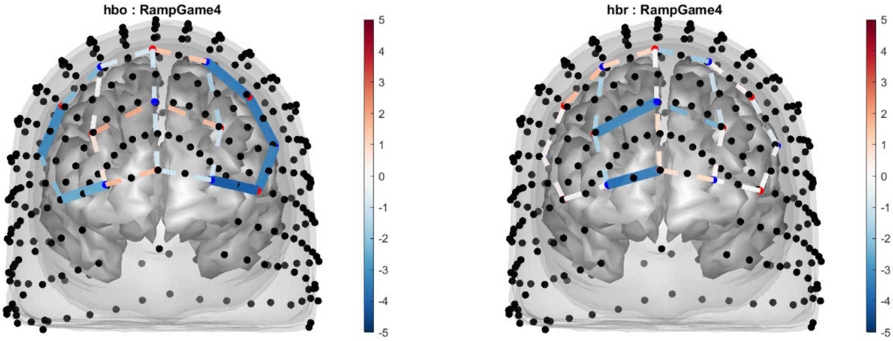


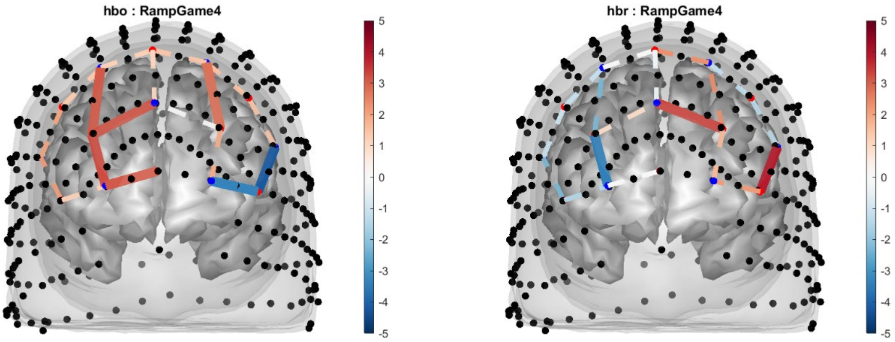
Figure 9 Contrast comparing the first minute of Hard and Easy Tetris game. The t-statistic is plotted onto the Colin27 atlas using NIRS Brain AnalyzIR toolbox (Santosa et al., 2018). The first minute of HardGame appears to be more cognitively demanding than the first minute of EasyGame, as primarily evidenced by a reduction in HbR (right side). HbO$_2$ is depicted to the left. The color bar represents the t-statistic scaled to [-5,5] with full red/blue lines indicating statistically significant increase/decrease at an FDR-corrected p-value q<0.05.

Figure 10 depicts individual activation maps for six participants during the 4th minute of RampGame[15]. Here, we see that though subjected to exactly the same level of difficulty, there is high individual variance between participants. It seems some participants still are engaged in the task, able to recruit neuronal activation (participant 2 in particular, but also participant 5), whereas others have completely disengaged, as exhibited by low cognitive activation (e.g., participants 1 and 3). Other participants (3 and 6) exhibit completely different activation patterns. As described in 2.1.2, neuroimaging and physiology sensors—here as represented by fNIRS—enables the ability to account for individual differences.

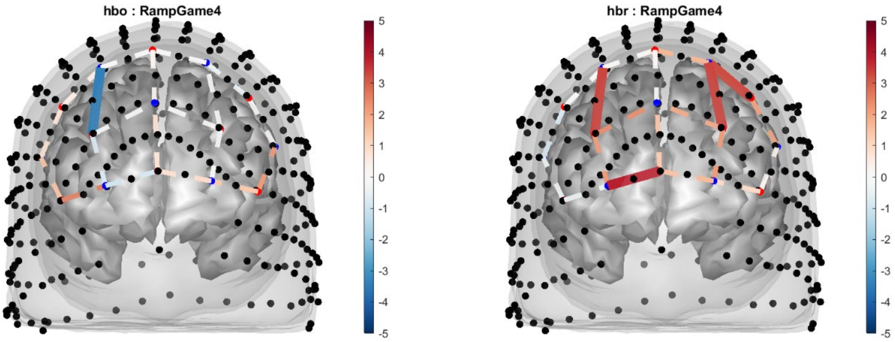


Participant 1

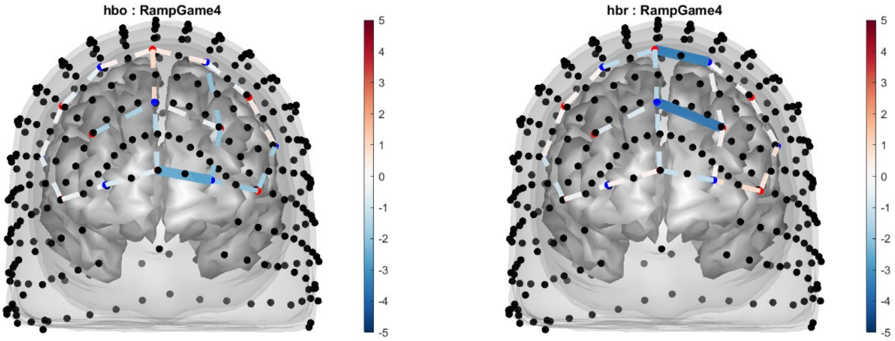
[15] These data come from the Tetris experiment **(Dybvik et al., n.d.)**, but are not presented in the appended manuscript.



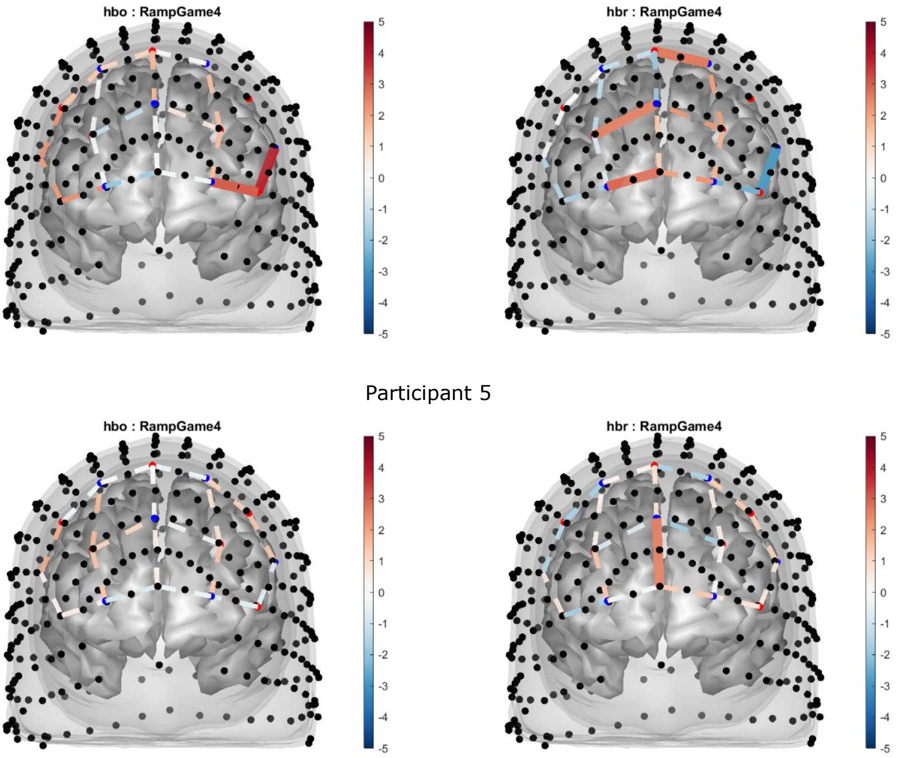
Participant 2



Participant 3



Participant 4



Participant 5

Participant 6

Figure 10 Individual participants in the 4th minute of RampGame. HbO_2 is depicted on the left, HbR is on the right. Rows represent individual participants. The t-statistic is plotted onto the Colin27 atlas using NIRS Brain AnalyzIR toolbox (Santosa et al., 2018). The color bar represents the t-statistic scaled to [-5,5] with full red/blue lines indicating statistically significant increase/decrease at an FDR-corrected p-value q<0.05.

In design, we could use this for researching designers working with- or testing out new tools, to identify which tool yielded the most beneficial cognitive activation for the individual designer, thereby adapting the design process to the individual. Furthermore, this could be useful in product evaluation of e.g., user interfaces for shore control centers. Here, we could compare different user interface designs on their effect on the user’s cognitive load during interaction. We could determine, also at an individual level, which interface yields highest cognitive activation and attention, and which interface, if any, was too difficult or cumbersome to use, yielding cognitive overload.

3.4 Design of experiments with fNIRS

fNIRS may be used in experiments to investigate causal relationships, and in correlational (observational) methods; however, this thesis focuses on experiments. Design of experiments with fNIRS begins by asking what the research question is. The selected task and construct sought to investigate identifies brain regions of interest.

3.4.1 Montage: How to place sources and detectors?

3.4.1.1 The montage

The montage is an arrangement of optodes (sources and detectors) for assessing cortical activation in regions relevant to the experimental hypothesis (NIRx Medical Technologies, 2019b). fNIRS experiments are most often designed with a limited number of sources and detectors, and it is therefore important that optodes are positioned on the portion of the scalp that most effectively captures relevant cortical brain regions. Deciding on a layout for the montage typically consists of identifying brain regions of interest (ROI) and using this information to either select a predefined montage or design a new one.

Given a certain number of optodes, there are, in practice, two interrelated problems one needs to solve: 1) Where to place optodes i.e., which brain areas do we want to cover?, and 2) How do we connect sources and detectors to create viable channels? Ideally, one maximizes coverage of the ROIs while minimizing coverage of no-interest areas, further maximizing channel density and optode stability to minimize motion. If fNIRS is combined with other modalities (e.g., EEG), one must ensure that there is room in the cap for these modalities. In fNIRS, it is most common to keep all channels at a fixed, standard length; this is known as topographic imaging and analysis. The standard distance between source and detector is 3-4 cm in adults and 2-3 in children and infants (Ferrari & Quaresima, 2012; NIRx Medical Technologies, 2019b; Quaresima & Ferrari, 2019). Short-channel distances are typically 8 mm for adults (Wyser et al., 2020); they must be included in the montage.

3.4.1.2 Brain-scalp correspondence

As described in section 3.2.2, although individual MRIs of each participant's brain would be ideal, this is often not possible. In most cases, one must use other approaches to identify which position an optode must have on the scalp to cover the desired ROI.

The brain-scalp correspondence—so-called cranio-cerebral correlation—has been computed based on the standard 10-20 system for EEG electrode placement for adults (Okamoto et al., 2004a) and infants (Kabdebon et al., 2014; Tsuzuki & Dan, 2014). As such, the correspondence between Montreal Neurological Institute (MNI) and Talairach MRI coordinates, 10-20 EEG electrode locations, and anatomical atlases can be found in literature (Cutini et al., 2011; Okamoto et al., 2004a). Therefore, the 10-5/10-10/10-20 electrode positions (Oostenveld & Praamstra, 2001) can be used as a proxy for a given brain ROI. fNIRS channels can be located around those positions. Optodes are, therefore, often placed on a cap that follows the standard 10-20 or 10-10 system for EEG electrode placement. However, this depends on the available fNIRS system—some systems only have predefined or static montages.

Several tools are available for exploring potential ROIs, montage design, or sensitivity analysis of the montage (Figure 11). These include software: e.g., NIRSite (NIRx Medical Technologies, Berlin, Germany), fOLD toolbox (Zimeo Morais et al., 2018), AtlasViewer

(Aasted et al., 2015), and anatomical atlases: e.g., Colin27, and ICBM152 (fMRI community's brain template) (Fonov et al., 2009, 2011).

In our studies **(Dybvik et al., n.d., 2021a, 2021b; Dybvik & Steinert, 2021)** we have used a pre-defined montage that covers the prefrontal cortex, as we are interested in measuring cognitive load. Figure 11 is reproduced from **Dybvik & Steinert (2021)** displays the sensitivity profile of this montages, and the montage rendered onto the ICBM152 brain atlas.

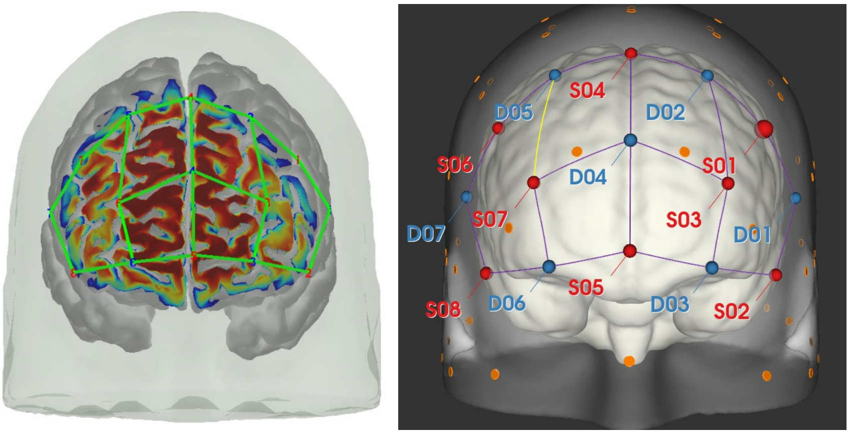


Figure 11 An example of a montage covering the prefrontal cortex. Left: The sensitivity profile of the probe, created with AtlasViewer (Aasted et al., 2015). Right: The montage rendered onto the ICBM 152 Nonlinear atlases version 2009 (Fonov et al., 2009, 2011). [Figure reproduced from Dybvik & Steinert (2021)]

3.4.2 Types of experimental design

There are several main types of experimental designs. An emphasis is placed on block design.

The experimental design must take into account the hemodynamic response (HR). Inter-stimulus interval (ISI) (also known as interblock interval) is therefore really important for all designs (Yücel et al., 2021). It is recommended to avoid fixed ISI, and an ISI of around 4 s, as this does not allow the hemodynamic response function (HRF) to return to baseline. HRF exhibits considerable variability between participants, and while responses are more consistent within participants, there is some variability between sessions (Aguirre et al., 1998; Zohdi et al., 2021).

3.4.2.1 Event-related designs

Event-related designs assume that the shape of HRF is predictable. Event-related potentials (ERP) come from EEG and are based on averaging many responses over many trials. The same is possible with fNIRS. ERP-designs associate brain processes with discrete events, typically occurring at any point in time during the session, through which we can detect transient hemodynamic responses. ERP-designs are suitable when the experimental tasks are naturally event-related. They allow studying within-trial effects, and may better explain the relation with behavioral factors, as behavioral changes may be masked within blocks. However, analyses are more complex, dependent on accurate modeling of the HRF, relying on selective averaging and GLMs. ERP-designs may include single or multiple trials, be periodic (constant, long ISI) or jittered (ISI varies between trials).

Considerations for effective ERP-designs: the ISI should be at least 2 seconds (not faster than 0.5 Hz). ERP-designs depend on accurate modeling of the HRF. To increase statistical power in an ERP-design, one may use a jittered ISI, include more repetitions per trial type, and (as always) increase the number of participants. A typical event-related design is the go/no-go task (NIRx Medical Technologies, 2019a).

3.4.2.2 Block design

A block design separates tasks into distinct time periods (blocks). During each block, participants might perform the same task several times with a small ISI. Alternatively, one task may constitute the entire block. Block designs may be "tight" (more than 1 task, including baseline, which enables multiple contrasts—a preferred design) or "loose" (1 task, one contrast with baseline—this might be weak). Two example designs: A design comparing task A to task B allows for distinguishing differential activation between conditions, but cannot identify activity common to both tasks. For that, we need a control condition. A design comparing task A to a non-task (to potentially task B, task C, etc.) allows investigation of the activity associated with the task, but may introduce erroneous results if task and non-task are not appropriately matched, e.g., rest data acquired while eyes were closed, while task had eyes open. Crucially, block designs assume that HR adds linearly (Dale & Buckner, 1997).

Important considerations for choosing block length include the following. Longer block lengths allow stability of extended responses, but the HR saturates after extended stimulation: activation plateaus after about 10 s. Shorter blocks move the signal to higher temporal frequencies and away from low-frequency noise (e.g., Mayer waves). Periodic blocks may result in aliasing of other periodic signals in the data. This could be a problem if the aliased signal falls within the range of desired signals, e.g., breathing rate of 12 per min and blocks are 10 seconds long (6 blocks per min). Periodic blocks could induce anticipation effects. Generally, it is recommended to avoid stimulus frequencies that fall within ranges that usually are filtered out. Mayer wave oscillations (blood vessels) occur around ~0.1 Hz (10s), which may overlap with stimuli presentation (Pinti et al., 2019).

Block designs may be limited by poor choice of conditions/baseline. Further, many tasks cannot be conducted well repeatedly; linearity of the HR cannot always be assumed for all paradigms; and habitation effects may be induced. Block designs are more likely than ERP-designs to trigger systemic responses (i.e., evoke non-brain activity effects).

Considerations for effective block designs: It is necessary to allow time for the HR to return to baseline, without allowing for baseline wandering, as this may result in uncontrolled activation. The HR typically returns to baseline 15-20 seconds after task offset. Baseline wander may occur if participants are left without a task for ~30 seconds or more. Longer block durations may be (mentally) difficult for participants. The linearity assumption is violated when saturation takes place. Thus, block durations over 60 seconds are not recommended. Habituation, learning effects, or anticipation effects may occur with repeated tasks: participants may anticipate or learn the task to a degree that reduces HR. For block designs in particular, an ISI of 4 seconds should be avoided.

In the Tetris experiment **(Dybvik et al., n.d.),** we used a block design composed of three blocks, each with one Tetris gameplay. As each Tetris game lasted for 4 minutes, i.e., longer than the recommended block length, we separated each condition into 4 blocks of 60 seconds each in the analysis. The block order followed a 3x3 Latin Square Design. As such, order effects at the group level should be avoided. To minimize learning effects, we included a practice session before the blocks. Participants filled out questionnaires between blocks. The conventional laptop case demonstration **(Dybvik et al., 2021a)** followed the same experiment design, except that a two-minute resting period interspersed blocks. In the yoga study **(Dybvik & Steinert, 2021),** we employed the Ashtanga Vinyasa Yoga primary series, a well-documented, standardized sequence of postures. Each posture is performed in a standard order with a transition (a vinyasa) between. Even though the experiment was constrained to follow the Ashtanga practice, the experiment design closely resembled a block design where each block consisted of one posture. Our driving experiment **(Dybvik et al., 2021a)** had a more naturalistic design; it involved a two-minute baseline inside the parked car, before 20 minutes of city center driving and 20 minutes of highway driving.

3.4.2.3 Mixed designs

Mixed designs combine block and event-related designs purposefully. This allows investigation of state-dependent effects (block) and item-related effects. If cognitive processes are independent, they may be modeled as separate phenomena, or one may model their interaction.

For all experimental designs, it is generally always good to jitter the ISI. Block designs are powerful for detecting activation and useful for examining state changes. ERP-designs are powerful for defining the activity's time course, tend to avoid habituation effects, be less predictable, but may be more time-consuming, and have smaller HR (worse signal-to-noise ratio (SNR)). Mixed designs represent the best combination of detection and estimation, but the analyses are more complicated than block- and ERP-designs.

3.4.3 Control condition and baseline

Scientific experiments fundamentally rely on including a control or baseline condition, to which the alternative condition (intervention, condition of interest) must be compared. Within cognitive neuroscience, this has posed a problem. Brain activity has been shown to predictably vary when left unconstrained (Gusnard & Raichle, 2001), producing the collection of resting state brain processes called the Default Mode Network (DMN)[16]. The DMN was discovered when researchers found greater activation in baseline conditions compared to task conditions. For example, there exists evidence for greater (and differential) activation during task interpretation than in the actual task. Fortunately, fluctuations in resting state brain activity are consistent across participants (Damoiseaux et al., 2006). Because of the DMN (Raichle et al., 2001), and that different brain areas may have different baselines (Gusnard & Raichle, 2001), it is important to search for the best control condition or baseline.

Simple resting periods (even 20 minutes or more) do not guarantee baseline stability (Jennings et al., 1992). As such, "vanilla baselines," which are minimally demanding tasks, have been proposed as alternates, having shown better (or equal) between- and within-baseline-stability, response amplitude and stability between different session days (Jennings et al., 1992). These tasks could, for example, ask participants to count the

[16] or resting state networks

number of times a certain color occupied the screen while watching a video, having the color change every 10 seconds (Jennings et al., 1992).

One of our in-situ demonstration cases **(Dybvik et al., 2021a)** tasked participants to drive around Trondheim city center and on the highway. Here, we recorded a two-minute baseline while the participant relaxed inside the parked car before they started driving. In the Tetris experiment **(Dybvik et al., n.d.),** we attempted to record 40 seconds of resting-state as a control condition before commencing the experiment. But, the preliminary analysis of fNIRS, GSR, and ECG implied that participants were highly alert, likely affected by being in an experiment setting, and thus that the measurement would not be representative of a resting state. Fortunately, as we mainly were interested in the difference between conditions, a resting-state control condition was not strictly necessary.

3.5 Analysis

As mentioned, fNIRS is a relatively young field that has inherited and learned many analysis methods from fMRI, EEG, physics, and other communities. Consequently—and important—there is no established standard processing pipeline nor signal quality assessment. Some pre-processing steps require many input parameters (Hocke et al., 2018; Pinti et al., 2019), the choice of which considerably influences the results (Pfeifer et al., 2018). Although there is high heterogeneity in methodological procedures, we attempt to provide a general overview of fNIRS analysis here.

Analyses aim to, first, obtain an accurate estimate of oxygenated and deoxygenated hemoglobin. Second, to obtain valid statistical inferences. To reduce type 1 (false-positive) and type 2 (false-negative) errors, we need to deal with sources of noise. As detailed previously (section 3.2.5), apart from measurement noise from the fNIRS system, there are generally two sources of noise: motion artifacts and systemic physiology. Methods for removing noise stemming from physiology and motion are discussed together (in section 3.5.3) and can be categorized into 1) prefiltering approaches and 2) statistical approaches (Santosa et al., 2020; Yücel et al., 2021). Before that, we briefly outline the necessary fundamentals and signal quality assessment.

3.5.1 The necessary fundamentals: Computation of concentration changes

There are two fundamentally necessary steps in all fNIRS analysis: 1) Optical density calculation, which converts from raw voltages measured by the detectors to optical density, and 2) Modified Beer-Lambert Law (Delpy et al., 1988), which converts from optical density to hemoglobin concentrations.

3.5.2 Signal quality assessment

There are several ways to assess the signal quality of fNIRS data.

> One commonly used criterion is the visual inspection of the […] signal for presence of cardiac pulsation, in either the time or frequency domain. The rationale being that the presence of a cardiac oscillation, which is robust and consistent, indicates that changes in optical density are coupled with physiological hemodynamic changes (Hocke et al., 2018).

The time series data may be visually inspected, in the form of raw light intensities, optical density, or concentration changes, for clear cardiac fluctuations (Hocke et al., 2018; Yücel et al., 2021). Although visual inspection is a popular option (Hocke et al., 2018; Nguyen et al., 2021; Yücel et al., 2021), the drawbacks are that the cardiac pulsation of the wavelength weighted to deoxyhemoglobin often is less robust, it is subjective, and time-consuming (Hocke et al., 2018). Hence, the fNIRS community seeks objective, quantitative signal quality measures, but no standardized criteria have been established yet (Sappia et al., 2020). Automated methods for signal quality control that assess presence of noise include coefficient of variation (CV) and signal-to-noise ratio (SNR). The scalp coupling index (SCI) quantifies presence of cardiac pulsation by cross-correlation between the two wavelengths for each channel, obtaining a correlation coefficient between 0 and 1.0 (Hernandez & Pollonini, 2020; Olds et al., 2016; Pollonini et al., 2016). SCI aims at assessing optode coupling with the scalp. Another proposed measure of cardiac signal strength is the spectral power of the cross-correlated signal. Specifically, the peak [peak spectral power (PSP)] focuses on assessing presence of motion artifacts, as motion artifacted signals have lower peak values, whereas clean signals typically have higher peak values (above 0.1) (Hernandez & Pollonini, 2020; Pollonini et al., 2016). SCI and PSP both discriminate binarily between good and bad quality signal, and they may be used together (Hernandez & Pollonini, 2020; Pollonini et al., 2016). Aiming to add resolution to this binary classification scheme, the signal quality index (SQI) assesses quality on a numeric scale from 1 (very low quality) to 5 (very high quality). The ability to discern good-, medium-, and low-quality data can be crucial in challenging experimental settings (Sappia et al., 2020), e.g., in-situ, and it is called for in the fNIRS community.

3.5.3 Removing noise

3.5.3.1 Prefiltering approaches

Prefiltering approaches entail a two-step process. Noise is first attempted removed or corrected by converting motion-contaminated signals into uncontaminated signals (Fishburn et al., 2019), before statistical analysis. Statistical estimates of brain activity are here inherently dependent on the selected preprocessing technique. This may remove systemic noise, but reduce sensitivity and introduce additional type 2 errors (false-negatives) (Santosa et al., 2020). It is outside the scope of this thesis to provide a comprehensive review of all preprocessing techniques. However, we provide a list of the most common techniques (see Table 4) and refer to others for comprehensive reviews (Huang et al., 2022; Pinti et al., 2019; Tak & Ye, 2014). Nevertheless, several points are worth noting. Some filtering techniques are a combination of several algorithms; BCI applications typically distinguish between noise removal techniques that can be applied online (during data collection for real-time feedback) and offline (processing post data collection); some filtering techniques must be applied at a particular stage in the preprocessing pipeline, while some may be applied at any stage (Huang et al., 2022). E.g., CBSI (Cui et al., 2010) must be applied to hemoglobin concentration changes, but wavelet-based methods can be applied to optical intensities, optical densities, and concentration changes (Huang et al., 2022; Molavi & Dumont, 2012).

3.5.3.2 Statistical approaches

Statistical approaches entail one single step where noise corrections are integrated directly into the statistical model. Here, a model of brain activity is generalized to also account for systemic physiology. Most commonly, this involves adding regressors of no interest (nuisance regressors) to a linear regression model that attempts at estimating brain activity. These nuisance regressors could either be external physiology measurements (e.g., pulse oximetry or respiratory belt) or estimation of scalp hemodynamics from short-channel measurements. The issue with these regression models is collinearity introduced between task and nuisance regressors, which may occur if the systemic physiological response is correlated with the task. Collinearity destabilizes regression analyses, potentially producing unpredictable results (Santosa et al., 2020).

Another alternative, precoloring and prewhitening approaches, modifies the assumptions of the statistical model (Huppert, 2016; Santosa et al., 2020). Both approaches generalize the linear model to correct errors because of the violated statistical assumptions that physiological noise is uncorrelated, normally distributed, and white (cf. structured and colored noise, see section 3.5.4 for details). However, these methods do not account for errors due to nonstationary noise or noise synchronized with the task (Santosa et al., 2020).

Important, "statistical corrections are not exclusive to the use of nuisance regressors in the model or additional preprocessing stages and all approaches can be used together [ultimately] creating a large array of possible analysis pipelines (Huppert, 2016)" (Santosa et al., 2020). It is thus not surprising that a standard analysis pipeline has not been established yet.

Table 4 Different processing techniques to remove noise in fNIRS measurements

Processing technique	**Citation**	**Article including method in comparison**
Adaptive filter	(Zhang et al., 2009)	-
AR-IRLS	(Barker et al., 2013, 2016)	(Santosa et al., 2020)
Baseline-derived PCA	(Franceschini et al., 2006)	(Santosa et al., 2020)
Bandpass filter	(Mesquita et al., 2010; Santosa et al., 2013)	(Santosa et al., 2020)
CBSI	(Cui et al., 2010)	(Fishburn et al., 2019)
GLM using SS filter	(Sato et al., 2016)	(Santosa et al., 2020)
ICA	(Santosa et al., 2013)	-
ICA using SS channels	(Funane et al., 2014)	-
Kalman filter	(Diamond et al., 2005; Hu et al., 2010)	-
PCA	(Zhang et al., 2016, 2005)	(Fishburn et al., 2019; Santosa et al., 2020)
SS as regressors	(Gagnon et al., 2012a; Sato et al., 2016)	(Santosa et al., 2020)
Spline interpolation	(Jahani et al., 2018; Scholkmann et al., 2010)	(Fishburn et al., 2019)
TDDR	(Fishburn et al., 2019)	(Fishburn et al., 2019)
OLS	(Gratton & Corballis, 1995)	(Santosa et al., 2020)
Wavelet analysis	(Holper et al., 2015; Molavi & Dumont, 2012)	(Fishburn et al., 2019)

3.5.4 The GLM and its statistical assumptions

The generalized linear model (GLM) framework is commonly used for fNIRS. The GLM analysis approach may be used on two levels, the individual, and the group level. The statistical assumptions of the GLM are that responses add linearly, and that noise (the error term) is normally distributed with zero mean, homoscedastic, independent and identically distributed (i.i.d), not autocorrelated, is stationary and ergodic. However, the sources of noise and artifacts occurring in fNIRS measurements violates these assumptions—this will be explained briefly in the following.

There exists evidence for nonlinear brain responses, in particular for short stimulus intervals (less than 4 seconds ISI). Noise arising from physiological changes (such as cardiac, respiratory and blood pressure changes) is serially correlated (each time point depends on the previous, i.e., exhibits autocorrelation) and has a colored structure, which means that specific temporal frequencies are overrepresented (frequencies corresponding with cardiac, respiratory and Mayer wave fluctuations). Measurements from different channels are not independent of each other. Correlations between channels occur due to low spatial frequency of systemic and superficial physiological signals. Motion artifacts may moreover induce a strong spatial noise structure. Furthermore, oxygenated and deoxygenated hemoglobin are partially correlated variables. Motion artifacts are often considerably greater in magnitude than physiological noise, giving rise to outliers contributing to a heavy-tailed noise distribution. The noise distribution (structure) may also vary considerably across channels as it is a function of how well an optode is placed on the scalp, which may differ greatly depending on hair thickness and other factors (cf. section 3.2.4) In summary, noise is typically not normally distributed and exhibits heteroscedasticity. Systemic physiology may change according to the task (i.e., be task-dependent), thus, the noise structure of any "baseline" may differ from the task. As such, noise may be nonstationary or nonergodic (Huppert, 2016; Tachtsidis & Scholkmann, 2016).

3.5.5 Statistical inference

After calculation of hemoglobin concentration changes, one usually estimates the HR by simple block averaging or linear regression models. Depending on experiment design and aim, one could employ other analytical techniques, e.g., functional connectivity or multivariate analysis.

3.5.5.1 Block averaging approaches

Block averaging consists of averaging signals across conditions before the application of classical statistical tests (e.g., t-test and ANOVAs) and amplitude-based methods. Block averaging avoids a priori assumptions about the shape of the HRF (Yücel et al., 2021), but does not utilize the high temporal resolution of fNIRS data (Tak & Ye, 2014). Block averaging assumes that channel positions are constant across participants, which is dependent on researchers' accurate placement of cap and individual brain anatomy, and thus most often not accurate (Tak & Ye, 2014).

3.5.5.2 GLM based approaches

The GLM assumes that data can be represented as a linear combination of several sources. Based on regression, it allows simultaneous modeling of different confounding variables and the HR of the signal. A less biased estimate of the HRF is provided by the GLM's simultaneous estimation of the contribution of all the components of an fNIRS signal. The entire fNIRS time-course is considered, providing more statistical power, but requiring assumptions about the shape and timing of the HRF. The HRF is typically modeled as a fixed canonical shape or a linear combination of multiple basis functions (Tak & Ye, 2014; Yücel et al., 2021).

3.5.5.3 Functional connectivity

Functional connectivity (FC) is defined as the temporal co-occurrence of spatially distant neurophysiological events (Eickhoff & Müller, 2015). I.e., two brain regions are considered functionally connected if there is a statistical relationship (correlation) between their activity measures (in fNIRS' case hemoglobin concentrations) (Eickhoff &

Müller, 2015). While a body of research investigates resting state FC (the spontaneous correlations of brain activity during rest detailed in section 3.4.3) (Lu et al., 2010; Tak & Ye, 2014), a growing number of studies are also investigating FC during cognitive or other tasks (Baker et al., 2018; Hu & Shealy, 2019)

Functional connectivity analysis of fNIRS data may be conducted in either the time domain—using correlation (e.g., Pearson's correlation) or causality models (e.g., Granger causality)—or in the frequency domain—using spectral coherence or phase-locking measurements (Santosa et al., 2017; Tak & Ye, 2014). The autocorrelative, colored noise structure of fNIRS data and the presence of motion artifacts challenges FC analysis, and must be addressed (Santosa et al., 2017). It is outside the scope of this thesis to provide a more comprehensive review of approaches to FC analysis in fNIRS.

3.5.5.4 The multiple comparisons problem

A selected statistical threshold (p-value), or confidence level, generally applies to individual statistical tests. The multiple comparisons problem arises when more than one statistical test is considered simultaneously, because it increases the chances of false-positives (type 1 error). Thus, if statistical inference is assessed based on multiple channels, regions, or network components, it should be adjusted to reflect that multiple statistical comparisons were made (Tak & Ye, 2014; Yücel et al., 2021). There are several approaches to control for multiple comparisons. Family-wise error rate (FWER) control adjusts the required threshold for significance in some way, and includes e.g., the Bonferroni correction, Least Significant Differences, and Tukey (Tak & Ye, 2014; Yücel et al., 2021). False discovery rate (FDR) controls the expected proportion of falsely declared-active detections among the total declared-active detections (Benjamini & Hochberg, 1995; Tak & Ye, 2014).

3.5.5.5 Multivariate analysis

There is a plethora of multivariate analysis techniques used for fNIRS analysis, classification, prediction, etc. Apart from the examples below, it is outside the scope of this thesis to provide a comprehensive review.

Partial Least Squares correlation has been used to examine brain-behavior associations. Specifically, significant correlations between fNIRS brain activity and task performance measures as a function of task difficulty were investigated (Meidenbauer et al., 2021), showing N-back level dependent changes in the relationship between HbR and task performance. The authors' demonstration of a load-performance interaction in recruitment of the prefrontal cortex suggests that the metabolic demand placed upon the prefrontal cortex to attain high performance varies depending on the level of difficulty.

Multivariate pattern analysis has been used to decode participants' (in this case, infants') minds. A correlation-based decoding method created group models of activation patterns across multiple fNIRS channels, accurately predicting activation patterns at the participant-, and individual trial- level (Emberson et al., 2017). The authors further found that multivariate analysis was able to distinguish between conditions where univariate methods could not, and that there was a difference in which channels were most informative depending on the stimuli type presented.

3.5.6 Comparing and selecting methods

3.5.6.1 Comparisons of analysis pipelines

Several works have compared the performance of various analysis pipelines. We list two of these works here.

Santosa et al. (2020) compared performance of different analysis pipelines[17] for removing systemic physiological noise, by using sensitivity-specificity analysis (also known as receiver operator characteristic (ROC) curves) on synthetic responses (truth) added to experimental data that featured resting state and a breath holding task. They found that using all available short-channels (8 in their case) as no-interest regressors in a mixed-effects variation of the GLM using autoregressive prewhitening followed by robust least squares regression (the AR-IRLS algorithm) best controlled type 1 errors; but concluded it would be best to use this approach without the mixed-effects variation as this increased computational time 10-fold (Santosa et al., 2020).

Fishburn et al. (2019) compared performance of different motion-correction algorithms[18] by using ROC curves and area under curve (AUC) for simulated data, and by creating t-statistic activation maps for real experimental data featuring an n-back task. They found that TDDR had greatest performance on simulated data (mean AUC value), followed by CBSI, tPCA, MARA, kWavelet, spline-SG, before the uncorrected signal. TDDR also performed best on experimental data, as evaluated by greatest t-statistic, followed by kWavelet, tPCA, spline-SG, CBSI, MARA, and as evaluated by positive significant $p<0.05$ values followed by spline-SG, kWavelet, MARA, CBSI, tPCA. (Fishburn et al., 2019). This illustrates that different algorithms perform better depending on evaluation metric and whether it is evaluated on real or simulated data.

The fNIRS community seems to converge on that using short-channels as regressors in some way[19], is the best way to correct for superficial systemic noise (Paranawithana et al., 2022; Santosa et al., 2020; Sato et al., 2016; Wyser et al., 2020).

3.5.6.2 Issues with comparing analysis pipelines

Despite convergence towards short-channels, several problematic issues exist when comparing and selecting an analysis pipeline. That none of the many different software and toolboxes includes all analysis approaches, makes it difficult to compare approaches across software/toolbox. No single work compares all pipelines. Moreover, the performance of any one pipeline is dependent on the data (and thus task and participants). While a pipeline might perform very well on a breath holding task as in Santosa et al. (2020), it might perform poorly on an eyes-closed resting-state as in Pinti et al. (2019). Furthermore, many pipeline performance comparisons are based on simulated or artificial data (Pinti et al., 2019; Santosa et al., 2020), which, obviously, is different from real data. We cannot be certain that any given pipeline will perform equally well on real and simulated data. Selection of techniques and whether it is included in a review depends on the application. Source codes are not available for all filtering techniques. The way performance is evaluated may differ between papers (Huang et al., 2022). Not all evaluation techniques can be applied to real experimental data (Huang et al., 2022). Finally, because there is a lack of standardized analysis pipelines and a vast

[17] The pipelines included both prefiltering and statistical approaches.

[18] The pipelines included only prefiltering approaches.

[19] i.e., different studies have differing number of regressors and additional nuisance regressors.

selection of methods with/without additional parameter selection, different results will be obtained depending on the selected method (Hocke et al., 2018; Pfeifer et al., 2018).

3.5.6.3 Our recommended preprocessing and analysis pipeline

Based on our experience analyzing fNIRS data in present contributions **(Dybvik et al., n.d., 2021a; Dybvik & Steinert, 2021)** and existing literature (Barker et al., 2016; Santosa et al., 2020), we recommend the following preprocessing and analysis pipeline for experiments where one seeks to obtain univariate estimation of concentration changes in response to a task/condition and whether those differ between tasks/conditions. We prefer the GLM framework because it is more flexible and requires fewer assumptions[20] than block averaging. We prefer statistical approaches to noise correction (as described in 3.5.3.2) because they appear to best correct violations of the GLM assumptions (as described in 3.5.4). For 1st-level statistics (i.e., participant level), we recommend using the GLM framework with the AR-IRLS algorithm (Barker et al., 2013, 2016), with short-channel measurements as nuisance regressors accounting for scalp physiology and motion, or accelerometer data as nuisance regressors accounting for motion. If one has both accelerometer data and short-channel measurements, we suggest conducting a ROC analysis to ascertain whether it is beneficial to use both as nuisance regressors, or if short-channel-regressors alone are better. 1st-level statistics should be fed to 2nd-level statistics (group model) that uses a robust mixed-effects model. In the NIRS Brain AnalyzIR toolbox (Santosa et al., 2018), this allows to e.g., control for age or include participant as a random variable in a relatively straight forward manner. We have used the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) to correct for multiple comparisons as this is implemented in the NIRS Brain AnalyzIR toolbox.

In the Tetris experiment **(Dybvik et al., n.d.)** for 1st level statistics we used a GLM with a canonical HRF to which we added accelerometer data as regressors and used the AR-IRLS algorithm. For 2nd level statistics we used a robust mixed-effects model. In our yoga study **(Dybvik & Steinert, 2021**) we also used a GLM with a canonical HRF and AR-IRLS to estimate 1st level statistics. 2nd level statistics employed a mixed-effects model. We used the Benjamini–Hochberg procedure for FDR correction in both studies **(Dybvik et al., 2021a; Dybvik & Steinert, 2021)**.

3.6 fNIRS in-situ

As described earlier (section 3.2.1) the low weight of wearable fNIRS devices, combined with relatively high robustness to motion artifacts, increased participant comfort and compliance, renders fNIRS a popular choice for in-situ and studies with high ecological validity. This section provides some examples of such fNIRS studies and discusses remaining challenges.

3.6.1 Examples fNIRS applications in-situ

Balardin et al. (2017) performed a series of proof-of-concept studies in unconstrained environments. This included an adult male playing table tennis, a female professional pianist with 45 years of regular practice playing the piano, human interaction within a professional violin duo, and continuous monitoring (4 hours) of a participant during their daily activities. The authors found which motor cortex regions were more active in an unpredictable table tennis condition compared to a predictable table tennis condition.

[20] assumptions that may be inaccurate

Polyrhythmic piano playing required greater cognitive effort and sustained attention, as evidenced by PFC activation, compared to simple rhythmic playing. Regarding the violin duo's brain synchronization, the authors found higher synchronization in parietal and frontal regions than in premotor and somatomotor areas. Continuous monitoring used FC analysis, but the authors did provide any conclusive findings.

Xu et al. (2019) used a one-channel fNIRS to investigate prefrontal cortex hemodynamics of anesthesiologists in operating rooms during a high-fidelity simulation-based training of crisis event management. The authors found higher levels of neural synchrony in the team phase than in the single provider phase of the simulation scenario. Moreover, neural synchrony level was sensitive to scenario difficulty, as more difficult scenarios exhibited increased HbO_2 synchrony. This implies that more difficult simulation scenarios require closer team cooperation than less difficult scenarios, yielding higher team engagement.

Tachibana et al. (2011) studied participants playing a dance video game, similar to Dance Dance Revolution^{TM} (Konami Corp., Tokyo, Japan), in three different stepping conditions ranging from more to less difficult (conditions included 1) using all 4 arrows for stepping, 2) using 2 arrows for stepping, 3) rhythmic stepping to music). The greatest HbO_2 levels were found in the most difficult dancing condition and the authors suggest that the quick-onset, trapezoidal-shaped HbO_2 time-course reflects the on-off response of spatial orientation and rhythmic motor control.

3.6.2 In-situ studies conducted by the author

We have demonstrated multimodal neuroimaging, notably EEG+fNIRS, in three real-world cases **(Dybvik et al., 2021a**). The first case included a conventional laptop setup, which is frequently used in experimental design research (e.g., in **Dybvik et al. (2023, 2022)**). The second case consisted of driving in a city and on the highway. Here, noise from muscle movement and ocular artifacts increased compared to the conventional setup, but we could detect differences in cognitive load between different driving conditions (highway vs. city). Furthermore, we tested battery life and was able to record 66 minutes of continuous data from EEG, fNIRS and one web camera connected to a laptop. Our successful collection of multimodal neuroimaging data in these two use cases is novel in design research. The third use case involved Ashtanga Vinyasa Yoga. As a precursor to the article described below **(Dybvik & Steinert, 2021)**), we provided a visual inspection of overall trends in the raw data. Although we have used multimodal brain imaging with success in-situ, we have yet to apply it to a full-scale design experiment.

In one study **(Dybvik & Steinert, 2021)** we further aimed to extend fNIRS applications to real-world environments. We conducted a single-subject study of a vigorous yoga practice (Ashtanga Vinyasa yoga). The study entailed seven sessions of yoga practiced in the participant's own apartment (Figure 12). To our knowledge, it was the first demonstration of fNIRS measurements recorded during any moving yoga practice in a real-world setting. We investigated whether there were differences in cognitive load and cognitive state between different postures, and we reported both HbO_2 and HbR. Although this study exemplifies usage of neuroimaging measurements in the real world, exploring the boundaries of real-world cognitive load measurements, it has limitations. The analysis could have better corrected for noise from systemic physiology[21] and the

[21] One session recorded ECG data although the article doesn't mention it.

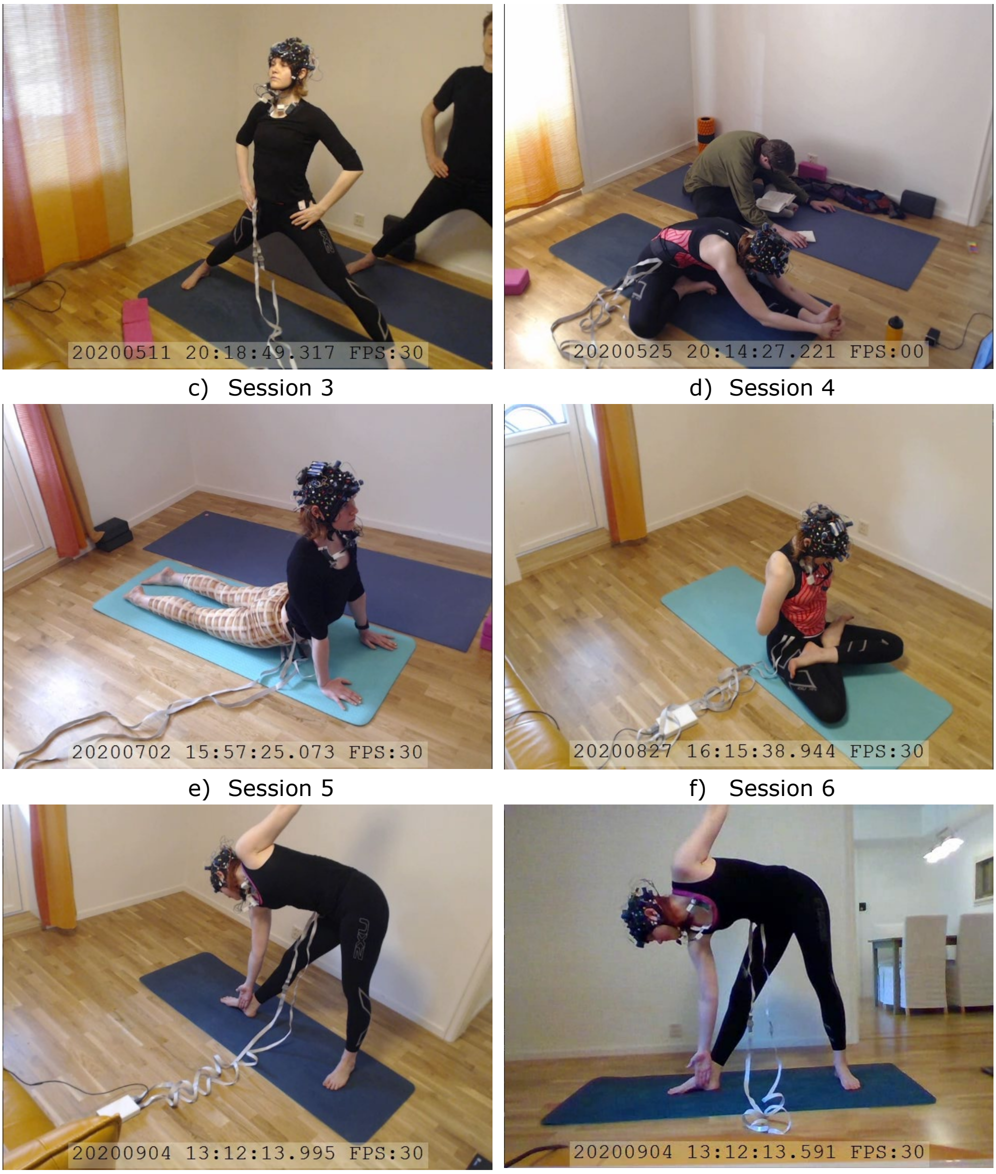
effects of gravitational changes[22]. If we were to redo the study, we would inspect main effects rather than contrasts. There was no signal quality check conducted after data collection, apart from visual inspection, which revealed motion artifacts across all channels consistent with movement in the yoga practice. Finally, not all postures were conducted as it would interfere with the fNIRS system (e.g., shoulder stand).

Together, our findings showcase that it is viable to transport fNIRS into the field, whether the field is a moving vehicle **(Dybvik et al., 2021a)** or a private apartment **(Dybvik & Steinert, 2021)**, and conduct studies where participants are moving. Compared to our studies, the environments and human movements of design research experiments are generally less demanding. This implies that it is possible to conduct design research experiments in-situ, e.g., studying established design teams and how they work on an everyday basis by transporting fNIRS to their workplace. Longitudinal studies, monitoring design teams develop over time, are possible, as was done in the yoga study **(Dybvik & Steinert, 2021)**. Testing and evaluating designed products, services, and interactions could also be conducted in-situ. E.g., if one designs new yoga mats, these could be brought to users' homes and tested there if the user usually practices at home. In this case, an at-home evaluation is likely more valid than an evaluation conducted in a design studio, because this is where the designed artifact would be used on an everyday basis. Further, if one designs yoga as a service or interaction, there are many elements, apart from the yoga mat itself, one must account for. The interior design, music and sounds, scents, lighting, temperature, and of course the instructor, all interact, creating one composite experience. These elements may be composed differently depending on the type of yoga (e.g., Ashtanga, Hatha, Kundalini, Yin, Bikram or hot yoga) and which yoga experience the service provider seeks to offer. The yogi's evaluation of the yoga service will be based on the joint effect of all these elements as well as other yogis attending the practice. As such, one cannot test these elements in isolation, and the yoga experience as a whole cannot be easily replicated in a design studio. It is best to test yoga as a service in-situ, e.g., in a yoga studio.

a) Session 1

b) Session 2

[22] We would have liked to have accelerometer data and explore using it as a regressor.


20200511 20:18:49.317 FPS:30
20200525 20:14:27.221 FPS:00
20200702 15:57:25.073 FPS:30
20200827 16:15:38.944 FPS:30
20200904 13:12:13.995 FPS:30
20200904 13:12:13.591 FPS:30

c) Session 3

d) Session 4

e) Session 5

f) Session 6

g) Session 7 (Web camera 1)

h) Session 7 (Web camera 2)

Figure 12 Photos from web camera recording of Ashtanga Vinyasa Yoga practices. A total of 7 sessions were recorded, illustrated in a) through g). Two web cameras were used, illustrated in g) and h).

Finally, we provide a demonstration of a multimodal (fNIRS and EEG) in-situ experiment in design research **(Aalto et al., Under review)**. In this proof-of-concept pilot experiment, we record multimodal fNIRS, EEG, and a web camera. We took a 1.5 hour walk through Nidaros Cathedral, which was closed to the public, see Figure 13. The walk included distinctly different spaces, such as the nave, catacombs, built-in-wall stairwells, secluded spaces of worship, and the rooftop. The analysis will, among other aims, consider the effect of large, open spaces compared to smaller and narrow spaces on hemodynamic activity in the prefrontal cortex. We show that it is feasible to examine architectural space sequences with fNIRS.

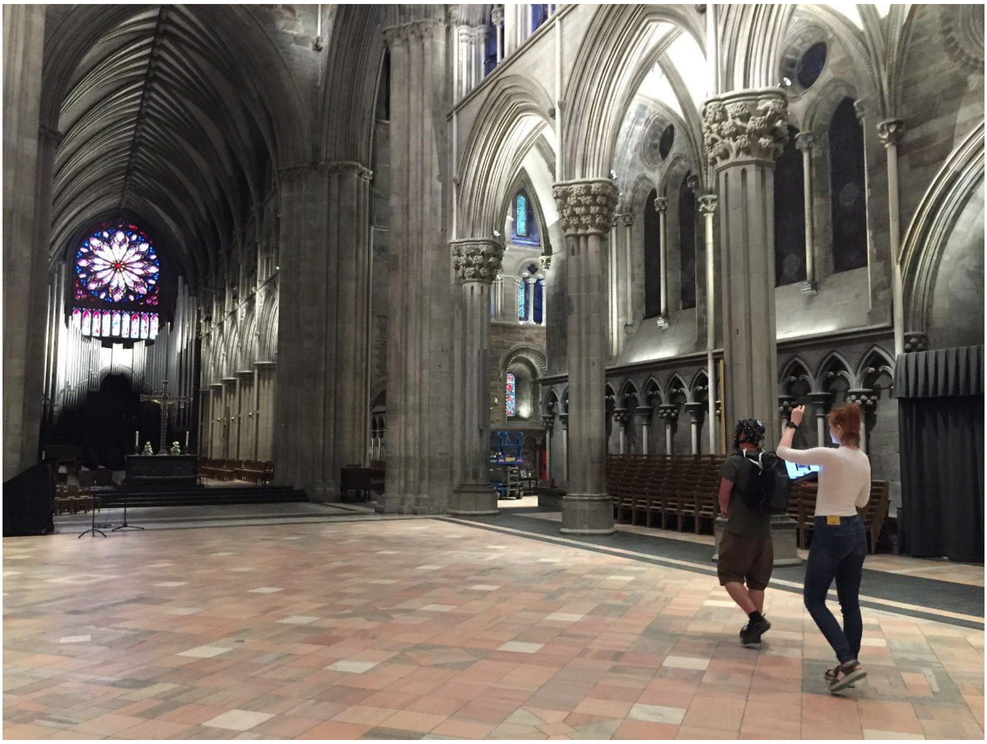


Figure 13 The stroll through Nidaros Cathedral. [Photo taken by Kristin Bjørlykke]

3.6.3 Remaining limitations in-situ

In-situ studies are being conducted. Nevertheless, they are limited by a set of remaining challenges that concern the complex factors in real-world data collection. In-situ will bring greater motions artefacts and larger changes in systemic physiology. We expect greater presence of uncontrolled changes in e.g., blood pressure, respiration, and gravitational pulling in unconstrained environments; we must better understand the impact those factors have on fNIRS' signal (Balardin et al., 2017; Pinti et al., 2015). In this undertaking, it becomes more important to simultaneously measure systemic physiology (von Lühmann et al., 2021). If recording multimodal EEG and fNIRS in-situ the level of noise in EEG data will increase **(Dybvik et al., 2021a)**. This could be mitigated by marking movements as captured by video and remove or correct this data segment **(Dybvik et al., 2021a)**. Timing and labeling of real-world events can be difficult (Hu & Shealy, 2019; Pinti et al., 2017), however, it is crucial to synchronize multimodal data and the stimuli's timing in the environment (von Lühmann et al., 2021). Concurrent video recordings could be used to label events (**Aalto et al., Under review; Dybvik et al., 2021a; Dybvik & Steinert, 2021**; Pinti et al., 2015), as could analytical approaches (Pinti et al., 2017). Bandwidth, battery life, optode weight, and participant comfort need to be further improved (von Lühmann et al., 2021). In our driving case demonstration we tested battery life of a laptop that recorded EEG, fNIRS, and video from an external web camera, and we managed to record 66 minutes of data **(Dybvik et**

al., 2021a). Of course, it will be difficult, if not impossible, to control all external factors in-situ—thus, analysis methods that handle real-world motion artifacts, changes in systemic physiology, and functional events ought to be developed[23].

3.7 fNIRS in design research

Given fNIRS' characteristics and the advantages presented in 3.2.1—notably, an objective measure of brain activity, capable of usage in-situ—fNIRS presents itself as a viable and promising technique for design research (Balters et al., 2023; Gero & Milovanovic, 2020; Hu & Shepley, 2022; Hu & Shealy, 2019). In this section, we provide examples of design research with fNIRS, followed by an overview of review articles of fNIRS in design research. fNIRS hyperscanning in design research is outlined, before we focus on two single studies and review their limitations. A discussion on avenues for future research close the section.

The design research community has used fNIRS for researching several broader research categories. These include recognized design methods and tools, with studies of sketching (Kato et al., 2018; Milovanovic et al., 2020) and concept generation techniques at various levels of structure (TRIZ, brainstorming, and morphological analysis) (Hu et al., 2018; Shealy et al., 2018, 2020b). The structuredness of concept generation techniques impacts designers' brain connectivity (Hu et al., 2018). Moreover, concept mapping helps students generate more concepts while requiring less cognitive load (Hu & Shealy, 2018). Research has found an effect of educational design training and experience (Hu et al., 2021; Shealy et al., 2017), and began to explore fNIRS neurofeedback (Shealy et al., 2020a). Although design team interaction, creative collaboration (Mayseless et al., 2019), and leader-follower interactions in teams (Jiang et al., 2015) have been researched, social interaction studies in design are still in their infancy.

3.7.1 Review articles of fNIRS in design research

Five recent articles have reviewed fNIRS studies in design research, three of which followed PRISMA (Balters et al., 2023; Hu & Shealy, 2019; Ohashi et al., 2022). Hu & Shealy (2019) focused on fNIRS in engineering decision-making and design cognition. Balters et al. (2023) focused on design cognition studies, covering only studies involving designers and design processes, but they included fMRI and EEG studies in addition to fNIRS. Borgianni & Maccioni (2020) included a broad range of biometric sensors, reviewing both design processes (measuring designers) and product evaluation (measuring users) in engineering design. Hu & Shepley (2022) review EEG, fMRI, and fNIRS design research, describe the technical principles, but does not go into detail on individual fNIRS studies. Ohashi et al. (2022) identified one fNIRS article, whose experiment we scrutinize further in section 3.7.3. Thus, we summarize the findings from the three first reviews here.

Hu & Shealy (2019) identified 32 articles with a mean sample size of 28 (*SD* 23), most of which used overly simplistic tasks[24] rather than real-world problems. Block designs with blocks ranging from 5-10 seconds to 60-120 seconds were most frequently employed, even though block design doesn't lend itself well to real-world applications. Mixed-designs might represent the real world better, but requires more assumptions to estimate HR. Behavioral data (questionnaires, task performance, response time, error

[23] Other researchers agree with this, e.g., (von Lühmann et al., 2021).
[24] a game or a dilemma

rate, eye-fixation information) was often collected in addition to fNIRS, which most often focused on the PFC. Most studies reported only HbO_2, while a few reported HbO_2 and HbR, or HbT. While the statistical inference methods used vary greatly, older averaging techniques dominate, with averaging conducted either across subjects, conditions, channels, functional areas, or probabilistic brain atlases. The authors highlight that there is no consensus about the validity of these averaging techniques, as they are limited by requiring researchers to subjectively define activation periods, and ignore information related to the HRF's shape or time-course of the data. Averaging techniques remove, or mask, potentially great individual variance; this may or may not amplify group similarities.

Balters et al. (2023) identified 18 fNIRS studies[25] with a sample size ranging from 10-32 participants, most of which focused on the PFC using 16-22 channel systems. No study used multi-modal brain imaging or short-channels. There was a large methodological variation between studies. 12 of 17 fNIRS studies included GLM and activation analysis, four included FC analyses, four included graph-theory based network analyses only. Problematically, not all studies described the data processing approach sufficiently, nor reported statistical results properly, i.e., to a degree that allows meta-analyses of results to establish generalizable results. Participant cohorts varied greatly in population adherence, sample size, and demographics. Most studies comprised students both with and without engineering or design experience. While remaining studies comprised non-students, experts, or working professionals, not all specified their levels of expertise. Moreover, the definition of expertise or expert was inconsistent. Some studies lack participant demographics; some did not assess[26] whether participants used medication, had any neurological or psychiatric conditions, or a history of substance abuse; all of which could affect typical brain function. This is a severe limitation.

Both Hu & Shealy and Balters et al. suggest that current analysis approaches are too simple, and that future research should consider more advanced analyses (Balters et al., 2023) and sophisticated statistical techniques (Hu & Shealy, 2019).

Borgianni & Maccioni (2020) find fragmented use and interpretation of sensors. Large variations of deducted variables are present. Not all studies used, or described their usage of, statistical methods. Instead, some provided qualitative analyses. Contradictory findings still occur. Sample sizes are limited to tens. No in-situ study was identified, and the number of studies involving industry was negligible. No product evaluation study conducted with fNIRS has been identified[27]. It must be noted that the review is limited by using a subjective snowballing process combined with one literature search in Scopus to identify literature.

3.7.2 fNIRS hyperscanning in design research

Hyperscanning is particularly relevant to design. One study investigated collaborative social interaction between two individuals, specifically: an open-ended creative design task was compared to a goal-oriented 3D-building task that followed instructions (Li et al., 2021; Mayseless et al., 2019). Overall, the design task exhibited stronger IBS between several regional connections compared to the 3D-building task (Mayseless et al.,

[25] (22%) of a total of 82 studies, most of which used EEG (62%) while some used fMRI (16%)

[26] or report that they assessed

[27] although publications in 2019 and onward weren't considered.

2019). The authors further identified transient occurrence of, and sequential shifts between, several distinct "states" of dynamic interbrain synchrony (IBS) during the tasks. Interestingly, although both tasks involved cooperation, the dynamic IBS states in the design task differed from the building task. Specifically, connections between brain regions in the building task tended to be less condense compared to those of the design task. The global efficiency[28] of several dynamic IBS states in the design task was positively correlated with cooperation performance, but none of the 3D-building states were. This could signify which states are more central to cooperative creative design tasks (Li et al., 2021). The sequential shifts between dynamic IBS stats could explain shifts in communication patters and cooperation between individuals in a creative design task (Li et al., 2021).

> we suggest that successful team cooperation leading to creative ideas relies on cyclic back and forth interactions between cognitive control and socio-emotional processes (including experiencing what the other is feeling and inferring emotions and intentions) (Mayseless et al., 2019, p. 8)

Collaboration, competition, and trust are examples of social interactions important to understand in the context of design (Balters et al., 2021) as research has demonstrated that team emotional intelligence cultivates trust, group- identity, and efficacy, yielding enhanced team performance and creativity. Another hyperscanning study is currently investigating the warm-up games/activities that aim at increasing team trust and collaboration frequently employed before engaging in collaborative design tasks as a team. Teams' creativity outcome (e.g., product/innovation) is hypothesized to increase when preceded by an interpersonal trust activity (Balters et al., 2022). We look forward to the published results and implications for design.

> The future of fNIRS hyperscanning is limitless and very well may be a key component of our understanding of the neurobiological underpinnings of social behavior (Balters et al., 2020).

3.7.3 Limitations of selected design studies

Having summarized reviews of fNIRS design research and associated limitations, we provide more detail here, focusing on a few studies only.

Shealy et al. (2020b) investigated the effect of using three well-known and distinct concept generation techniques: brainstorming, morphological analysis, and TRIZ on PFC hemodynamic activity amongst 30 first-year graduate students[29].

Their analysis pipeline used Shimandzu fNIRS software[30], employing a bandpass filter [0.01-0.1 Hz] to remove high-frequency instrumental noise and low-frequency physiological noise, and an ICA with a 0.5 spatial uniformity coefficient to remove motion artifacts. Baseline correction was performed by subtracting mean HbO_2 from the baseline rest period from the HbO_2 time-series during the task for each channel. Afterwards, averaging techniques are employed—e.g., mean value across the entire condition was used as a proxy for cognitive activation—followed by classical statistical tests. The authors acknowledge the limitations associated with these averaging techniques, notably that they ignore individual differences. The authors suggest using sliding windows to better capture temporal dynamics. Other mentioned limitations include: only

[28] a measure of how actively and efficiently a network can share information between regions

[29] 3 of which had to be excluded due to a weak signal

[30] which presumably came with the 8x8 fNIRS device and is proprietary

investigating the PFC though other brain regions might also be involved[31], the analysis did not account for possible differences in the students' design outcomes[32], sample size, and that the many variations of the techniques and associated tools may lead to varying outcomes. Finally, it is suggested to investigate the difference between novice and expert.

Although the authors mention that removing noise and motion artifacts is critical to avoid false-positives, there is no mention of FDR corrections or other multiple comparisons corrections in the statistical analysis. This is expected when performing statistical analysis on several channels, and is vital for reducing the risk of false-positives (Tak & Ye, 2014; Yücel et al., 2021). Additional limitations associated with block averaging are that the entire duration of conditions was averaged, but possible plateau effects of the HRF were not discussed. Possible explanations for their findings are offered based on reverse inference. That none of the students were familiar with TRIZ might have affected the results.

FC analysis used Pearson's correlation metrics for each task, averaged across participants. Correlations greater than a set of threshold coefficients (0.6-0.7) were considered indicative of FC between nodes. The authors show that FC maps for the two thresholds differ greatly. Although a literature-based threshold was applied, FC maps for the two thresholds are reported without providing any specific reason for why they did so. We speculate that this could be because fNIRS is a relatively new field. Each condition was additionally segmented into 20 equal, non-overlapping segments. A FC network was calculated for each segment. Network centrality and density were calculated to provide descriptive measures of the network. Node centrality was shown to vary, but no significance test appears to have been conducted.

The Pearson's correlation approach used for FC analysis is problematic because it has been demonstrated that even in absence of spatially global systemic physiology the autocorrelative structure of the hemodynamic signal results in >70% FDR rates. In a comparison of correlation models, Pearson's correlation performed among the worst[33]. Without controlling for these serial correlations, this means that the p-value reported by the standard correlation method is completely inaccurate, in that it's wrong 70% to 85% of the time when displayed at a threshold of $p < 0.05$. Note further that this is not a multiple comparisons issue since this inaccuracy applies to a single statistical test (Santosa et al., 2017). Uncontrolled type 1 error (false-positives) in the standard analysis methods for FC analysis in resting state fNIRS data is a well-known problem (Santosa et al., 2017). Moreover, it is problematic that the article does not state whether only significant correlations, or all correlations regardless of significance level, were considered.

A hyperscanning study (Mayseless et al., 2019) investigated inter-brain synchrony of two individuals while designing "a product that would motivate people to vote". Although the authors state it is an exploratory investigation of naturalistic creative teamwork, the 10-minute design task was compared with a 3D-model building task following instructions. Despite excellent choice of tasks, both were completed in a white-room laboratory setting. Participants were not acquainted prior to the study, nor did the article state whether participants had relevant design education. The study could arguably have been

[31] This is an fNIRS inherent limitation when whole-head montage is unavailable.
[32] e.g., more compared to less novel designs.
[33] only a lowpass filtered signal performed slightly worse.

more naturalistic by recruiting designers who were acquainted, and conducting it in a design studio.

3.7.4 Avenues for future research

fNIRS allow us to look into the black box of humans in experimental design research. fNIRS in design research is a young and upcoming field. There are not many experiments, Hu & Shealy (2019) identified 32 decision-making articles, Balters et al. (2023) identified 18 design cognition articles. The latter 18 articles appear to consist of 6-8 unique experiment designs. It seems researchers are starting to learn how to use fNIRS, as evidenced by the employment of simpler and older processing and statistical techniques, among other limitations.

Overall, limitations of design studies using fNIRS include a general bias towards the PFC—not surprising as it is easier to obtain high-quality fNIRS recordings from this location. No study appears to include measures of systemic physiology nor other sensor modalities. None fully follow best practices in fNIRS publications (Yücel et al., 2021).

Future fNIRS research would benefit from short-channels and other technological advances, although there is a tradeoff: while whole-head coverage could aid in obtaining holistic cortical mechanisms underlying design activities, small and ultra-portable montages and systems allow longitudinal and in-situ studies (Balters et al., 2023). Future research will benefit from triangulating fNIRS with measures of systemic physiology and behavioral measures. Further, environmental (ambient noise and light, room infrastructure etc.) and technical (e.g., computer frame rate) information ought to be monitored and included to complete our understanding of human behavior during design (Balters et al., 2023; Borgianni & Maccioni, 2020). No individual neuroimaging tool is perfect, necessitating multimodal neuroimaging studies for a comprehensive understanding of the neurological underpinnings of design (Balters et al., 2023; **Dybvik et al., 2021b**; Hu & Shepley, 2022).

Several researchers, us included, suggest avenues for future fNIRS research (Balters et al., 2023; Borgianni & Maccioni, 2020; **Dybvik et al., 2021a, 2021b**; Hu & Shepley, 2022; Hu & Shealy, 2019; Mayseless et al., 2019). We believe there is a great unexploited potential in conducting product evaluation studies with fNIRS, considering that no product evaluation study yet has used fNIRS (Borgianni & Maccioni, 2020). Design teams composed of three or more individuals should be investigated, because it captures more complex social elements of teamwork (Mayseless et al., 2019), and likely better represent real-life design teams. We would have found it interesting to investigate an established design team of practicing designers in a company working in their normal environment, i.e., in-situ. The built environment and architecture could benefit from fNIRS neuroimaging, e.g., informing designers about human responses to built environment, how building shape, ceiling height, lighting, object placements, acoustics, etc., impacts well-being (Hu & Shepley, 2022). Future in-situ fNIRS studies could benefit from using eye-tracking to determine relevant time intervals for analysis[34]. Qualitative, explorative fNIRS studies could be a good first step to investigating user-product interactions, services, user experience, and other sensorial experiences in-situ.

[34] similar to a study highlighted by Borgianni & Maccioni (2020) which used eye-tracking in combination with EEG to determine relevant time intervals for analysis.

“The important thing is not to stop questioning. Curiosity has its own reason for existence. One cannot help but be in awe when [s]he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery each day”

Albert Einstein[35]

[35] In Old Man’s Advice to Youth: ’Never Lose a Holy Curiosity. *LIFE Magazine*, 1955, May 2, pg. 64.

4 Summary, discussion, and outlook

4.1 Contributions and limitations

This thesis's contributions to introducing fNIRS to multimodal in-situ experiments in design research are two-fold, relating to foundational aspects (articles pertaining to Chapter 2) and fNIRS-specific aspects (articles pertaining to Chapter 3) of experimental design research. Relating to foundational aspects, we have 1) detailed how one might construct hypotheses for design experiments from qualitative inquiry **(Dybvik & Steinert, Under review)**, 2) established objective physiological measurements corresponding to the subjective constructs of workload and stress in an engineering design product evaluation context **(Wulvik et al., 2019)**, and 3) replicated and extended a design experiment involving a well-researched design topic **(Dybvik et al., 2023, 2022)**. Relating to fNIRS specifically, we have 1) demonstrated fNIRS usage in two in-situ scenarios and in more traditional design contexts **(Dybvik et al., 2021a)**, 2) conducted a full-scale in-situ fNIRS study **(Dybvik & Steinert, 2021)**, 3) lowered the barrier to including multiple modalities in design experiments by proposing a low-cost EEG+fNIRS sensor setup **(Dybvik et al., 2021b)**, and 4) explored use of multiple modalities in conjunction with fNIRS **(Dybvik et al., n.d.)**.

In summary, we have all the necessary components for attaining this thesis's objective: multimodal data capture, an fNIRS study conducted in-situ, an experiment involving a well-researched design topic, and a proof-of-concept of a multimodal in-situ design research experiment. Taken together, we conclude it is possible to use fNIRS in multimodal, in-situ experiments—and, that such studies will greatly improve design research.

Of course, there are limitations associated with all articles, and the contributions to the thesis' overall objective. For limitations pertaining to any one article in particular, the reader is referred to the article itself. Regarding the limitations of this thesis, the greatest drawback is the absence of a full-scale multimodal in-situ design experiment. Although in-situ studies have been demonstrated [studies **(Dybvik et al., 2021a; Dybvik & Steinert, 2021)**], they lack sufficient signal-quality check post data collection. Noise from systemic physiology and motion artifacts could have been more meticulously inspected and corrected for. Short-channels were unavailable but desired. The studies **(Dybvik et al., 2021a; Dybvik & Steinert, 2021)** are case studies with a small sample size, which, while capable of demonstrating existence, have non-generalizable results. While one experiment (N=30) collected and analyzed multiple sensor modalities **(Dybvik et al., n.d.)**, it laid contextually outside design. This thesis, could, as always, have been better and shorter, while encompassing more.

4.2 fNIRS in design research

Design research has used fNIRS for researching design methods and tools, e.g., sketching (Kato et al., 2018; Milovanovic et al., 2020) and concept generation techniques at various structural levels (Hu et al., 2018; Shealy et al., 2018, 2020b); the effect of educational design training and experience (Hu et al., 2021; Shealy et al., 2017); design ideation aided by sketching (Milovanovic et al., 2020); fNIRS neurofeedback (Shealy et al., 2020a); and, design team interaction and collaboration (Mayseless et al., 2019).

Experimental design research with fNIRS has several limitations and remaining milestones before attaining the objective of conducting multimodal in-situ studies. First, to our knowledge, there exists no design experiment collecting systemic measurements and multimodal brain imaging (e.g., EEG+fNIRS), nor an in-situ design experiment with fNIRS. The design studies that do record fNIRS have several limitations and associated aspects needing rectification and quality elevation. Accurate reporting of statistical results and full disclosure of data processing pipeline is required to assess the validity of results and address potential methodological issues that could restrict generalizability. Complete, accurate reporting is necessary for meta-analysis of experimental design research to collect and connect theories and establish generalizable results (Balters et al., 2023). Detailed descriptions of experimental procedure are needed for reproducibility, benefitting future research. To further facilitate open science, transparency, and generalizable results, we recommend sharing data sets and code, and using open-access analytical tools. Mass univariate statistical methods are often used to evaluate multiple dependent variables, when multivariate statistical approaches ought to be used. Future research should place greater emphasis on multivariate techniques.

Participant samples ought to be of sufficient sizes and population representative. Student participants are still far overused, while experiments involving practicing designers in industry are needed. Moreover, participants in fNIRS studies must be screened for neurological and psychiatric conditions, use of medications, history of drug or alcohol usage, or other aspects that could affect typical brain function. Not all current studies complete, or declare completion of, such a screening (Balters et al., 2023).

Future design research must consider several aspects of fNIRS analysis. It is necessary to develop and validate analysis methods specifically for design research. Although such research efforts exist—e.g., focusing on developing analyses that allow greater temporal details of design activities (Balters et al., 2023; Li et al., 2021)—they are scarce in numbers. The fNIRS community experience fractioned analytical practices (Yücel et al., 2021), which is true for fNIRS design research as well (Balters et al., 2023; Hu & Shealy, 2019). The lack of a standardized fNIRS analysis pipeline is problematic for design research as it could lead to misunderstandings, and poor quality, irreproducible studies (Hocke et al., 2018; Pinti et al., 2020). Without a focused initiative on standardizing fNIRS analysis pipelines, we foresee a greater dispersion of analytical approaches in fNIRS design research, challenging the field.

fNIRS short-channel technology development will increase design research’s ability to fully utilize fNIRS in-situ and naturalistic settings (Balters et al., 2023). To our knowledge, none of the studies reviewed by Balters et al. (2023) had included short-channels. Future research ought to include short-channels whenever possible.

4.3 Outlook and thoughts on future research

This section expands upon section 3.7.4. Current fNIRS studies in neuroscience are not related to realistic design settings, and fNIRS design studies are concerned with simple aids and processes. Realistic scenarios are needed, both when investigating designers and users' interactions with designed artifacts. We are currently exploring the effects of real architecture on humans as they experience the space **(Aalto et al., Under review)**.

fNIRS hyperscanning will crucially aid design research because it enables investigation of team interactions in-situ. Because much understanding and methods in design rely on the notion that groups of individuals working together on a design problem more effectively create better solutions and superior products (Balters et al., 2023; Mayseless et al., 2019; Stempfle & Badke-Schaub, 2002), we need additional hyperscanning research that addresses this topic efficiently. E.g., unraveling the neural dynamics of established high-performing design teams. We need studies of design teams of three or more people in their everyday work environment to accurately ascertain the neural underpinnings of real-life design. Further, we need longitudinal studies of design teams, reflecting that most products are not taken from initial idea to finished product within the timeframe of an experimental session. Longitudinal studies that compare multiple design teams over time would allow us to infer how team composition, dynamics, and interactions affect the design process and whether that relates to the designed outcome's success.

It will be crucial to relate team dynamics to design outcome, because this will help us understand how to design and develop the products, services, interactions, and experiences that most effectively solve real problems. Ultimately, we want to reverse-engineer the design process of the very best technological solutions: predicting the success of a design outcome based on team interactions at the earliest stages of the design process—and establish quantitative measures of team interactions that can be used to steer the development process in the right direction, ensuring a successful outcome from the very beginning. We believe such research will contribute to developing better solutions serving the current needs of the planet.

Current hyperscanning uses one neuroimaging modality. Multimodal studies, involving multiple neuroimaging modalities (e.g., both fNIRS and EEG) and physiology sensors, are necessary to triangulate brain-body interactions on the individual and team level. We term the recording of multiple physiology and neuroimaging modalities: ***heightened hyperscanning***. To reach heightened hyperscanning's full potential in-situ, we must continue to improve hardware, analysis methods, and experimental designs. We must continue to reduce weight and size, while increasing participant comfort and the optodes' head coverage. No modality should be considered in isolation in any data analysis when it was recorded along with other modalities. Of course, multimodal data warrant advances in multivariate data analysis. Together, we believe this will enable a shift in experimental paradigms—from laboratory to in-situ—that allows accurate results, directly generalizable to the real world.

References

Aalto, P., Dybvik, H., & Steinert, M. (Under review). A stroll through a cathedral: FNIRS and space sequences in architecture. *Frontiers in Neuroergonomics*.

Aasted, C. M., Yücel, M. A., Cooper, R. J., Dubb, J., Tsuzuki, D., Becerra, L., Petkov, M. P., Borsook, D., Dan, I., & Boas, D. A. (2015). Anatomical guidance for functional near-infrared spectroscopy: AtlasViewer tutorial. *Neurophotonics*, *2*(2). https://doi.org/10.1117/1.NPh.2.2.020801

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). The Variability of Human, BOLD Hemodynamic Responses. *NeuroImage*, *8*(4), 360–369. https://doi.org/10.1006/nimg.1998.0369

Ahn, S., & Jun, S. C. (2017). Multi-Modal Integration of EEG-fNIRS for Brain-Computer Interfaces – Current Limitations and Future Directions. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00503

Ahn, S., Nguyen, T., Jang, H., Kim, J. G., & Jun, S. C. (2016). Exploring Neuro-Physiological Correlates of Drivers' Mental Fatigue Caused by Sleep Deprivation Using Simultaneous EEG, ECG, and fNIRS Data. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00219

Al-Shargie, F., Kiguchi, M., Badruddin, N., Dass, S. C., Hani, A. F. M., & Tang, T. B. (2016). Mental stress assessment using simultaneous measurement of EEG and fNIRS. *Biomedical Optics Express*, *7*(10), 3882. https://doi.org/10.1364/BOE.7.003882

APA Dictionary of Psychology. (n.d.). Retrieved February 8, 2023, from https://dictionary.apa.org/

Auernhammer, J., & Roth, B. (n.d.). *What is Design Thinking?* https://www.researchgate.net/publication/365926664_What_is_Design_Thinking

Auernhammer, J., & Roth, B. (2021). The origin and evolution of Stanford University's design thinking: From product design to design thinking in innovation management. *Journal of Product Innovation Management*, *38*(6), 623–644. https://doi.org/10.1111/jpim.12594

Babiloni, F., & Astolfi, L. (2014). Social neuroscience and hyperscanning techniques: Past, present and future. *Neuroscience & Biobehavioral Reviews*, *44*, 76–93. https://doi.org/10.1016/j.neubiorev.2012.07.006

Baker, J. M., Bruno, J. L., Gundran, A., Hosseini, S. M. H., & Reiss, A. L. (2018). FNIRS measurement of cortical activation and functional connectivity during a visuospatial working memory task. *PLOS ONE*, *13*(8), e0201486. https://doi.org/10.1371/journal.pone.0201486

Balardin, J. B., Zimeo Morais, G. A., Furucho, R. A., Trambaiolli, L., Vanzella, P., Biazoli, C. J., & Sato, J. R. (2017). Imaging Brain Function with Functional Near-Infrared Spectroscopy in Unconstrained Environments. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00258

Balters, S., Baker, J. M., Hawthorne, G., & Reiss, A. L. (2020). Capturing Human Interaction in the Virtual Age: A Perspective on the Future of fNIRS Hyperscanning. *Frontiers in Human Neuroscience*, *14*, 458. https://doi.org/10.3389/fnhum.2020.588494

Balters, S., Mayseless, N., Hawthorne, G., & Reiss, A. L. (2021). The Neuroscience of Team Cooperation Versus Team Collaboration. In C. Meinel & L. Leifer (Eds.), *Design Thinking Research: Interrogating the Doing* (pp. 203–217). Springer International Publishing. https://doi.org/10.1007/978-3-030-62037-0_9

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

Balters, S., Weinstein, T. J., Hawthorne, G., & Reiss, A. L. (2022). Interpersonal Trust Activity to Increase Team Creativity Outcome: An fNIRS Hyperscanning Approach. In C. Meinel & L. Leifer (Eds.), *Design Thinking Research: Achieving Real Innovation* (pp. 19–36). Springer International Publishing. https://doi.org/10.1007/978-3-031-09297-8_2

Balters, S., Weinstein, T., Mayseless, N., Auernhammer, J., Hawthorne, G., Steinert, M., Meinel, C., Leifer, L. J., & Reiss, A. L. (2023). Design science and neuroscience: A systematic review of the emergent field of Design Neurocognition. *Design Studies*, *84*, 101148. https://doi.org/10.1016/j.destud.2022.101148

Barker, J. W., Aarabi, A., & Huppert, T. J. (2013). Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS. *Biomedical Optics Express*, *4*(8), 1366. https://doi.org/10.1364/BOE.4.001366

Barker, J. W., Rosso, A. L., Sparto, P. J., & Huppert, T. J. (2016). Correction of motion artifacts and serial correlations for real-time functional near-infrared spectroscopy. *Neurophotonics*, *3*(3), 031410. https://doi.org/10.1117/1.NPh.3.3.031410

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Blessing, L. T., & Chakrabarti, A. (2009). *DRM, a design research methodology*. Springer Science & Business Media.

Bordens, K. S., & Abbott, B. B. (2016). *Research design and methods: A process approach* (9th ed.). McGraw-Hill.

Borgianni, Y., & Maccioni, L. (2020). Review of the use of neurophysiological and biometric measures in experimental design research. *AI EDAM*, *34*(2), 248–285. https://doi.org/10.1017/S0890060420000062

Cairns, P. E., & Cox, A. L. (2008). *Research methods for human-computer interaction.* Cambridge University Press.

Cash, P., Isaksson, O., Maier, A., & Summers, J. (2022). Sampling in design research: Eight key considerations. *Design Studies*, *78*, 101077. https://doi.org/10.1016/j.destud.2021.101077

Cash, P. J. (2018). Developing theory-driven design research. *Design Studies*, *56*, 84–119. https://doi.org/10.1016/j.destud.2018.03.002

Cash, P., & Maier, A. (2021). Understanding representation: Contrasting gesture and sketching in design through dual-process theory. *Design Studies*, *73*, 100992. https://doi.org/10.1016/j.destud.2021.100992

Cash, P., Stanković, T., & Štorga, M. (Eds.). (2016). *Experimental Design Research*. Springer International Publishing. https://doi.org/10.1007/978-3-319-33781-4

Cisler, D., Greenwood, P. M., Roberts, D. M., McKendrick, R., & Baldwin, C. L. (2019). Comparing the Relative Strengths of EEG and Low-Cost Physiological Devices in Modeling Attention Allocation in Semiautonomous Vehicles. *Frontiers in Human Neuroscience*, *13*. https://doi.org/10.3389/fnhum.2019.00109

Consolvo, S., Harrison, B., Smith, I., Chen, M. Y., Everitt, K., Froehlich, J., & Landay, J. A. (2007). Conducting In Situ Evaluations for and With Ubiquitous Computing Technologies. *International Journal of Human–Computer Interaction*, *22*(1–2), 103–118. https://doi.org/10.1080/10447310709336957

Cui, X., Bray, S., Bryant, D. M., Glover, G. H., & Reiss, A. L. (2011). A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. *NeuroImage*, *54*(4), 2808–2821. https://doi.org/10.1016/j.neuroimage.2010.10.069

Cui, X., Bray, S., & Reiss, A. L. (2010). Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *NeuroImage*, *49*(4), 3039–3046. https://doi.org/10.1016/j.neuroimage.2009.11.050

Cui, X., Bryant, D. M., & Reiss, A. L. (2012). NIRS-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation. *NeuroImage*, *59*(3), 2430–2437. https://doi.org/10.1016/j.neuroimage.2011.09.003

Cutini, S., Scatturin, P., & Zorzi, M. (2011). A new method based on ICBM152 head surface for probe placement in multichannel fNIRS. *NeuroImage*, *54*(2), 919–927. https://doi.org/10.1016/j.neuroimage.2010.09.030

Dale, A. M., & Buckner, R. L. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping*, *5*(5), 329–340. https://doi.org/10.1002/(SICI)1097-0193(1997)5:5<329::AID-HBM1>3.0.CO;2-5

Damoiseaux, J. S., Rombouts, S. A. R. B., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., & Beckmann, C. F. (2006). Consistent resting-state networks across healthy subjects. *Proceedings of the National Academy of Sciences*, *103*(37), 13848–13853. https://doi.org/10.1073/pnas.0601417103

Davies, D. J., Su, Z., Clancy, M. T., Lucas, S. J. E., Dehghani, H., Logan, A., & Belli, A. (2015). Near-Infrared Spectroscopy in the Monitoring of Adult Traumatic Brain Injury: A Review. *Journal of Neurotrauma*, *32*(13), 933–941. https://doi.org/10.1089/neu.2014.3748

Delpy, D. T., Cope, M., Zee, P. van der, Arridge, S., Wray, S., & Wyatt, J. (1988). Estimation of optical pathlength through tissue from direct time of flight measurement. *Physics in Medicine & Biology*, *33*(12), 1433. https://doi.org/10.1088/0031-9155/33/12/008

Diamond, S. G., Huppert, T. J., Kolehmainen, V., Franceschini, M. A., Kaipio, J. P., Arridge, S. R., & Boas, D. A. (2005). Physiological System Identification with the Kalman Filter in Diffuse Optical Tomography. In J. S. Duncan & G. Gerig (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005* (pp. 649–656). Springer. https://doi.org/10.1007/11566489_80

Dong, A. (2005). The latent semantic approach to studying design team communication. *Design Studies*, *26*(5), 445–461. https://doi.org/10.1016/j.destud.2004.10.003

Dong, A., Kleinsmann, M. S., & Deken, F. (2013). Investigating design cognition in the construction and enactment of team mental models. *Design Studies*, *34*(1), 1–33. https://doi.org/10.1016/j.destud.2012.05.003

Druckman, J. N., & Kam, C. D. (2011). Students as experimental participants. *Cambridge Handbook of Experimental Political Science*, *1*, 41–57.

Dybvik, H., Abelson, F. G., Aalto, P., Goucher-Lambert, K., & Steinert, M. (2023). Inspirational Stimuli Attain Visual Allocation: Examining Design Ideation with Eye-Tracking. In J. S. Gero (Ed.), *Design Computing and Cognition'22* (pp. 463–480). Springer International Publishing. https://doi.org/10.1007/978-3-031-20418-0_28

Dybvik, H., Abelson, F. G., Aalto, P., Goucher-Lambert, K., & Steinert, M. (2022). Inspirational Stimuli Improve Idea Fluency during Ideation: A Replication and Extension Study with Eye-Tracking. *Proceedings of the Design Society*, *2*, 861–870. https://doi.org/10.1017/pds.2022.88

Dybvik, H., Erichsen, C. K., & Steinert, M. (n.d.). *Tetris' effect on cognitive load, performance, and systemic neurophysiology*.

Dybvik, H., Erichsen, C. K., & Steinert, M. (2021a). Demonstrating the Feasibility of Multimodal Neuroimaging Data Capture with a Wearable Electoencephalography + Functional Near-Infrared Spectroscopy (EEG+FNIRS) in Situ. *Proceedings of the Design Society*, *1*, 901–910. https://doi.org/10.1017/pds.2021.90

Dybvik, H., Erichsen, C. K., & Steinert, M. (2021b). Description of a Wearable Electroencephalography + Functional Near-Infrared Spectroscopy (EEG+FNIRS) for In-Situ Experiments on Design Cognition. *Proceedings of the International Conference on Engineering Design (ICED21)*, *1*, 943–952. https://doi.org/10.1017/pds.2021.94

Dybvik, H., Løland, M., Gerstenberg, A., Slåttsveen, K. B., & Steinert, M. (2021c). A low-cost predictive display for teleoperation: Investigating effects on human performance and workload. *International Journal of Human-Computer Studies*, *145*, 102536. https://doi.org/10.1016/j.ijhcs.2020.102536

Dybvik, H., & Steinert, M. (2021). Real-World fNIRS Brain Activity Measurements during Ashtanga Vinyasa Yoga. *Brain Sciences*, *11*(6), 742. https://doi.org/10.3390/brainsci11060742

Dybvik, H., & Steinert, M. (Under review). *Operationalized hypotheses build bridges between qualitative and quantitative design research*.

Dybvik, H., Veitch, E., & Steinert, M. (2020). EXPLORING CHALLENGES WITH DESIGNING AND DEVELOPING SHORE CONTROL CENTERS (SCC) FOR AUTONOMOUS SHIPS. *Proceedings of the Design Society: DESIGN Conference*, *1*, 847–856. https://doi.org/10/ggz7zb

Dybvik, H., Wulvik, A., & Steinert, M. (2018). STEERING A SHIP - INVESTIGATING AFFECTIVE STATE AND WORKLOAD IN SHIP SIMULATIONS. *Proceedings of the Design Society: DESIGN Conference*, 2003–2014. https://doi.org/10.21278/idc.2018.0459

Eickhoff, S. B., & Müller, V. I. (2015). Functional Connectivity. In A. W. Toga (Ed.), *Brain Mapping* (pp. 187–201). Academic Press. https://doi.org/10.1016/B978-0-12-397025-1.00212-8

Emberson, L. L., Zinszer, B. D., Raizada, R. D. S., & Aslin, R. N. (2017). Decoding the infant mind: Multivariate pattern analysis (MVPA) using fNIRS. *PLOS ONE*, *12*(4), e0172500. https://doi.org/10.1371/journal.pone.0172500

Erichsen, C. K., Dybvik, H., & Steinert, M. (2020). Integration of low-cost, dry-comb EEG-electrodes with a standard electrode cap for multimodal signal acquisition during human experiments. *DS 101: Proceedings of NordDesign 2020, Lyngby, Denmark, 12th - 14th August 2020*, 1–12. https://doi.org/10.35199/NORDDESIGN2020.19

Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K.-R., & Blankertz, B. (2012). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *NeuroImage*, *59*(1), 519–529. https://doi.org/10.1016/j.neuroimage.2011.07.084

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

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th edition). SAGE Publications.

Fishburn, F. A., Ludlum, R. S., Vaidya, C. J., & Medvedev, A. V. (2019). Temporal Derivative Distribution Repair (TDDR): A motion correction method for fNIRS. *NeuroImage*, *184*, 171–179. https://doi.org/10.1016/j.neuroimage.2018.09.025

Fishburn, F. A., Norr, M. E., Medvedev, A. V., & Vaidya, C. J. (2014). Sensitivity of fNIRS to cognitive state and load. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00076

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, *54*(1), 313–327. https://doi.org/10.1016/j.neuroimage.2010.07.033

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Franceschini, M. A., Joseph, D. K., Huppert, T. J., Diamond, S. G., & Boas, D. A. (2006). Diffuse optical imaging of the whole head. *Journal of Biomedical Optics*, *11*(5), 054007. https://doi.org/10.1117/1.2363365

Funahashi, S. (2017). Working Memory in the Prefrontal Cortex. *Brain Sciences*, *7*(5), 49. https://doi.org/10.3390/brainsci7050049

Funane, T., Atsumori, H., Katura, T., Obata, A. N., Sato, H., Tanikawa, Y., Okada, E., & Kiguchi, M. (2014). Quantitative evaluation of deep and shallow tissue layers' contribution to fNIRS signal using multi-distance optodes and independent component analysis. *NeuroImage*, *85*, 150–165. https://doi.org/10.1016/j.neuroimage.2013.02.026

Gagnon, L., Cooper, R. J., Yücel, M. A., Perdue, K. L., Greve, D. N., & Boas, D. A. (2012a). Short separation channel location impacts the performance of short channel regression in NIRS. *NeuroImage*, *59*(3), 2518–2528. https://doi.org/10.1016/j.neuroimage.2011.08.095

Gagnon, L., Yücel, M. A., Dehaes, M., Cooper, R. J., Perdue, K. L., Selb, J., Huppert, T. J., Hoge, R. D., & Boas, D. A. (2012b). Quantification of the cortical contribution to the NIRS signal over the motor cortex using concurrent NIRS-fMRI measurements. *NeuroImage*, *59*(4), 3933–3940. https://doi.org/10.1016/j.neuroimage.2011.10.054

Gemignani, J. (2019). *Expanding the analysis of functional Near-Infrared Spectroscopy (fNIRS) data with multivariate techniques* [Doctoral dissertation, Technischen Universität Berlin]. https://depositonce.tu-berlin.de/handle/11303/10247

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

Gonçalves, M., & Cash, P. (2021). The life cycle of creative ideas: Towards a dual-process theory of ideation. *Design Studies*, *72*, 100988. https://doi.org/10.1016/j.destud.2020.100988

Goucher-Lambert, K., & McComb, C. (2019). Using Hidden Markov Models to Uncover Underlying States in Neuroimaging Data for a Design Ideation Task. *Proceedings of the Design Society: International Conference on Engineering Design*, *1*(1), 1873–1882. https://doi.org/10.1017/dsi.2019.193

Goucher-Lambert, K., Moss, J., & Cagan, J. (2019). A neuroimaging investigation of design ideation with and without inspirational stimuli—Understanding the meaning of near and far stimuli. *Design Studies*, *60*, 1–38. https://doi.org/10.1016/j.destud.2018.07.001

Gratton, G., & Corballis, P. M. (1995). Removing the heart from the brain: Compensation for the pulse artifact in the photon migration signal. *Psychophysiology*, *32*(3), 292–299. https://doi.org/10.1111/j.1469-8986.1995.tb02958.x

Gusnard, D. A., & Raichle, M. E. (2001). Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience*, *2*(10), 685–694. https://doi.org/10.1038/35094500

Hamann, A., & Carstengerdes, N. (2022). Investigating mental workload-induced changes in cortical oxygenation and frontal theta activity during simulated flights. *Scientific Reports*, *12*(1), 6449. https://doi.org/10.1038/s41598-022-10044-y

Hassib, M., Schneegass, S., Eiglsperger, P., Henze, N., Schmidt, A., & Alt, F. (2017). EngageMeter: A System for Implicit Audience Engagement Sensing Using Electroencephalography. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5114–5119. https://doi.org/10.1145/3025453.3025669

Hatlem, L. A., Chen, J., Dybvik, H., & Steinert, M. (2020). A Modular Research Platform – Proof-of-Concept of a Flexible Experiment Setup Developed for Rapid Testing of Simulators, UIs and Human Physiology Sensors. *Procedia CIRP*, *91*, 407–414. https://doi.org/10.1016/j.procir.2020.02.193

Hay, L., Cash, P., & McKilligan, S. (2020). The future of design cognition analysis. *Design Science*, *6*. https://doi.org/10.1017/dsj.2020.20

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—Quantified in the prefrontal cortex using fNIRS. *Frontiers in Human Neuroscience*, *7*. https://www.frontiersin.org/article/10.3389/fnhum.2013.00935

Hernandez, S. M., & Pollonini, L. (2020). NIRSplot: A Tool for Quality Assessment of fNIRS Scans. *Biophotonics Congress: Biomedical Optics 2020 (Translational, Microscopy, OCT, OTS, BRAIN)*, BM2C.5. https://doi.org/10.1364/BRAIN.2020.BM2C.5

Herold, F., Wiegel, P., Scholkmann, F., & Müller, N. G. (2018). Applications of Functional Near-Infrared Spectroscopy (fNIRS) Neuroimaging in Exercise–Cognition Science: A Systematic, Methodology-Focused Review. *Journal of Clinical Medicine*, *7*(12), 466. https://doi.org/10.3390/jcm7120466

Hocke, L. M., Oni, I. K., Duszynski, C. C., Corrigan, A. V., Frederick, B. D., & Dunn, J. F. (2018). Automated Processing of fNIRS Data—A Visual Guide to the Pitfalls and Consequences. *Algorithms*, *11*(5), 67. https://doi.org/10.3390/a11050067

Holper, L., Scholkmann, F., & Seifritz, E. (2015). Time–frequency dynamics of the sum of intra- and extracerebral hemodynamic functional connectivity during resting-state and respiratory challenges assessed by multimodal functional near-infrared spectroscopy. *NeuroImage*, *120*, 481–492. https://doi.org/10.1016/j.neuroimage.2015.07.021

Hu, L., & Shepley, M. M. (2022). Design Meets Neuroscience: A Preliminary Review of Design Research Using Neuroscience Tools. *Journal of Interior Design*, *47*(1), 31–50. https://doi.org/10.1111/joid.12213

Hu, M., & Shealy, T. (2018). *Systems versus Linear Thinking: Measuring Cognitive Networks for Engineering Sustainability*. 726–736. https://doi.org/10.1061/9780784481301.072

Hu, M., & Shealy, T. (2019). Application of Functional Near-Infrared Spectroscopy to Measure Engineering Decision-Making and Design Cognition: Literature Review and Synthesis of Methods. *Journal of Computing in Civil Engineering*, *33*(6), 04019034. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000848

Hu, M., Shealy, T., & Gero, J. S. (2018, June 23). Neuro-cognitive Differences Among Engineering Students when Using Unstructured, Partially Structured, and Structured Design Concept Generation Techniques. *2018 ASEE Annual Conference & Exposition*. 2018 ASEE Annual Conference & Exposition. https://peer.asee.org/neuro-cognitive-differences-among-engineering-students-when-using-unstructured-partially-structured-and-structured-design-concept-generation-techniques

Hu, M., Shealy, T., & Milovanovic, J. (2021). Cognitive differences among first-year and senior engineering students when generating design solutions with and without additional dimensions of sustainability. *Design Science*, *7*, e1. https://doi.org/10.1017/dsj.2021.3

Hu, X.-S., Hong, K.-S., Ge, S. S., & Jeong, M.-Y. (2010). Kalman estimator- and general linear model-based on-line brain activation mapping by near-infrared spectroscopy. *BioMedical Engineering OnLine*, *9*(1), 82. https://doi.org/10.1186/1475-925X-9-82

Huang, R., Hong, K.-S., Yang, D., & Huang, G. (2022). Motion artifacts removal and evaluation techniques for functional near-infrared spectroscopy signals: A review. *Frontiers in Neuroscience*, *16*. https://www.frontiersin.org/articles/10.3389/fnins.2022.878750

Huppert, T. J. (2016). Commentary on the statistical properties of noise and its implication on general linear models in functional near-infrared spectroscopy. *Neurophotonics*, *3*(1), 010401. https://doi.org/10.1117/1.NPh.3.1.010401

Jacko, J. A. (2012). *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications* (3rd ed.). CRC Press.

Jahani, S., Setarehdan, S. K., Boas, D. A., & Yücel, M. A. (2018). Motion artifact detection and correction in functional near-infrared spectroscopy: A new hybrid method based on spline interpolation method and Savitzky–Golay filtering. *Neurophotonics*, *5*(1), 015003. https://doi.org/10.1117/1.NPh.5.1.015003

Jennings, J. R., Kamarck, T., Stewart, C., Eddy, M., & Johnson, P. (1992). Alternate Cardiovascular Baseline Assessment Techniques: Vanilla or Resting Baseline. *Psychophysiology*, *29*(6), 742–750. https://doi.org/10.1111/j.1469-8986.1992.tb02052.x

Jiang, J., Chen, C., Dai, B., Shi, G., Ding, G., Liu, L., & Lu, C. (2015). Leader emergence through interpersonal neural synchronization. *Proceedings of the National Academy of Sciences*, *112*(14), 4274–4279. https://doi.org/10.1073/pnas.1422930112

Kabdebon, C., Leroy, F., Simmonet, H., Perrot, M., Dubois, J., & Dehaene-Lambertz, G. (2014). Anatomical correlations of the international 10–20 sensor placement system in infants. *NeuroImage*, *99*, 342–356. https://doi.org/10.1016/j.neuroimage.2014.05.046

Kato, T., Okada, H., & Izu, Y. (2018). MEASUREMENT OF BRAIN ACTIVITIES OF IDEA GENERATION (SKETCH). *DS 92: Proceedings of the DESIGN 2018 15th International Design Conference*, 2027–2034. https://doi.org/10.21278/idc.2018.0133

Kelle, U. (1997). Theory Building in Qualitative Research and Computer Programs for the Management of Textual Data. *Sociological Research Online*, *2*(2), 10–22. https://doi.org/10.5153/sro.86

Kirk, R. (2013). *Experimental Design: Procedures for the Behavioral Sciences*. SAGE Publications, Inc. https://doi.org/10.4135/9781483384733

Lee, M.-H., Fazli, S., Mehnert, J., & Lee, S.-W. (2015). Subject-dependent Classification for Robust Idle State Detection Using Multi-modal Neuroimaging and Data-fusion Techniques in BCI. *Pattern Recogn.*, *48*(8), 2725–2737. https://doi.org/10.1016/j.patcog.2015.03.010

Lee, S.-H., & Park, Y. (2018). Computational EEG Analysis for the Diagnosis of Psychiatric Illnesses. In C.-H. Im (Ed.), *Computational EEG Analysis* (pp. 149–175). Springer Singapore. https://doi.org/10.1007/978-981-13-0908-3_7

Leifer, L. J., & Steinert, M. (2011). Dancing with ambiguity: Causality behavior, design thinking, and triple-loop-learning. *Information Knowledge Systems Management*, *10*(1–4), 151–173. https://doi.org/10.3233/IKS-2012-0191

Leithner, C., & Royl, G. (2014). The oxygen paradox of neurovascular coupling. *Journal of Cerebral Blood Flow & Metabolism*, *34*(1), 19–29. https://doi.org/10.1038/jcbfm.2013.181

Li, R., Mayseless, N., Balters, S., & Reiss, A. L. (2021). Dynamic inter-brain synchrony in real-life inter-personal cooperation: A functional near-infrared spectroscopy hyperscanning study. *NeuroImage*, *238*, 118263. https://doi.org/10.1016/j.neuroimage.2021.118263

Li, R., Yang, D., Fang, F., Hong, K.-S., Reiss, A. L., & Zhang, Y. (2022). Concurrent fNIRS and EEG for Brain Function Investigation: A Systematic, Methodology-Focused Review. *Sensors*, *22*(15), 5865. https://doi.org/10.3390/s22155865

Liu, N., Mok, C., Witt, E. E., Pradhan, A. H., Chen, J. E., & Reiss, A. L. (2016). NIRS-Based Hyperscanning Reveals Inter-brain Neural Synchronization during Cooperative Jenga Game with Face-to-Face Communication. *Frontiers in Human Neuroscience*, *10*. https://www.frontiersin.org/articles/10.3389/fnhum.2016.00082

Lu, C.-M., Zhang, Y.-J., Biswal, B. B., Zang, Y.-F., Peng, D.-L., & Zhu, C.-Z. (2010). Use of fNIRS to assess resting state functional connectivity. *Journal of Neuroscience Methods*, *186*(2), 242–249. https://doi.org/10.1016/j.jneumeth.2009.11.010

Lukanov, K., Maior, H. A., & Wilson, M. L. (2016). Using fNIRS in Usability Testing: Understanding the Effect of Web Form Layout on Mental Workload. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4011–4016. https://doi.org/10.1145/2858036.2858236

Mayseless, N., Hawthorne, G., & Reiss, A. L. (2019). Real-life creative problem solving in teams: FNIRS based hyperscanning study. *NeuroImage*, *203*, 116161. https://doi.org/10.1016/j.neuroimage.2019.116161

McComb, C., Cagan, J., & Kotovsky, K. (2017a). Mining Process Heuristics From Designer Action Data via Hidden Markov Models. *Journal of Mechanical Design*, *139*(11). https://doi.org/10.1115/1.4037308

McComb, C., Cagan, J., & Kotovsky, K. (2017b). Utilizing Markov Chains to Understand Operation Sequencing in Design Tasks. In John. S. Gero (Ed.), *Design Computing and Cognition '16* (pp. 401–418). Springer International Publishing. https://doi.org/10.1007/978-3-319-44989-0_22

McKim, R. H. (1959). Designing for the whole man. In *Creative engineering* (pp. 198–217). Stanford University.

Meidenbauer, K. L., Choe, K. W., Cardenas-Iniguez, C., Huppert, T. J., & Berman, M. G. (2021). Load-dependent relationships between frontal fNIRS activity and performance: A data-driven PLS approach. *NeuroImage*, *230*, 117795. https://doi.org/10.1016/j.neuroimage.2021.117795

Mesquita, R. C., Franceschini, M. A., & Boas, D. A. (2010). Resting state functional connectivity of the whole head with near-infrared spectroscopy. *Biomedical Optics Express*, *1*(1), 324–336. https://doi.org/10.1364/BOE.1.000324

Metz, A. J., Klein, S. D., Scholkmann, F., & Wolf, U. (2017). Continuous coloured light altered human brain haemodynamics and oxygenation assessed by systemic physiology augmented functional near-infrared spectroscopy. *Scientific Reports*, *7*(1), 10027. https://doi.org/10.1038/s41598-017-09970-z

Miller, E. K., Freedman, D. J., & Wallis, J. D. (2002). The prefrontal cortex: Categories, concepts and cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *357*(1424), 1123–1136. https://doi.org/10.1098/rstb.2002.1099

Milovanovic, J., Hu, M., Shealy, T., & Gero, J. (2020, November 3). *Evolution of Brain Network Connectivity in the Prefrontal Cortex During Concept Generation Using Brainstorming for a Design Task*. ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. https://doi.org/10.1115/DETC2020-22563

Molavi, B., & Dumont, G. A. (2012). Wavelet-based motion artifact removal for functional near-infrared spectroscopy. *Physiological Measurement*, *33*(2), 259. https://doi.org/10.1088/0967-3334/33/2/259

Nguyen, T. A., & Zeng, Y. (2014). A physiological study of relationship between designer’s mental effort and mental stress during conceptual design. *Computer-Aided Design*, *54*, 3–18. https://doi.org/10.1016/j.cad.2013.10.002

Nguyen, T., Hoehl, S., & Vrtička, P. (2021). A Guide to Parent-Child fNIRS Hyperscanning Data Processing and Analysis. *Sensors*, *21*(12), 4075. https://doi.org/10.3390/s21124075

NIRx Medical Technologies. (2019a). *Getting Started Guide: FNIRS Experimental Design & Stimulus Presentation*. NIRx Medical Technologies. https://support.nirx.de/documentation

NIRx Medical Technologies. (2019b). *Getting Started Guide: Montage Design for fNIRS Experiments*. NIRx Medical Technologies. https://support.nirx.de/software

Norwegian Ministry of Education and Research. (2017, August 22). *National goals and guidelines for open access to research articles* [Retningslinjer]. Government.No; regjeringen.no. https://www.regjeringen.no/en/dokumenter/national-goals-and-guidelines-for-open-access-to-research-articles/id2567591/

Norwegian Ministry of Education and Research. (2018, February 13). *National strategy on access to and sharing of research data* [Plan]. Government.No; regjeringen.no. https://www.regjeringen.no/en/dokumenter/national-strategy-on-access-to-and-sharing-of-research-data/id2582412/

NTNU. (2020, October 2). *Policy for Open Science at NTNU*. https://www.ntnu.edu/policy-for-open-science

Obrig, H., Neufang, M., Wenzel, R., Kohl, M., Steinbrink, J., Einhäupl, K., & Villringer, A. (2000). Spontaneous Low Frequency Oscillations of Cerebral Hemodynamics and Metabolism in Human Adults. *NeuroImage*, *12*(6), 623–639. https://doi.org/10.1006/nimg.2000.0657

Ohashi, T., Auernhammer, J., Liu, W., Pan, W., & Leifer, L. (2022). NeuroDesignScience: Systematic Literature Review of Current Research on Design Using Neuroscience Techniques. In J. S. Gero (Ed.), *Design Computing and Cognition’20* (pp. 575–592). Springer International Publishing. https://doi.org/10.1007/978-3-030-90625-2_34

Okamoto, M., Dan, H., Sakamoto, K., Takeo, K., Shimizu, K., Kohno, S., Oda, I., Isobe, S., Suzuki, T., Kohyama, K., & Dan, I. (2004a). Three-dimensional probabilistic anatomical cranio-cerebral correlation via the international 10–20 system oriented for transcranial functional brain mapping. *NeuroImage*, *21*(1), 99–111. https://doi.org/10.1016/j.neuroimage.2003.08.026

Okamoto, M., Dan, H., Shimizu, K., Takeo, K., Amita, T., Oda, I., Konishi, I., Sakamoto, K., Isobe, S., Suzuki, T., Kohyama, K., & Dan, I. (2004b). Multimodal assessment of cortical activation during apple peeling by NIRS and fMRI. *NeuroImage*, *21*(4), 1275–1288. https://doi.org/10.1016/j.neuroimage.2003.12.003

Olds, C., Pollonini, L., Abaya, H., Larky, J., Loy, M., Bortfeld, H., Beauchamp, M. S., & Oghalai, J. S. (2016). Cortical activation patterns correlate with speech understanding after cochlear implantation. *Ear and Hearing*, *37*(3), e160–e172. https://doi.org/10.1097/AUD.0000000000000258

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, *112*(4), 713–719. https://doi.org/10.1016/S1388-2457(00)00527-7

Paranawithana, I., Mao, D., Wong, Y. T., & McKay, C. M. (2022). Reducing false discoveries in resting-state functional connectivity using short channel correction: An fNIRS study. *Neurophotonics*, *9*(1), 015001. https://doi.org/10.1117/1.NPh.9.1.015001

Pfeifer, M. D., Scholkmann, F., & Labruyère, R. (2018). Signal Processing in Functional Near-Infrared Spectroscopy (fNIRS): Methodological Differences Lead to Different Statistical Results. *Frontiers in Human Neuroscience*, *11*. https://www.frontiersin.org/articles/10.3389/fnhum.2017.00641

Phillips, A. A., Chan, F. H., Zheng, M. M. Z., Krassioukov, A. V., & Ainslie, P. N. (2016). Neurovascular coupling in humans: Physiology, methodological advances and clinical implications. *Journal of Cerebral Blood Flow & Metabolism*, *36*(4), 647–664. https://doi.org/10.1177/0271678X15617954

Pinti, P., Aichelburg, C., Lind, F., Power, S., Swingler, E., Merla, A., Hamilton, A., Gilbert, S., Burgess, P., & Tachtsidis, I. (2015). Using Fiberless, Wearable fNIRS to Monitor Brain Activity in Real-world Cognitive Tasks. *JoVE (Journal of Visualized Experiments)*, *106*, e53336. https://doi.org/10.3791/53336

Pinti, P., Merla, A., Aichelburg, C., Lind, F., Power, S., Swingler, E., Hamilton, A., Gilbert, S., Burgess, P. W., & Tachtsidis, I. (2017). A novel GLM-based method for the Automatic IDentification of functional Events (AIDE) in fNIRS data recorded in naturalistic environments. *NeuroImage*, *155*, 291–304. https://doi.org/10.1016/j.neuroimage.2017.05.001

Pinti, P., Scholkmann, F., Hamilton, A., Burgess, P., & Tachtsidis, I. (2019). Current Status and Issues Regarding Pre-processing of fNIRS Neuroimaging Data: An Investigation of Diverse Signal Filtering Methods Within a General Linear Model Framework. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00505

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2020). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1464*(1), 5–29. https://doi.org/10.1111/nyas.13948

Pollonini, L., Bortfeld, H., & Oghalai, J. S. (2016). PHOEBE: A method for real time mapping of optodes-scalp coupling in functional near-infrared spectroscopy. *Biomedical Optics Express*, *7*(12), 5104–5119. https://doi.org/10.1364/BOE.7.005104

Popper, K. (2002). *The Logic of Scientific Discovery* (2nd ed.). Routledge.

Published fNIRS Research with NIRx Systems. (n.d.). NIRx Medical Technologies. Retrieved January 26, 2023, from https://nirx.net/publications

Purves, D., Mooney, R. D., & Platt, M. L. (2012). *Neuroscience* (5th ed.). Sinauer Associates.

Quaresima, V., & Ferrari, M. (2019). Functional Near-Infrared Spectroscopy (fNIRS) for Assessing Cerebral Cortex Function During Human Behavior in Natural/Social Situations: A Concise Review. *Organizational Research Methods*, *22*(1), 46–68. https://doi.org/10.1177/1094428116658959

Rahman, T. T., Polskaia, N., St-Amant, G., Salzman, T., Vallejo, D. T., Lajoie, Y., & Fraser, S. A. (2021). An fNIRS Investigation of Discrete and Continuous Cognitive Demands During Dual-Task Walking in Young Adults. *Frontiers in Human Neuroscience*, *15*. https://www.frontiersin.org/article/10.3389/fnhum.2021.711054

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(2), 676–682. https://doi.org/10.1073/pnas.98.2.676

Rørvik, S. B., Auflem, M., Dybvik, H., & Steinert, M. (2021). Perception by Palpation: Development and Testing of a Haptic Ferrogranular Jamming Surface. *Frontiers in Robotics and AI*, *8*, 311. https://doi.org/10.3389/frobt.2021.745234

Rupawala, M., Dehghani, H., Lucas, S. J. E., Tino, P., & Cruse, D. (2018). Shining a Light on Awareness: A Review of Functional Near-Infrared Spectroscopy for Prolonged Disorders of Consciousness. *Frontiers in Neurology*, *9*. https://www.frontiersin.org/articles/10.3389/fneur.2018.00350

Salkind, N. (2010). *Encyclopedia of Research Design*. SAGE Publications, Inc. https://doi.org/10.4135/9781412961288

Santosa, H., Aarabi, A., Perlman, S. B., & Huppert, T. (2017). Characterization and correction of the false-discovery rates in resting state connectivity using functional near-infrared spectroscopy. *Journal of Biomedical Optics*, *22*(5), 055002. https://doi.org/10.1117/1.JBO.22.5.055002

Santosa, H., Jiyoun Hong, M., Kim, S.-P., & Hong, K.-S. (2013). Noise reduction in functional near-infrared spectroscopy signals by independent component analysis. *Review of Scientific Instruments*, *84*(7), 073106. https://doi.org/10.1063/1.4812785

Santosa, H., Zhai, X., Fishburn, F., & Huppert, T. (2018). The NIRS Brain AnalyzIR Toolbox. *Algorithms*, *11*(5), 73. https://doi.org/10.3390/a11050073

Santosa, H., Zhai, X., Fishburn, F., Sparto, P. J., & Huppert, T. J. (2020). Quantitative comparison of correction techniques for removing systemic physiological signal in functional near-infrared spectroscopy studies. *Neurophotonics*, *7*(3), 035009. https://doi.org/10.1117/1.NPh.7.3.035009

Sappia, M. S., Hakimi, N., Colier, W. N. J. M., & Horschig, J. M. (2020). Signal quality index: An algorithm for quantitative assessment of functional near infrared spectroscopy signal quality. *Biomedical Optics Express*, *11*(11), 6732. https://doi.org/10.1364/BOE.409317

Sato, T., Nambu, I., Takeda, K., Aihara, T., Yamashita, O., Isogaya, Y., Inoue, Y., Otaka, Y., Wada, Y., Kawato, M., Sato, M., & Osu, R. (2016). Reduction of global interference of scalp-hemodynamics in functional near-infrared spectroscopy using short distance probes. *NeuroImage*, *141*, 120–132. https://doi.org/10.1016/j.neuroimage.2016.06.054

Scholkmann, F., Kleiser, S., Metz, A. J., Zimmermann, R., Mata Pavia, J., Wolf, U., & Wolf, M. (2014). A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *NeuroImage*, *85*, 6–27. https://doi.org/10.1016/j.neuroimage.2013.05.004

Scholkmann, F., Spichtig, S., Muehlemann, T., & Wolf, M. (2010). How to detect and reduce movement artifacts in near-infrared imaging using moving standard deviation and spline interpolation. *Physiological Measurement*, *31*(5), 649–662. https://doi.org/10.1088/0967-3334/31/5/004

Scholkmann, F., & Vollenweider, F. X. (2022). Psychedelics and fNIRS neuroimaging: Exploring new opportunities. *Neurophotonics*, *10*(1), 013506. https://doi.org/10.1117/1.NPh.10.1.013506

Shealy, T. (1), Gero, J. (2), Milovanovic, J. (3), & Hu, M. (1). (2020a). SUSTAINING CREATIVITY WITH NEURO-COGNITIVE FEEDBACK: A PRELIMINARY STUDY. *Proceedings of the Sixth International Conference on Design Creativity (ICDC 2020)*, 084–091. https://doi.org/10.35199/ICDC.2020.11

Shealy, T., Gero, J., Hu, M., & Milovanovic, J. (2020b). Concept generation techniques change patterns of brain activation during engineering design. *Design Science*, *6*. https://doi.org/10.1017/dsj.2020.30

Shealy, T., Grohs, J. R., Hu, M., Maczka, D. K., & Panneton, R. (2017, June 24). Investigating Design Cognition during Brainstorming Tasks with Freshmen and Senior Engineering Students using Functional Near Infrared Spectroscopy. *2017 ASEE Annual Conference & Exposition*. 2017 ASEE Annual Conference & Exposition. https://peer.asee.org/investigating-design-cognition-during-brainstorming-tasks-with-freshmen-and-senior-engineering-students-using-functional-near-infrared-spectroscopy

Shealy, T., Hu, M., & Gero, J. (2018, November 2). *Patterns of Cortical Activation When Using Concept Generation Techniques of Brainstorming, Morphological Analysis, and TRIZ*. ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. https://doi.org/10.1115/DETC2018-86272

Solovey, E. T., Girouard, A., Chauncey, K., Hirshfield, L. M., Sassaroli, A., Zheng, F., Fantini, S., & Jacob, R. J. K. (2009). Using fNIRS brain sensing in realistic HCI settings: Experiments and guidelines. *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*, 157–166. https://doi.org/10.1145/1622176.1622207

Steinert, M., & Jablokow, K. (2013). *Triangulating front end engineering design activities with physiology data and psychological preferences*. 109–118.

Stempfle, J., & Badke-Schaub, P. (2002). Thinking in design teams—An analysis of team communication. *Design Studies*, *23*(5), 473–496. https://doi.org/10.1016/S0142-694X(02)00004-2

Surma-aho, A., & Hölttä-Otto, K. (2022). Conceptualization and operationalization of empathy in design research. *Design Studies*, *78*, 101075. https://doi.org/10.1016/j.destud.2021.101075

Tachibana, A., Noah, J. A., Bronner, S., Ono, Y., & Onozuka, M. (2011). Parietal and temporal activity during a multimodal dance video game: An fNIRS study. *Neuroscience Letters*, *503*(2), 125–130. https://doi.org/10.1016/j.neulet.2011.08.023

Tachtsidis, I., & Scholkmann, F. (2016). False positives and false negatives in functional near-infrared spectroscopy: Issues, challenges, and the way forward. *Neurophotonics*, *3*(3), 031405. https://doi.org/10.1117/1.NPh.3.3.031405

Tak, S., & Ye, J. C. (2014). Statistical analysis of fNIRS data: A comprehensive review. *NeuroImage*, *85*, 72–91. https://doi.org/10.1016/j.neuroimage.2013.06.016

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 4–4. https://doi.org/10.1167/7.14.4

Tsuzuki, D., & Dan, I. (2014). Spatial registration for functional near-infrared spectroscopy: From channel position on the scalp to cortical location in individual and group analyses. *NeuroImage*, *85*, 92–103. https://doi.org/10.1016/j.neuroimage.2013.07.025

Vanderhasselt, M.-A., De Raedt, R., & Baeken, C. (2009). Dorsolateral prefrontal cortex and Stroop performance: Tackling the lateralization. *Psychonomic Bulletin & Review*, *16*(3), 609–612. https://doi.org/10.3758/PBR.16.3.609

Veitch, E., Dybvik, H., Steinert, M., & Alsos, O. A. (2022). Collaborative Work with Highly Automated Marine Navigation Systems. *Computer Supported Cooperative Work (CSCW)*. https://doi.org/10.1007/s10606-022-09450-7

von Lühmann, A., Zheng, Y., Ortega-Martinez, A., Kiran, S., Somers, D. C., Cronin-Golomb, A., Awad, L. N., Ellis, T. D., Boas, D. A., & Yücel, M. A. (2021). Toward Neuroscience of the Everyday World (NEW) using functional near-infrared spectroscopy. *Current Opinion in Biomedical Engineering*, *18*, 100272. https://doi.org/10.1016/j.cobme.2021.100272

Wijeakumar, S., Huppert, T. J., Magnotta, V. A., Buss, A. T., & Spencer, J. P. (2017). Validating an image-based fNIRS approach with fMRI and a working memory task. *NeuroImage*, *147*, 204–218. https://doi.org/10.1016/j.neuroimage.2016.12.007

Wulvik, A. S., Dybvik, H., & Steinert, M. (2019). Investigating the relationship between mental state (workload and affect) and physiology in a control room setting (ship bridge simulator). *Cognition, Technology & Work*. https://doi.org/10.1007/s10111-019-00553-8

Wyser, D., Mattille, M., Wolf, M., Lambercy, O., Scholkmann, F., & Gassert, R. (2020). Short-channel regression in functional near-infrared spectroscopy is more effective when considering heterogeneous scalp hemodynamics. *Neurophotonics*, *7*(3), 035011. https://doi.org/10.1117/1.NPh.7.3.035011

Xu, J., Slagle, J. M., Banerjee, A., Bracken, B., & Weinger, M. B. (2019). Use of a Portable Functional Near-Infrared Spectroscopy (fNIRS) System to Examine Team Experience During Crisis Event Management in Clinical Simulations. *Frontiers in Human Neuroscience*, *13*. https://doi.org/10.3389/fnhum.2019.00085

Yeung, M. K., & Chu, V. W. (2022). Viewing neurovascular coupling through the lens of combined EEG–fNIRS: A systematic review of current methods. *Psychophysiology*, *59*(6), e14054. https://doi.org/10.1111/psyp.14054

Yücel, M. A., Lühmann, A. v, Scholkmann, F., Gervain, J., Dan, I., Ayaz, H., Boas, D., Cooper, R. J., Culver, J., Elwell, C. E., Eggebrecht, A., Franceschini, M. A., Grova, C., Homae, F., Lesage, F., Obrig, H., Tachtsidis, I., Tak, S., Tong, Y., … Wolf, M. (2021). Best practices for fNIRS publications. *Neurophotonics*, *8*(1), 012101. https://doi.org/10.1117/1.NPh.8.1.012101

Yücel, M. A., Selb, J., Aasted, C. M., Lin, P.-Y., Borsook, D., Becerra, L., & Boas, D. A. (2016). Mayer waves reduce the accuracy of estimated hemodynamic response functions in functional near-infrared spectroscopy. *Biomedical Optics Express*, *7*(8), 3078–3088. https://doi.org/10.1364/BOE.7.003078

Zhang, Q., Strangman, G. E., & Ganis, G. (2009). Adaptive filtering to reduce global interference in non-invasive NIRS measures of brain activation: How well and when does it work? *NeuroImage*, *45*(3), 788–794. https://doi.org/10.1016/j.neuroimage.2008.12.048

Zhang, X., Noah, J. A., & Hirsch, J. (2016). Separation of the global and local components in functional near-infrared spectroscopy signals using principal component spatial filtering. *Neurophotonics*, *3*(1), 015004. https://doi.org/10.1117/1.NPh.3.1.015004

Zhang, Y., Brooks, D. H., Franceschini, M. A., & Boas, D. A. (2005). Eigenvector-based spatial filtering for reduction of physiological interference in diffuse optical imaging. *Journal of Biomedical Optics*, *10*(1), 011014. https://doi.org/10.1117/1.1852552

Zimeo Morais, G. A., Balardin, J. B., & Sato, J. R. (2018). fNIRS Optodes’ Location Decider (fOLD): A toolbox for probe arrangement guided by brain regions-of-interest. *Scientific Reports*, *8*(1), 3341. https://doi.org/10.1038/s41598-018-21716-z

Zohdi, H., Scholkmann, F., & Wolf, U. (2021). Individual Differences in Hemodynamic Responses Measured on the Head Due to a Long-Term Stimulation Involving Colored Light Exposure and a Cognitive Task: A SPA-fNIRS Study. *Brain Sciences*, *11*(1), 54. https://doi.org/10.3390/brainsci11010054

Appendix C1: Academic contribution 1

Wulvik, A. S., Dybvik, H., & Steinert, M. (2019). Investigating the relationship between mental state (workload and affect) and physiology in a control room setting (ship bridge simulator). Cognition, Technology & Work. https://doi.org/10.1007/s10111-019-00553-8

Cognition, Technology & Work (2020) 22:95–108
https://doi.org/10.1007/s10111-019-00553-8

ORIGINAL ARTICLE

Investigating the relationship between mental state (workload and affect) and physiology in a control room setting (ship bridge simulator)

Andreas Simskar Wulvik[1] · Henrikke Dybvik[1] · Martin Steinert[1]

Received: 28 June 2018 / Accepted: 25 March 2019 / Published online: 6 April 2019

Abstract
This paper discusses how to investigate the human element in a control room setting in terms of situational settings (monitoring and active control) and mental state (workload and affect). We show an explorative experiment in a ship bridge simulator context to investigate measurement practices and uncover correlations between mental state and changes in physiology. 31 participants from an engineering student population participated in the experiment. Data were collected from two scenarios through surveys (workload and affect) and physiology sensors (electrocardiography and electrodermal activity). We highlight the following findings from our experiment: One, there is a significant difference in variables measuring mental and physiological states between two regularly occurring scenarios in the context of large ship navigation. With changes in mental and physiological states, the capacity and reaction pattern of users change, so there are different demands of the user interface and system behavior. Two, elements of mental state are correlated with changes in physiological state. Most prominently, stress and workload covary with electrodermal activity and elements of heart rate variability. This finding can support designers in evaluating different solutions by enabling them to assess changes in the mental state of users working in control rooms through physiology sensor data.

Keywords Control room · Interaction design evaluation · Workload · Affect · Heart rate variability · Electrodermal activity

1 Introduction: including users' mental state

In modern industry, many aspects of operation have moved into the control room. When combined with increasing degrees of automation, users must keep track of more information and are responsible for an increasing number of systems. Woods et al. (2002) highlight this issue of data overload, describing the issue of users being presented with an abundance of data without being able to process and act on the data in an efficient manner. In addition, control room activities are shifting toward monitoring systems, ensuring that operation is within normal limits. In the event of anomalies, the user must be able to perform appropriate actions to ensure safe and efficient operation. Bainbridge (1983) highlights a challenge of automation—namely, that of users mainly monitoring systems instead of actively controlling them. This can be problematic when action must be taken, because users are less familiar with active control and might not have a feel for the process that they are controlling. When designing such systems, the user is usually modeled as having stable, rational behavior—i.e., responding to external events in a predictable manner (Balters and Steinert 2015). However, it has been shown that humans do not behave in stable, rational patterns (Kahneman and Tversky 1979, 1984). Behavior can be influenced in part by the mental state of users. We define mental state as the combination of affective state and experienced workload. A positive affect has been shown to influence decision making by increasing risk averseness (Isen 2001; Isen and Reeve 2005). Hart and Staveland (1988) claim that experienced workload influences problem-solving strategies. If a user

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10111-019-00553-8) contains supplementary material, which is available to authorized users.

✉ Andreas Simskar Wulvik
andreas.wulvik@ntnu.no

[1] Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

experiences a high workload, he might start shedding tasks or adopt a lower criterion for performance.

This knowledge—namely, the notion of behavior being influenced by mental state—has been adopted by the fields of engineering design and human–computer interaction. Affective computing (Picard 1997) aims to improve the interaction between users and computers by including knowledge of the user's mental state in the design and behavior of the system. Jiao et al. (2017) highlight the importance of integrating affective and cognitive needs when designing the user experience. A promising method for measuring mental state is through physiology. Many studies explore this correlation in strictly controlled experiments (Hjortskov et al. 2004; McDuff et al. 2014; Nasoz et al. 2004; Zhai and Barreto 2006; Zhou et al. 2011). In an engineering context, users are more prone to noise, which might influence the correlation between physiology and mental state (Balters and Steinert 2015). There are several studies that investigate more complex tasks, such as driving (Healey and Picard 2005), aviation (Nixon and Charles 2017; Wilson 2002), nuclear power plants (Gao et al. 2013), and ship navigation (Cohen et al. 2015).

Ship navigation is of special interest for the authors, especially because there is increasing activity in the field of remote and autonomous shipping. The YARA Birkeland will be one of the first autonomous ships in operation and will be operating on the Norwegian coast as of 2019 with a small crew for supervision, and then autonomously as of 2020. We believe that this is only the beginning of a shift in the maritime industry toward remote and autonomous vessels. This shift means that there will be people operating and monitoring ships from onshore control rooms. There is little research on how this type of operation will affect the people performing the tasks, which in turn may have unknown consequences on safety and efficiency of operation.

The aim of this paper is to investigate the mental state aspect of people who remotely operate and monitor ships. More specifically, we set out to understand how mental state and physiology change between different operational situations, and if mental state can be estimated from physiological responses. This can then form a basis for understanding how mental state is related to performance by monitoring physiological state in future experiments, which again can help designers when developing and evaluating new concepts.

We pursue this goal through an experiment investigating how physiological changes are related to changes in mental state in the context of ship bridges. The experiment was conducted with consumer ship simulation software. 31 participants from a student population were tasked to navigate a large ship in two scenarios: on open sea and in a narrow harbor. Stimuli were designed and verified in cooperation with an industry partner, one of the world's largest suppliers of ship bridge systems. Our contacts were professional ship simulator instructors with extensive experience operating large ships. The stimuli were created to be as realistic as possible—i.e., to replicate demands and actions encountered in a real context under normal operation. Tasks were consciously selected to represent the majority of activities on large ships, not extreme events that occur seldom, if ever. This will most likely result in smaller effect sizes in changes of mental state and physiology compared to extreme situations. Data were collected through surveys and physiology sensors. Participants rated their own affective state and workload through survey questions. Physiological state was evaluated through electrocardiography (ECG) and electrodermal activity (EDA). The results showed significant changes in mental and physiological state between the two scenarios. Concepts of stress and workload were correlated with EDA and elements of heart rate variability (HRV). These results indicate that users may have changing demands in their interfaces and system behavior as their affective state changes and that the experienced stress and workload of users are related to EDA and HRV.

2 Background: Mental state

In this work, we are interested in the mental state of users. In our definition of mental state, we include the constructs of affective state and workload. Both are of interest to the human–computer interaction community, because these concepts might influence users' behavior and how they perceive their situations. Below we present the constructs of affective state and workload, along with common subjective (Table 1) and physiological (Table 2) measurement tools for the respective constructs.

2.1 Affective state

We use the definition of Balters and Steinert (2015), which describes affect, or emotions, as a set of variables that might moderate behavior. We define affective state as the manifestation of the concept affect—i.e., the value of these variables at a specific time for a specific individual.

There are two main schools on how to describe affect in the field of psychology. The first considers affect as a set of discrete categories (Tomkins 1962; Ekman and Friesen 1969, 1971; Ekman 1992). Tomkins and McCarter (1964) describe eight categories of emotions, naming them at medium and high levels. They are interest–excitement, enjoyment–joy, surprise–startle, distress–anguish, fear–terror, shame–humiliation, contempt–disgust, and anger–rage. Ekman et al. (1972) define the six basic emotions as happiness, surprise, anger, disgust, and fear. Later, Ekman expanded the number of emotions to amusement, anger,

Table 1 Subjective evaluation of mental state

	Model	Dimensions	Method	Authors
Affective state	Circumplex model of affect	Arousal–sleepiness, Pleasure–displeasure	Affect grid, rate along separate dimensions	Russell (1980)
	PANAS	Positive affect, Negative affect	Two ten-item mood scales	Watson et al. (1988)
	AD-ACL	Energetic arousal, Tense arousal	20-25 activation-descriptive adjectives	Thayer (1986)
Workload	NASA TLX	Mental-, Physical-, Temporal demand, Performance, Effort, Frustration	Weighted sum of dimensions	Hart and Staveland (1988)
	SWAT	Time load, Mental effort load, Stress load	Weighted sum of dimensions	Reid and Nygren (1988)
	OW	Overall workload	Single scale	Vidulich and Tsang (1987)
	MCH	Overall workload	Decision tree	Wierwille and Casali (1983)

Table 2 Studies using physiological measurements to evaluate mental state

	Authors	Measurement
Affective state	Healey and Picard (2005)	Heart rate variability, electrodermal activity, respiration, muscle tension
	Pedrotti et al. (2014)	Pupil dilation, electrodermal activity
	Zhou et al. (2011)	Electrodermal activity, respiration, muscle tension, brain activity
	Katsis et al. (2008)	Heart rate variability, electrodermal activity, respiration, muscle tension
	Kim et al. (2004)	Heart rate variability, electrodermal activity, skin temperature,
	Schmidt et al. (2016)	Heart rate variability, electrodermal activity, skin temperature, respiration
	Baltaci and Gokcay (2016)	Pupil dilation, skin temperature
	Horlings et al. (2008)	Brain activity
	Mandryk and Atkins (2007)	Heart rate variability, electrodermal activity, muscle tension
Workload	Nixon and Charles (2017)	Heart rate variability, respiration
	McDuff et al. (2014)	Heart rate variability
	Gao et al. (2013)	Pupil dilation, Heart rate variability, eye blink
	Hjortskov et al. (2004)	Heart rate variability, blood pressure
	Nourbakhsh et al. (2012)	Electrodermal activity
	Brookings et al. (1996)	Heart rate variability, respiration, brain activity

contempt, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure, and shame (Ekman Ekman 1999). The second school of thought considers affect as a combination of multiple dimensions (Russell 1980; Russell and Barrett 1999; Thayer 1967; Watson and Tellegen 1985). Watson and Tellegen describe affect as a combination of the dimensions of positive affect and negative affect (Watson et al. 1988; Watson and Tellegen 1985). Positive affect is related to the extent a person feels enthusiastic, active, and alert. Negative affect is related to subjective feelings of distress and unpleasurable engagement. Positive affect and negative affect are measured through the Positive and Negative Affect Schedule (PANAS) scales, asking people to rate two ten-item mood scales assessing positive and negative affect on a five-point scale. Thayer (1967, 1978, 1986) explains affect as the two dimensions of energetic arousal and tense arousal, consisting of the four factors general activation (energy), deactivation-sleep (tiredness), high activation (tension), and general deactivation (calmness). This is measured through the activation–deactivation adjective checklist by rating 20–25 activation-descriptive adjectives on a four-point scale. Russell et al. (Russel et al. 1989; Russell 1979, 1980) propose a model of affect with the main dimensions arousal and pleasure–displeasure in which different affective states can be described as combinations of these two dimensions—e.g., stress being the combination of displeasure and high arousal. Levels of arousal and pleasure can be assessed either through the single-item affect grid (Russel et al. 1989) or through rating arousal and pleasure along separate dimensions. Russell (1979) shows that Thayer's dimensions of energetic arousal and tense arousal can be seen as an approximate rotation of arousal and pleasure. Watson and Tellegen's dimensions can—according to Russel et al. (1989)—be seen as a 45° rotation of the same dimensions.

2.2 Workload

The notion of workload or cognitive load has been used in human factor research in relation to performance. Parasuraman et al. (2008) argue that workload is one of the few constructs that is predictive of both performance in complex human–machine interactions and of the mental state of the operator. Cooper and Harper (1969) define workload as the sum of physical and mental effort and attention required to maintain a given level of performance. When viewing workload as a function of effort, one should consider both the capabilities of operators and their state. This could be how the skill level of operators might influence their effort, as well as their physical and mental state—i.e., tired or stressed. Parasuraman et al. (2008) describe workload as a function of the demand on mental resources in relation to the resources available from the human operator. Hart and Staveland (1988) describe workload as a multidimensional construct describing the cost incurred by a human operator to achieve a particular level of performance. Workload is not an inherent property but emerges from the interaction among task requirements, context, operator skills, behavior, perceptions, and affective state (Hart and Staveland 1988; Sheridan and Stassen 1979; Xie and Salvendy 2000). These definitions of workload are human centered, focusing on the subjective perception of workload. The notion of workload as a subjective experience is supported by Johanssen et al. (1979) and Sheridan (1980). The reason why the subjective experience of workload is important, according to Hart and Staveland (1988), is that this might alter behavior. Should an operator experience a situation as a high workload, they might adopt strategies to mitigate workload and experience distress.

Subjective workload is usually measured through surveys. Commonly used tools are the NASA Task Load Index (TLX) (Hart and Staveland 1988), Subjective Workload Assessment Technique (SWAT) (Reid and Nygren 1988), Modified Cooper–Harper scale (MCH) (Wierwille and Casali 1983), and Overall Workload (OW) (Vidulich and Tsang 1987). TLX and SWAT use multiple dimensions to assess workload, offering better diagnostic properties than one dimension when trying to assess the underlying mechanisms of workload. MCH and OW are unidimensional, providing less detail, but with the advantage of being faster to fill out. In addition, there is evidence that univariate methods have greater sensitivity than multivariate methods when estimating OW (Hendy et al. 1993; Vidulich and Tsang 1987).

2.3 Measuring mental state through physiology

Affective state and workload have been shown to be reflected in physiological responses (Andreassi 2010; Boucsein 2012; Ekman et al. 1983; Levenson 2003; Wilson and Eggemeier 1991). These physiological responses are controlled by the autonomic nervous system (ANS). The ANS is divided into two branches, the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). The PNS and SNS work antagonistically to regulate physiological arousal (Appelhans and Luecken 2006). The SNS can be described as responsible for the body's fight-or-flight reactions, whereas the PNS has been said to be in charge of rest and digest. Sympathetic and parasympathetic activities are expressed in various physiological phenomena, such as heart rate, respiration, EDA, brain activity, muscle tension, pupil dilation, skin temperature, and blood pressure. For an exhaustive overview of the use of physiological measurements in an engineering context, we refer to Balters and Steinert (2015). In Table 2, we show examples of previous work on correlating mental state with physiological changes.

3 Method: experiment investigating mental state in a ship simulator

An experiment was created to investigate how self-reported mental state relates to physiology in large ship navigation settings. Our focus has been to develop an experiment that addresses regularly occurring situations when operating a large ship. This means that stimuli have been designed to replicate reality rather than to elicit specific mental states or reactions. The experiment was conducted in commercial ship simulation software with student participants. For an exhaustive description of the experiment setup and execution, we refer to Dybvik et al. (2018).

4 Stimuli: open sea and harbor

Participants were asked to steer a 200-m-long cruise ship using a commercial ship simulation software in two scenarios. These scenarios were created in cooperation with experienced ship simulator instructors from one of the world's largest suppliers of ship bridge systems. The stimuli were created to be as realistic as possible—i.e., to replicate demands and actions encountered in a real context under normal operation. It was our goal to create scenarios that represent the majority of activities on large ships, not extreme events that occur seldom, if ever. This decision will likely result in smaller effect sizes in changes of mental state and physiology compared to extreme situations.

The first scenario was designed to replicate the task of sailing on an open sea. This can be described as long periods of time with little activity, mostly spent monitoring systems. Participants were tasked to sail the ship out from port, across an empty body of water, and into a new port. The only additional stimuli given during this scenario was a low-frequency

(LF) engine rumble to create a realistic backdrop. The scenario lasted 15 min, but the duration was not communicated to participants in advance to reduce any expectation effects towards the end. When the scenario ended, the ship would be approximately halfway between the two ports.

The second scenario was supposed to recreate tasks associated with harboring. This includes navigating narrow channels and performing secondary tasks related to going to harbor under a time constraint. Participants were instructed to navigate to a berth marked on their map. When the ship started moving, a 10-min visible timer could be seen on the screen, instructing participants to reach their destination before time ran out. The secondary tasks consisted of asking prerecorded questions regarding crew and cargo over the radio at regular intervals. Crew and cargo lists were printed and placed face-down on the table in front of the participants. The lists were turned over by the participants at the beginning of the second scenario after the participants were given instructions to do so on screen. Participants were instructed in advance to give their answers using a handheld walkie-talkie. Questions were repeated after 90 s if no reply was given or upon request from participants. Should participants reach their designated berth, a new destination would be given. This was intended to make it nearly impossible to finish the primary task within the allotted 10 min. Throughout the second scenario, radio chatter and LF engine noise were added to create a realistic backdrop.

5 Data collection: subjective and physiological

42 participants were sampled from an engineering student population. Eleven were excluded because of technical issues or failure to follow instructions. 31 participants were included in the analysis—13 females and 18 males. Ages of the participants ranged from 19 to 33 years (24.0 ± 2.74). To address our research questions, we collected both physiological data and subjectively assessed affective state and workload.

5.1 Subjective measurements

Affective state and workload were assessed through survey questions. We adopted the circumplex model of affect (Russell 1980), evaluating arousal and pleasure, as a framework for affective state. Pilot studies showed that the participants had trouble understanding the concept of arousal. Thus, alertness and awakeness were added to arousal and pleasure in an attempt to triangulate the concept of arousal described by the circumplex model of affect. Because stress is a known term for most people, it was included to see how a subjective rating of arousal and pleasure would relate to perceived levels of stress. According to the circumplex model of affect, stress should show up as a combination of high arousal and displeasure. Workload was assessed in two ways: as a single dimension of OW (Vidulich and Tsang 1987) and through the multidimensional TLX (Hart and Staveland 1988) scheme. These were selected because TLX and OW have been shown to be superior in terms of sensitivity and user acceptance (Hill et al. 1992). OW can provide an overview of subjective workload, whereas TLX gives a more detailed view of which dimensions influence subjective workload in our context. Affective state and overall workload were rated on eleven-point Likert scales. TLX was rated on seven-point Likert scales plus pairwise comparisons of the six dimensions.

5.2 Physiological measurements

Selection of the physiological data type and related sensors to use in the experiment was guided by the feasibility of integration in future products. By this, we mean sensors that could be worn by users without interfering with normal operation—e.g., wireless and comfortable to wear. In this experiment, physiological data were collected through ECG and EDA. Heart rate and HRV can be calculated from ECG. Heart rate and HRV are associated with both sympathetic and parasympathetic nervous system activity. Sympathetic activity tends to increase heart rate and decrease HRV, and vice versa for parasympathetic activity (Appelhans and Luecken 2006; Berntson et al. 1997; Camm et al. 1996). High-frequency (HF) variations (0.04–0.15 Hz) in heart rate are believed to be parasympathetically mediated, whereas LF variations (0.15–0.4 Hz) are considered a product of both parasympathetic and sympathetic activities (Berntson et al. 1997; Camm et al. 1996; Malliani et al. 1991). The normalized frequency components of LF and HF HRV are supposed to assess sympathetic and parasympathetic activity, respectively (Furlan et al. 2000; Pagani et al. 1997).

EDA is influenced only by sympathetic nervous system activity (Boucsein 2012; Dawson et al. 2007). EDA can be divided into phasic and tonic activity. Phasic and tonic activities are, according to Dawson et al. (2007), related to attention and activation, respectively. Phasic activity, or skin conductance response (SCR), is elicited by almost any stimulus that is novel, unexpected, or potentially important (Siddle 1991). Tonic activity, or skin conductance level (SCL), is related to continuous stimuli—e.g., performing a task (Bohlin 1976). ECG data were collected using the Shimmer3 ECG device (Shimmer3 ECG/EMG Unit, 2017) with a sampling rate of 512 Hz. EDA data were collected with the Shimmer3 GSR + device (Shimmer3 GSR + Unit, 2017) with a sampling rate of 128 Hz. Both devices transmitted data wirelessly via Bluetooth to a central computer running iMotions 6.4 (iMotions 2017) for synchronization and storage.

5.3 Procedure

The IMotions (iMotions 2017) software platform was used to present stimuli and collect and synchronize subjective and physiological data. After the participants expressly consented to the experiment, physiology sensors were attached to them, and they were seated in front of a computer screen. The ECG sensor was attached with five leads on the chests of participants per the instructions provided by Shimmer, with the V_x lead in position six. The EDA sensor was attached to the middle part of the index and middle fingers on the left hand. After the sensors were attached, the participants were seated in front of a computer screen in the simulator environment (Fig. 1). The participants were instructed on how the experiment would proceed and told that they should follow instructions given either onscreen or via audio.

Figure 2 shows the experimental procedure as a timeline. First, the participants were asked to fill out a survey on their affective state to provide a baseline. This was followed by information about the ship they were supposed to steer and a video demonstrating the controls of the ship. After the instructions were completed, the first scenario was presented, and the participants were supposed to sail the ship out of harbor and across an open expanse of sea. At the end of the first scenario, surveys on affective state and workload were filled out. This was repeated for the second scenario. After both scenarios were completed and surveys on mental state were filled out, an additional survey on demographics was filled out. This concluded the experiment, and the participants were debriefed and sensors were disconnected.

5.4 Analysis: classical statistics and multivariate analysis

Subjective measurements used for analysis were collected after each scenario. Measurements used for analyzing the physiological state were sampled from the last 5 min of each scenario (see Fig. 2). This was intended to avoid carryover effects from previous stimuli.

Heart rate and HRV in the time and frequency domains were calculated from ECG data using Kubios HRV

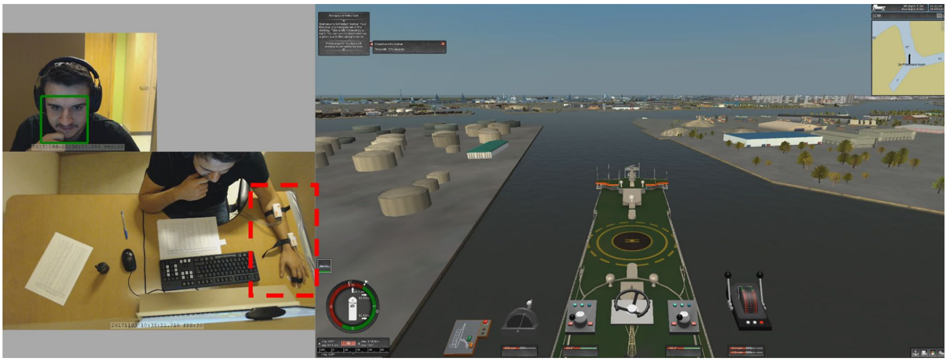
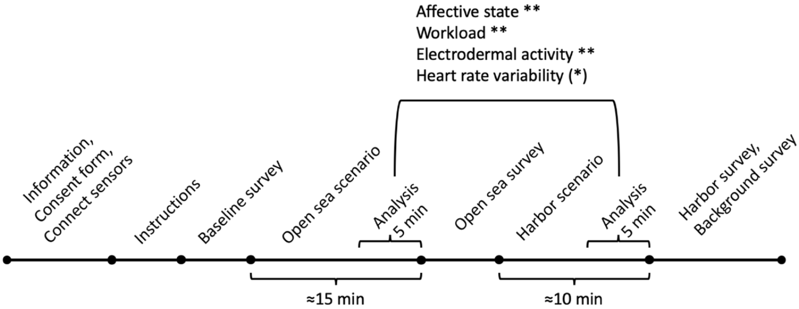


Fig. 1 Experiment environment (Harbor scenario), both physical and virtual. Physiological sensors are highlighted in the red dashed rectangle, ECG (top) and GSR (bottom) (Dybvik et al. 2018)

Fig. 2 Experimental procedure



(Tarvainen et al. 2014). The frequency domain of HRV is typically calculated using either fast Fourier transformation or autoregressive modeling (AR). We have selected AR because of its increased robustness and accuracy for shorter periods (Malliani et al. 1991; Montano et al. 2009). EDA data were processed in Ledalab using continuous decomposition analysis (CDA) (Benedek and Kaernbach 2010). CDA has been shown to be more sensitive to peak detection and estimation of tonic activity as opposed to through-to-peak algorithms (Benedek and Kaernbach 2010).

Table 3 provides an overview of all physiological variables included in the analysis. Assumptions of normal distribution, significant outliers, and skewness were evaluated, and statistical tests selected accordingly. The paired-sample *t* test investigates whether there is a difference in mean values between populations. The Wilcoxon signed-rank test and sign test investigate whether there is a difference in median values between populations. In addition to statistical tests comparing populations, the results were analyzed through correlation tests and multivariate analysis (i.e., principal component analysis (PCA) and partial least-squares regression (PLSR)) to investigate the relationship between different variables. For PCA and PLSR, each variable was mean centered and standardized (1/std.dev. of each variable). Multivariate analysis was performed using the software program (The Unscrambler X 2018).

6 Results

Three outliers were removed from the arousal values owing to inconsistencies in subjective reporting of arousal compared to awakeness and alertness. In these cases, arousal levels were reported much lower than awakeness and alertness, and we assume the nonnative English-speaking participants misinterpreted arousal as sexual arousal. This was a known misunderstanding from pilot studies. The data used for both classical statistics and multivariate analysis can be found in Online Resource 1.

7 Subjective variables: affective state and workload

Table 4 shows the results from statistical tests of subjective variables. Data used for the analysis are based on surveys filled out after each scenario. Participants show significant ($p < 0.01$) changes in self-reported affective state and workload between the two scenarios, open sea and harbor conditions, for paired samples with the exception of subjective experience of performance ($p = 0.694$), as previously shown by Dybvik et al. (2018).

Table 3 Physiology variables

	Variable	Description
Heart rate variability	Mean RR (ms)	Average RR interval
	SDNN (ms)	Standard deviation of RR interval
	Mean HR (bpm)	Average heart rate
	SD HR (bpm)	Standard deviation of heart rate
	RMSSD (ms)	Root mean square of successive differences. Measure of short-term variability
	NN50 (n.u.)	Number of successive intervals that differ more than 50 ms
	pNN50 (%)	Relative number of successive intervals that differ more than 50 ms
	LFpeak, HFpeak (Hz)	Peak frequencies for LF and HF bands
	LFpow, HFpow (ms^2)	Absolute power in LF and HF bands
	LFpow, HFpow (n.u.)	Powers of LF and HF bands in normalized units $LF[n.u.] = \frac{LF[ms^2]}{LF[ms^2]+HF[ms^2]}$ $HF[n.u.] = \frac{HF[ms^2]}{LF[ms^2]+HF[ms^2]}$
	TOTpow (ms^2)	Total spectral power
	LF_HF_ratio (n.u.)	Ratio between LF and HF powers
Electrodermal activity	nSCR (n.u.)	Number of significant skin conductivity responses (SCR)
	AmpSum (μS)	Sum of SCR amplitudes
	SCR (μS)	Average phasic driver
	PhasicMax (μS)	Maximum value of phasic activity
	Tonic (μS)	Mean tonic activity
	RawMean (μS)	Mean skin conductivity (SC) value

Table 4 Subjective variable results

Variable	$\bar{X}_{\text{sea}}$	σ_{sea}	$\bar{X}_{\text{harbour}}$	σ_{harbour}	$\bar{\Delta}$	σ_{Δ}	Paired samples	Effect size (d)
Arousal 0–10[d]	6.07	1.98	7.32	1.89	1.25	1.71	0.001^{b**}	0.73
Awake 0–10	6.48	2.05	7.61	2.04	1.13	1.09	$< 0.001^{c**}$	1.04
Alert 0–10	6.26	1.95	7.42	1.86	1.16	1.61	$< 0.001^{c**}$	0.72
Pleasant 0–10	6.35	1.62	4.68	1.80	− 1.68	1.54	$< 0.001^{a**}$	1.09
Stress 0–10	3.23	1.94	6.39	2.14	3.16	1.88	$< 0.001^{a**}$	1.68
Overall Workload 0–10	2.03	1.70	8.03	1.78	6.00	2.93	$< 0.001^{c**}$	2.05
Mental demand 1–7	2.55	1.41	6.13	0.85	3.58	1.43	$< 0.001^{a**}$	2.50
Physical demand 1–7	1.58	0.85	3.74	1.91	2.16	1.83	$< 0.001^{c**}$	1.18
Temporal demand 1–7	1.94	0.93	6.26	0.96	4.32	1.28	$< 0.001^{c**}$	3.39
Performance 1–7	3.90	2.07	4.13	1.88	0.23	3.17	0.694^{a}	0.07
Effort 1–7	2.71	1.47	5.90	1.14	3.19	1.64	$< 0.001^{b**}$	1.95
Frustration 1–7	2.74	1.69	4.42	1.82	1.68	2.09	$< 0.001^{a**}$	0.80
TLX 1–7	2.86	1.23	5.60	0.74	2.73	1.37	$< 0.001^{a**}$	1.99

[a] Paired samples t test, [b] Wilcoxon signed-rank test, [c] Sign test, $*p < 0.05$, $**p < 0.01$

[d] Three outliers were removed due to inconsistencies in subjective reporting of arousal compared to awakeness and alertness. 28 data points used for this specific analysis

8 Physiological variables: heart rate variability and electrodermal activity

The results from the statistical tests can be seen in Table 5. Heart rate variables such as mean and standard deviation of heart rate show significant ($p < 0.05$) changes. Significant changes are found in the frequency domain for normalized LF and HF powers, whereas nonnormalized powers do not exhibit this change. The LF to HF ratio of HRV, which is a common metric used to describe physiological arousal, has a positive change from the open sea scenario

Table 5 Physiological variable results

Variable	$\bar{X}_{\text{sea}}$	σ_{sea}	$\bar{X}_{\text{harbour}}$	σ_{harbour}	$\bar{\Delta}$	σ_{Δ}	Paired samples	Effect size (d)
Mean RR (ms)	802.83	134.31	726.55	116.73	− 76.29	66.05	$< 0.001^{a**}$	1.16
SDNN (ms)	41.86	18.16	42.52	17.14	0.66	12.49	0.782^{a}	0.05
Mean HR (bpm)	76.63	11.87	84.58	13.00	7.95	6.79	$< 0.001^{a**}$	1.17
SD HR (bpm)	4.63	1.65	5.57	1.67	0.94	1.17	$< 0.001^{a**}$	0.80
RMSSD (ms)	34.34	19.95	30.91	16.77	− 3.43	9.67	0.072^{a}	0.35
NN50 (beats)	46.14	43.48	43.71	42.33	− 2.43	29.47	0.400^{b}	0.08
pNN50 (%)	13.75	14.55	11.67	12.12	− 2.08	8.69	0.215^{a}	0.24
LFpeak (Hz)	0.093	0.014	0.097	0.018	0.004	0.020	0.036^{c*}	0.18
HFpeak (Hz)	0.214	0.070	0.155	0.024	− 0.059	0.072	0.001^{b**}	0.82
LFpow (ms^2)	1273.34	1030.42	1308.92	877.50	35.58	929.32	0.855^{b}	0.04
HFpow (ms^2)	549.38	851.69	497.28	655.89	− 52.10	309.36	0.185^{c}	0.17
LFpow (n.u.)	71.81	13.47	76.60	10.17	4.79	9.57	0.013^{a*}	0.50
HFpow (n.u.)	28.13	13.45	23.34	10.15	− 4.79	9.55	0.013^{a*}	0.50
TOTpow (ms^2)	1989.34	1836.75	2017.95	1558.11	28.60	1225.82	0.964^{b}	0.02
LF_HF_ratio	3.78	3.15	4.45	3.50	0.68	2.44	0.053^{b}	0.28
nSCR (n.u.)	62.10	44.60	111.45	35.96	49.35	44.79	$< 0.001^{a**}$	1.10
AmpSum (µS)	5.99	8.04	12.96	10.27	6.97	6.84	$< 0.001^{c**}$	1.02
SCR (µS)	0.006	0.006	0.011	0.008	0.005	0.005	$< 0.001^{a**}$	1.00
PhasicMax (µS)	2.03	2.05	2.77	2.28	0.73	1.91	0.041^{a*}	0.38
Tonic (µS)	2.47	1.39	3.75	1.70	1.28	0.86	$< 0.001^{c**}$	1.49
RawMean (µS)	2.56	1.44	3.93	1.76	1.37	0.90	$< 0.001^{a**}$	1.53

[a] Paired samples t test, [b] Wilcoxon signed-rank test, [c] Sign test, $*p < 0.05$, $**p < 0.01$

to the harbor scenario ($p = 0.053$). Data on electrodermal activity show highly significant changes between the two scenarios for all variables except for the maximum phasic driver ($p = 0.041$).

8.1 Relation within subjective variables

In the following two sections, we highlight interesting correlations from the analysis. The full correlation table can be found in Online Resource 2. Stress is defined according to Russell (1980) as a combination of high arousal and displeasure—i.e., the opposite of feeling pleasant. Our results show that arousal has a stronger correlation to stress ($r = 0.47$, $n = 62$, $p < 0.001$) than displeasure ($r = 0.25$, $n = 62$, $p = 0.01$). For this context, we interpret arousal as the main factor of stress. In a professional context such as large ship navigation, this makes sense, because we expect affective changes to be more linked to arousal—i.e., energy level—than feelings of displeasure. We find that workload, both OW and TLX, is associated with levels of stress, displeasure, and arousal, in that order.

The subjective variables were also analyzed using principal component analysis (PCA). Values are given as cross-validated results followed by calibration in parentheses. Cross-validation was performed by leaving one participant out at a time. The results show that the first component accounted for 46% (52%) of the variance in the subjective variables. This component was mainly dominated by workload and stress, as seen in Fig. 3a. The second component, accounting for 16% (18%) of variance, consisted mainly of arousal and displeasure (Fig. 3b). These first two components, explaining 62% (70%) of the total variance in the subjective data, align with Watson and Tellegen's (1985) model of affect consisting of the dimensions positive and negative affect.

What we see from both the statistical tests and the PCA is that the largest changes in mental state are in workload and stress. This could mean that in an environment where participants are supposed to perform professional tasks, emotions—i.e., arousal and displeasure—are less influenced than the more task-related concepts stress and workload. One other explanation for the smaller changes in self-reported affective state could be the semantic understanding of the different concepts, given that our experience reveals that it is more common to discuss concepts of stress and workload in a professional context as opposed to emotions. Evaluating one's own state of arousal and displeasure could be more challenging than stress and workload for nonnative English speakers owing to unfamiliarity of the concepts.

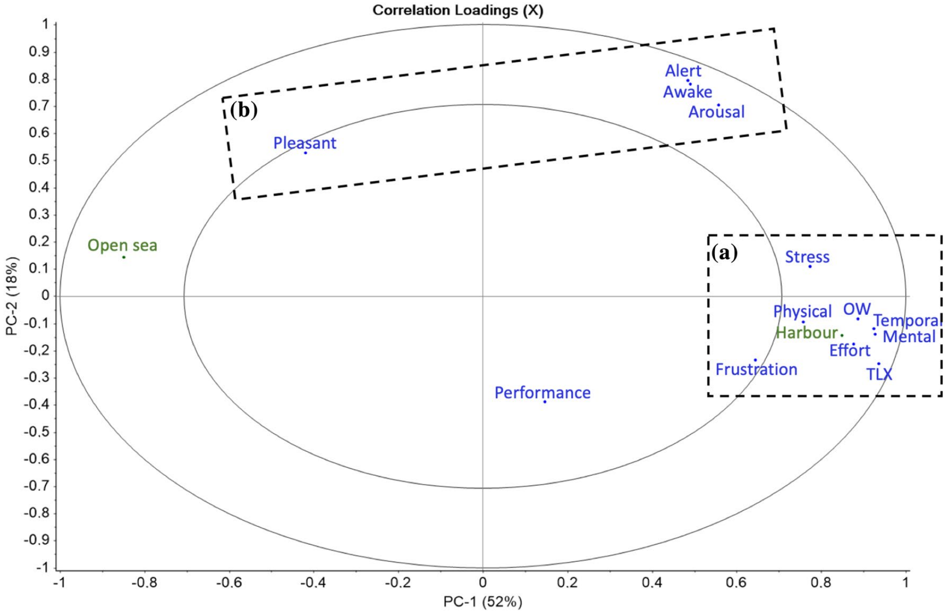


Fig. 3 PCA—correlation loadings for subjective variables

8.2 Relation between subjective and physiological variables

We find that the variables workload and stress have the strongest correlation to both HRV and EDA variables. Among the HRV variables, the HF peak frequency has the highest correlation to both stress and workload (stress: $r=-0.33$, $n=57$, $p<0.001$; OW: $r=-0.43$, $n=57$, $p<0.001$; TLX: $r=-0.61$, $n=57$, $p<0.001$). The number of skin conductivity responses (nSCRs), tonic level, and raw mean SC signal have the strongest correlation to stress and workload of the EDA variables. Displeasure was found to be correlated with nSCRs ($r=0.30$, $n=62$, $p=0.02$). Awakeness shows correlations to the LF peak of HRV ($r=0.37$, $n=57$, $p=0.01$) and HR standard deviation ($r=0.36$, $n=57$, $p=0.01$). Arousal and alertness were not correlated with any of the physiological variables ($r<0.30$). The strongest correlations to mental state were peak frequency in HF HRV and phasic and tonic EDA.

The relation between physiological and subjective variables was also investigated through PLSR (Fig. 4). The values are given as cross-validated results followed by calibration in parentheses. Cross-validation was performed by leaving one participant out at a time. We found that the first component accounts for 13% (19%) of variance in the subjective variables (marked in red) and 20% (29%) in the physiological variables (marked in blue). The second component accounted for 30% (30%) of the variance in the physiological variables but did not contribute to explaining the variance in the subjective variables. Here, the explained variance was −2% (2%), indicating a small amount of overfitting.

The first component was dominated by HF peak, EDA variables, mean RR/HR, and to some extent normalized LF and HF components in terms of physiological variables (Fig. 4a, b). The mental state variables in the first component mainly consisted of workload and stress variables (Fig. 4a). The second component consisted of absolute power and time-domain HRV variables (Fig. 4c). The two scenarios (marked in green), cruising on open sea and docking in a narrow harbor, were down-weighted when calculating the factors. By including these variables in the model, we see the tendencies of how the variables changed between the two scenarios. Although the explained variance was relatively low, we found an emerging pattern of how the variables covaried. We observed a decrease in peak frequency HF HRV when workload and stress increased—i.e., HRV moved toward lower frequencies, which are associated with lowered parasympathetic activity, and could be a sign of increased sympathetic activity. This corresponds well to the increase in EDA for increasing levels of workload and stress, which is mediated solely by the SNS. The increase in both

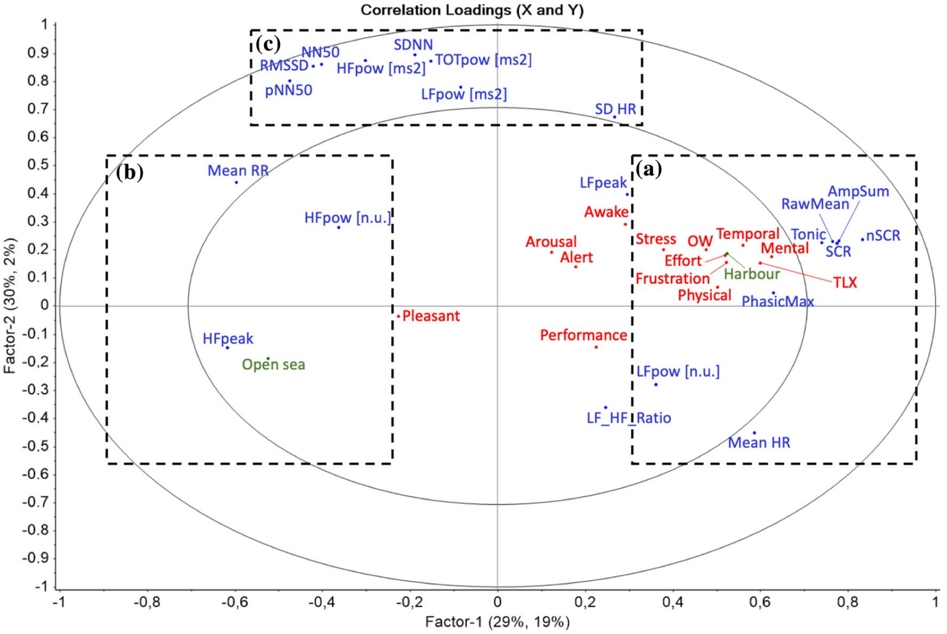


Fig. 4 PLSR—correlation loadings for physiological to subjective variables

phasic and tonic EDA could be explained by the increased number of external stimuli and task demands, which should influence phasic and tonic levels, respectively.

9 Discussion

Our results show significant changes in both mental and physiological states between the two scenarios in the experiment. There is a general increase in arousal, displeasure, stress, and workload in the scenario where participants navigated a large ship in a narrow harbor compared to the scenario where they navigated on open sea. The same increase is found for EDA and partially for HRV. Through PLSR, we find that workload and stress make up the majority of variations in the mental state of participants. Most variation in physiological data along the same component can be found in EDA and the HRV HF band peak frequency. When calculating correlations between mental and physiological state variables, we find that the strongest relations are between the dominant variables found in the PLSR analysis—namely, stress and workload for mental state and EDA and the HF HRV peak for physiological state. These correlation scores range from approximately 0.4 to 0.6. The relation between mental and physiological state variables might be influenced by the following factors: First, physiological responses are highly individual. If given the same stimuli, one participant might show strong changes in EDA, whereas another has little to no change. Second, participants might have different reactions to the stimuli. What is considered stressful for one person might be routine for the other. Third, there might be differences in how individuals rate concepts of affective state and workload, owing either to language barriers or their own understanding of the concept in question. Finally, the analysis in this paper is based on average results from participants, meaning that individual differences are not considered. Although we did not consider individual differences, we found correlations between mental and physiological states. Our interpretation is that when these results can be found despite the aforementioned factors, analysis of multiple data points from individual participants should provide even stronger results.

Changes in HRV LF/HF ratio are similar to the findings of Gao et al. (2013) and Hjortskov et al. (2004), although our LF/HF ratio is higher. McDuff et al. (2014) found a much larger change in the LF/HF ratio, doubling the mean value from approx. 0.5 to 1.0 between the resting and stress conditions. We believe that our large ratio values are due to a lower HF power in our data compared to that of Hjortskov et al. This could indicate parasympathetic withdrawal, which could be an indicator of high baseline physiological arousal. Total power is comparable to the studies of Hjortskov et al. and Gao et al., although the latter report a much lower total power for their low-complexity task. Healey and Picard (2005) show that driver stress level was correlated most strongly to mean skin conductivity and HRV, when compared to respiration rate and electromyography in addition to the aforementioned variables. Out of these, mean skin conductivity had the strongest correlation to their stress metric. Wilson (2002) found EDA and heart rate to be more sensitive to changes in cognitive demand than HRV, which is similar to our results when comparing physiological data to the subjective ratings.

From the experiment setup, we know that the harbor scenario, where higher levels of perceived stress and workload were measured, contained more stimuli. EDA, or more specifically SCR, has been shown to be highly reactive to external stimuli (Siddle 1991). We observe a relative increase in number of SCRs in our data between the open sea and harbor scenarios, corresponding to the known reactivity to external stimuli. At the same time, we observe an increase in the tonic level of EDA as well as stress and workload. We show that there is a relation among the number of stimuli, subjectively assessed mental state, and electrodermal activity in the context of large ship navigation. However, we do not claim causality between the variables, only that they covary.

By designing tasks that replicate normal working conditions onboard a ship, it is our belief that the results of this paper are representative of the real context. However, the results are limited to the relevance of the participants partaking in the experiment. When interpreting the results of this experiment, it is important to keep in mind that the data were collected from a student population performing tasks in an unfamiliar environment, steering a large ship in a simulator. Even if the results are consistent for a student population, we do not know how this relates to the responses of professionals if they were to participate in the same experiment. We believe that students' reactions to stimuli might be stronger than would be the case for professionals owing to their training and experience in handling the given tasks. Our assumption is that the direction of responses would be similar for students and professionals, although the magnitude might differ. This should be tested in a future experiment.

10 Conclusion: mental state and physiology

In this paper, we have presented an experiment aiming to investigate changes in mental state and physiology between different scenarios and how mental state and physiology are related in the context of large ship navigation. The motivation behind the experiment has been to create a foundation for future work on continuously monitoring users' mental state in the context of remote operation and monitoring of ships.

The experiment tests the feasibility of using changes in electrodermal activity and heart rate variability to act as a proxy for users' self-reported mental state. 31 participants from a student population were tasked with navigating a large ship in two scenarios: on open sea and in a narrow harbor. The results show that there are significant changes in the variables used to measure mental state between the two scenarios. We found significant changes in EDA and several variables representing HRV. We found that self-reported stress and workload were correlated with EDA and the peak frequency in the HF band of HRV. Awakeness was found to be positively correlated with the LF band of HRV and the standard deviation of heart rate. No correlation was found between arousal or alertness and physiological variables. Multivariate analysis showed that one component could explain 13% of the variance in mental state and 20% in physiological data. The second component did not contribute to explaining any variance in the mental state of users but did explain 30% of the variability of physiological data. The first component was dominated by workload, stress, EDA, and elements of HRV. The second component was dominated by absolute power- and time-domain heart rate variability.

We draw the following conclusions from our experiment: One, there is a significant difference in variables used to measure mental and physiological states between two regularly occurring scenarios in the context of large ship navigation. As these changes occur, the capacity and reaction pattern of the user change, and the user could have different or changing demands of the user interface and system behavior. Two, elements of mental state are correlated with changes in physiological state. Most prominently, stress and workload covary with EDA and elements of HRV. This finding can serve as a foundation for how to assess changes in mental state of users remotely operating and monitoring ships through measuring changes in physiological state.

We believe our findings to be representative of similar control rooms. To verify this, experiments should be conducted in situ by collecting data on professionals working in their normal environments, such as ship bridges, power plants, airplanes, or air traffic control. Such experiments would be more prone to noise and uncontrollable variables that may influence results. Despite the challenges of noisy data and uncontrolled influencing factors, this would be the real situation we are interested in. Results from such an experiment would have the highest ecological validity, because it is the real situation. For research on how mental and physiological states are related in a control room context, this is the necessary step to find an answer. Should a connection between mental and physiological states be found from in situ experiments, it would be possible to evaluate the mental state of users unobtrusively—i.e., not interfering with their tasks by having to fill out surveys or answer questions. With this information, designers could compare concepts on how they influence the mental state of users, which again may influence task performance. We hope this can be a tool in the designers' toolkit on how to evaluate designs when working on remote operation of ships.

References

Andreassi JL (2010) Psychophysiology: human behavior and physiological response. Psychology Press, London

Appelhans BM, Luecken LJ (2006) Heart rate variability as an index of regulated emotional responding. Rev General Psychol 10(3):229. https://doi.org/10.1037/1089-2680.10.3.229

Bainbridge L (1983) Ironies of automation. Automatica 19(6):775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Baltaci S, Gokcay D (2016) Stress detection in human–computer interaction: fusion of pupil dilation and facial temperature features. Int J Hum–Comput Stud 32(12):956–966. https://doi.org/10.1080/10447318.2016.1220069

Balters S, Steinert M (2015) Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices. J Intell Manuf. https://doi.org/10.1007/s10845-015-1145-2

Benedek M, Kaernbach C (2010) A continuous measure of phasic electrodermal activity. J Neurosci Methods 190(1):80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028

Berntson GG, Thomas Bigger J, Eckberg DL, Grossman P, Kaufmann PG, Malik M et al (1997) Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology 34(6):623–648. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x

Bohlin G (1976) Delayed habituation of the electrodermal orienting response as a function of increased level of arousal. Psychophysiology 13(4):345–351. https://doi.org/10.1111/j.1469-8986.1976.tb03088.x

Boucsein W (2012) Electrodermal activity. Springer Science & Business Media, New York

Brookings JB, Wilson GF, Swain CR (1996) Psychophysiological responses to changes in workload during simulated air traffic control. Biol Psychol 42(3):361–377. https://doi.org/10.1016/0301-0511(95)05167-8

Camm AJ, Malik M, Bigger JT, Breithardt G, Cerutti S, Cohen RJ et al (1996) Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Circulation 93(5):1043–1065. https://doi.org/10.1161/01.CIR.93.5.1043

Cohen I, Brinkman W-P, Neerincx MA (2015) Modelling environmental and cognitive factors to predict performance in a stressful training scenario on a naval ship simulator. Cogn Technol Work 17(4):503–519. https://doi.org/10.1007/s10111-015-0325-3

Cooper GE, Harper RP (1969) The use of pilot rating in the evaluation of aircraft handling qualities (No. AGARD-567). Advisory Group for Aerospace Research and Development Neuilly-Sur-Seine (France)

Dawson ME, Schell AM, Filion DL (2007) The electrodermal system. Handb Psychophysiol 2:200–223

Dybvik H, Wulvik A, Steinert M (2018) Steering a ship-investigating affective state and workload in ship simulations. In: DS92: Proceedings of the DESIGN 2018 15th international design conference (pp. 2003–2014). https://doi.org/10.21278/idc.2018.0459

Ekman Paul (1992) An argument for basic emotions. Cogn Emot 6(3–4):169–200. https://doi.org/10.1080/02699939208411068

Ekman Paul, Friesen WV (1969) The repertoire of nonverbal behavior: categories, origins, usage, and coding. Semiotica 1(1):49–98. https://doi.org/10.1515/semi.1969.1.1.49

Ekman Paul, Friesen WV (1971) Constants across cultures in the face and emotion. J Pers Soc Psychol 17(2):124. https://doi.org/10.1037/h0030377

Ekman P, Levenson RW, Friesen WV (1983) Autonomic nervous system activity distinguishes among emotions. Science 221(4616):1208–1210. https://doi.org/10.1126/science.6612338

Ekman P (1999) Basic emotions. In: Scientist TDR, MJP of C. Psychology (eds) Handbook of cognition and emotion. Wiley, New York, pp 45–60. https://doi.org/10.1002/0470013494.ch3

Ekman P, Friesen WV, Ellsworth P (1972) Emotion in the human face: guide-lines for research and an integration of findings: guidelines for research and an integration of findings. Pergamon

Furlan R, Porta A, Costa F, Tank J, Baker L, Schiavi R et al (2000) Oscillatory patterns in sympathetic neural discharge and cardiovascular variables during orthostatic stimulus. Circulation 101(8):886–892. https://doi.org/10.1161/01.CIR.101.8.886

Gao Q, Wang Y, Song F, Li Z, Dong X (2013) Mental workload measurement for emergency operating procedures in digital nuclear power plants. Ergonomics 56(7):1070–1085. https://doi.org/10.1080/00140139.2013.790483

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv Psychol 52:139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Healey JA, Picard RW (2005) Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans Intell Transp Syst 6(2):156–166. https://doi.org/10.1109/TITS.2005.848368

Hendy KC, Hamilton KM, Landry LN (1993) Measuring subjective workload: when is one scale better than many? Hum Factors 35(4):579–601. https://doi.org/10.1177/001872089303500401

Hill SG, Iavecchia HP, Byers JC, Bittner AC, Zaklade AL, Christ RE (1992) Comparison of four subjective workload rating scales. Hum Factors 34(4):429–439. https://doi.org/10.1177/001872089203400405

Hjortskov N, Rissén D, Blangsted AK, Fallentin N, Lundberg U, Søgaard K (2004) The effect of mental stress on heart rate variability and blood pressure during computer work. Eur J Appl Physiol 92(1–2):84–89. https://doi.org/10.1007/s00421-004-1055-z

Horlings R, Datcu D, Rothkrantz LJM (2008) Emotion recognition using brain activity. In: Proceedings of the 9th international conference on computer systems and technologies and workshop for PhD students in computing, CompSysTech '08. ACM, New York, NY, pp 6:II.1–6:1. https://doi.org/10.1145/1500879.1500888

Isen AM (2001) An influence of positive affect on decision making in complex situations: theoretical issues with practical implications. J Consum Psychol 11(2):75–85. https://doi.org/10.1207/S15327663JCP1102_01

Isen AM, Reeve J (2005) The influence of positive affect on intrinsic and extrinsic motivation: facilitating enjoyment of play, responsible work behavior, and self-control. Motiv Emotion 29(4):295–323. https://doi.org/10.1007/s11031-006-9019-8

Jiao RJ, Zhou F, Chu C-H (2017) Decision theoretic modeling of affective and cognitive needs for product experience engineering: key issues and a conceptual framework. J Intell Manuf 28(7):1755–1767. https://doi.org/10.1007/s10845-016-1240-z

Johanssen G, Moray N, Pew R, Rasmussen J, Sanders A, Wickens C (1979) Final report of experimental psychology group. Mental workload. Springer, Boston, pp 101–114. https://doi.org/10.1007/978-1-4757-0884-4_7

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47(2):263–292

Kahneman D, Tversky A (1984) Choices, values, and frames. Am Psychol 39(4):341–350

Katsis CD, Katertsidis N, Ganiatsas G, Fotiadis DI (2008) Toward emotion recognition in car-racing drivers: a biosignal processing approach. IEEE Trans Syst Man Cyber Part A: Syst Human 38(3):502–512. https://doi.org/10.1109/TSMCA.2008.918624

Kim KH, Bang SW, Kim SR (2004) Emotion recognition system using short-term monitoring of physiological signals. Med Biol Eng Comput 42(3):419–427

Levenson RW (2003) Blood, Sweat, and Fears. Ann N Y Acad Sci 1000(1):348–366. https://doi.org/10.1196/annals.1280.016

Malliani A, Pagani M, Lombardi F, Cerutti S (1991) Cardiovascular neural regulation explored in the frequency domain. Circulation 84(2):482–492. https://doi.org/10.1161/01.CIR.84.2.482

Mandryk RL, Stella Atkins M (2007) A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. Int J Hum-Comput Stud 65(4):329–347

McDuff D, Gontarek S, Picard R (2014) Remote measurement of cognitive stress via heart rate variability. In: 2014 36th annual international conference of the IEEE engineering in medicine and biology society, pp. 2957–2960. https://doi.org/10.1109/EMBC.2014.6944243

Montano N, Porta A, Cogliati C, Costantino G, Tobaldini E, Casali KR, Iellamo F (2009) Heart rate variability explored in the frequency domain: a tool to investigate the link between heart and behavior. Neurosci Biobehav Rev 33(2):71–80. https://doi.org/10.1016/j.neubiorev.2008.07.006

Nasoz F, Alvarez K, Lisetti CL, Finkelstein N (2004) Emotion recognition from physiological signals using wireless sensors for presence technologies. Cogn Technol Work 6(1):4–14. https://doi.org/10.1007/s10111-003-0143-x

Nixon J, Charles R (2017) Understanding the human performance envelope using electrophysiological measures from wearable technology. Cogn Technol Work 19(4):655–666. https://doi.org/10.1007/s10111-017-0431-5

Nourbakhsh N, Wang Y, Chen F, Calvo RA (2012) Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In: Proceedings of the 24th Australian computer-human interaction conference. New York, NY, USA: ACM, pp 420–423. https://doi.org/10.1145/2414536.2414602

Pagani M, Montano N, Porta A, Malliani A, Abboud FM, Birkett C, Somers VK (1997) Relationship between spectral components of cardiovascular variabilities and direct measures of muscle sympathetic nerve activity in humans. Circulation 95(6):1441–1448. https://doi.org/10.1161/01.CIR.95.6.1441

Parasuraman R, Sheridan TB, Wickens CD (2008) Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J Cognit Eng Decis Mak 2(2):140–160. https://doi.org/10.1518/155534308X284417

Pedrotti M, Mirzaei MA, Tedesco A, Chardonnet J-R, Mérienne F, Benedetto S, Baccino T (2014) Automatic stress classification with pupil diameter analysis. Int J Hum Comput Stud 30(3):220–236

Picard RW (1997) Affective computing. The MIT Press, Cambridge, p 167, 170

Reid GB, Nygren TE (1988) The subjective workload assessment technique: a scaling procedure for measuring mental workload. In: Hancock PA, Meshkati N (eds) Advances in psychology, vol 52. Elsevier, North-Holland, pp 185–218

Russel JA, Weiss A, Mendelsohn GA (1989) Affect grid: a single-item scale of pleasure and arousal. J Pers Soc Psychol 57(3):493–502

Russell JA (1979) Affective space is bipolar. J Pers Soc Psychol 37(3):345. https://doi.org/10.1037/0022-3514.37.3.345

Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39(6):1161–1178. https://doi.org/10.1037/h0077714

Russell JA, Barrett LF (1999) Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J Pers Soc Psychol 76(5):805. https://doi.org/10.1037/0022-3514.76.5.805

Schmidt E, Decke R, Rasshofer R (2016) Correlation between subjective driver state measures andpsychophysiological and vehicular data in simulated driving. In: 2016 IEEE intelligent vehicles symposium (IV), Gothenburg, Sweden, 19–22 June 2016, pp 1380–1385. https://doi.org/10.1109/IVS.2016.7535570

Sheridan T (1980) Mental workload: what is it? Why bother with it. Human Factors Soc Bull 23(2):1–2

Sheridan TB, Stassen HG (1979) Definitions, models and measures of human workload. Mental workload. Springer, Boston, pp 219–233. https://doi.org/10.1007/978-1-4757-0884-4_12

Shimmer3 ECG/EMG Unit (2017) Dublin, Ireland: Shimmersense. http://www.shimmersensing.com/products/shimmer3-ecg-sensor. Accessed 12 Nov 2017

Shimmer3 GSR + Unit (2017) Dublin, Ireland: Shimmersense. Retrieved from http://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor. Accessed 12 Nov 2017

Siddle DAT (1991) Orienting, habituation, and resource allocation: an associative analysis. Psychophysiology 28(3):245–259. https://doi.org/10.1111/j.1469-8986.1991.tb02190.x

Tarvainen MP, Niskanen J-P, Lipponen JA, Ranta-aho PO, Karjalainen PA (2014) Kubios HRV—Heart rate variability analysis software. Comput Methods Programs Biomed 113(1):210–220. https://doi.org/10.1016/j.cmpb.2013.07.024

Thayer RE (1967) Measurement of activation through self-report. Psychol Rep 20(2):663–678. https://doi.org/10.2466/pr0.1967.20.2.663

Thayer RE (1978) Toward a psychological theory of multidimensional activation (arousal). Motiv Emotion 2(1):1–34. https://doi.org/10.1007/BF00992729

Thayer RE (1986) Activation-deactivation adjective check list: current overview and structural analysis. Psychol Rep 58(2):607–614. https://doi.org/10.2466/pr0.1986.58.2.607

The Unscrambler X (2018) (Version 10.5). Camo Software, Inc., Woodbridge

Tomkins S (1962) Affect imagery consciousness: volume I: the positive affects. Springer, New York

Tomkins SS, McCarter R (1964) What and where are the primary affects? Some evidence for a theory. Percept Mot Skills 18(1):119–158. https://doi.org/10.2466/pms.1964.18.1.119

Vidulich MA, Tsang PS (1987) Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proc Hum Factors Soc Ann Meet 31(9):1057–1061. https://doi.org/10.1177/154193128703100930

Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol 54(6):1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Watson D, Tellegen A (1985) Toward a consensual structure of mood. Psychol Bull 98(2):219. https://doi.org/10.1037/0033-2909.98.2.219

Wierwille WW, Casali JG (1983) A validated rating scale for global mental workload measurement applications. Proc Hum Factors Soc Ann Meet 27(2):129–133. https://doi.org/10.1177/154193128302700203

Wilson GF (2002) An analysis of mental workload in pilots during flight using multiple psychophysiological measures. Int J Aviat Psychol 12(1):3–18. https://doi.org/10.1207/S15327108IJAP1201_2

Wilson GF, Eggemeier FT (1991) Psychophysiological assessment of workload in multi-task environments. In: Multiple-task performance. Taylor & Francis, London, pp 329–360

Woods DD, Patterson ES, Roth EM (2002) Can we ever escape from data overload? A cognitive systems diagnosis. Cogn Technol Work 4(1):22–36. https://doi.org/10.1007/s101110200002

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

Zhai J, Barreto A (2006) Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In: 2006 International conference of the IEEE engineering in medicine and biology society, pp 1355–1358. https://doi.org/10.1109/IEMBS.2006.259421

Zhou F, Qu X, Helander MG, Jiao JR (2011) Affect prediction from physiological measures via visual stimuli. Int J Hum Comput Stud 69(12):801–819. https://doi.org/10.1016/j.ijhcs.2011.07.005

iMotions (2017) (Version 6.4). iMotions, Boston

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16

Appendix C2: Academic contribution 2

Dybvik, H., Wulvik, A., & Steinert, M. (2018). STEERING A SHIP - INVESTIGATING AFFECTIVE STATE AND WORKLOAD IN SHIP SIMULATIONS. Proceedings of the Design Society: DESIGN Conference, 2003–2014. https://doi.org/10.21278/idc.2018.0459

INTERNATIONAL DESIGN CONFERENCE - DESIGN 2018
https://doi.org/10.21278/idc.2018.0459

STEERING A SHIP - INVESTIGATING AFFECTIVE STATE AND WORKLOAD IN SHIP SIMULATIONS

H. Dybvik, A. Wulvik and M. Steinert

Abstract

We present an experiment investigating concepts of affective state and workload in a large ship manoeuvring context. It is run on a consumer ship simulator software where student participants (N=31) perform two ecologically valid scenarios: sailing on open sea and in a harbour. Results from surveys show highly significant changes in terms of both affect and workload between the scenarios. Thus, one should consider varying affects and workloads from users in varying contexts, consequently demanding new design paradigms for product development, such as dynamically adaptive interfaces.

Keywords: human behaviour, emotional engineering, engineering design, empirical studies, ocean space

1. Introduction: The human element and the ship bridge

The ship bridge is where the captain and his crew controls the ship. Navigation, monitoring systems, and communicating with both internal and external personnel are important activities. Sea piloting, i.e. sailing on open sea normally consists of monitoring tasks and no active navigation at all. Harbour piloting, i.e. sailing in harbours, requires continuous adjustment of speed and course, monitoring ship systems, and communicating with both crew and external contacts. These scenarios range from the monotone to the highly complex (Norros, 2004; Nilsson et al., 2009).

Maritime accidents occur in either scenario (Nilsson et al., 2009), mostly as the result of human error. Research shows that 49 to 96 percent of all shipping incidents or marine causalities are caused by human error (Rothblum, 2000; Hetherington et al., 2006; Tzannatos, 2010). Given the large share of maritime accidents caused by human error, this paper aims to direct attention towards the human users and their mental state during ship operation with the goal of identifying opportunities for reducing accidents. The notion of human centred design (Woodson and Conover, 1970; Sanders and McCormick, 1987) has existed since the 1960s. When considering humans in engineering, they are usually represented by generic models based on certain boundary conditions (Balters and Steinert, 2017). Models often represent the "average" human, with a general and stable behaviour response. Kahneman and Tversky (1979, 1984) show that this is indeed not the case. They show that humans are not rational with stable behavioural responses to stimuli. Human behaviour is influenced by psychological, physiological and situational factors. This could be issues in personal life, lack of sleep, or suddenly demanding tasks that needs to be solved. Following the fact that humans are not static entities with known responses, but rather change over time and contexts, efforts should be made to gain insights about what might influence behaviour. Two potentially influential topics are the constructs of affect and workload. Knowledge about how affective state and workload influence operator performance could potentially aid engineers in their work to design and test new product solutions for the maritime industry. We believe that by taking these parameters into consideration, human error could be reduced by designing the system around the human, and not make the human adapt to the system.

The paper proposes and demonstrates an experimental setup to investigate differences in affective state and workload between two ecologically valid scenarios within the domain of large ship navigation. The goal of this paper is to show that there are measurable differences in affective state and workload between the two scenarios. This may influence new ship bridge designs. These tasks have been developed in cooperation with ship simulator instructors with extensive experience as ship navigators. The experiment is run in a consumer ship simulator software (N=31) where participants from a student population are tasked to steer a ship in the following scenarios: cruising on open sea and navigating a busy harbour. Data was collected through video, self-assessment surveys and physiology sensors. The paper is part of a larger study investigating the relationship between physiological data, affect, and workload. The foundation, description and analysis of the physiological data is not within the scope or aim of this paper, and will be discussed elsewhere.

The results from the self-assessments show highly significant differences in terms of both affect and workload for the two scenarios. Consequently, one will have to consider distinctly varying affects and workloads from users in varying contexts, which, if translated into GUI and UI design suggest new design paradigms such as dynamically adaptive interfaces.

2. Theoretical foundation

2.1. Affect

Psychology presents emotion or affect as a set of variables that may moderate behaviour (Balters and Steinert, 2017). There are two main schools of thought when describing affect. The first describe emotions as a set of discrete categories (Tomkins, 1962; Ekman and Friesen, 1971; Ekman, 1992). The second describe emotions as a combination of multiple dimensions (Thayer, 1967; Russell, 1980; Watson and Tellegen, 1985; Russell and Barrett, 1999). In this paper, we adopt the description of emotions or affect of Russell (1980), the Circumplex Model of Affect, later named the Affect Grid (Russell et al., 1989). Affect is described as a construct made up of the combination of the two dimensions, arousal-sleepiness and pleasure-displeasure, see Figure 1.

Several researchers have considered how stress might influence human performance (Westman and Eden, 1996; Healey and Picard, 2005; Balters and Steinert, 2017). Russell et al. (1989) describe the construct of stress as the combination of arousal and displeasure. This is also referred to as distress as opposed to eustress which is the combination of arousal and pleasure (Healey and Picard, 2005; Balters and Steinert, 2014, 2017). Baddeley (1972) shows that increased arousal seems to narrow attention, which in term increases performance on the task that is deemed most important, but decrease performance on all other tasks.

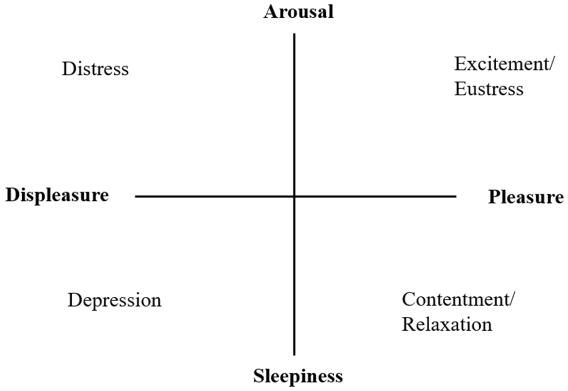


Figure 1. The Affect grid (adapted from Russell, 1980; and Russell et al., 1989)

2.1.1. Subjective measurements of affect

Assessing the subjective experience of affect is commonly done through self-report surveys. Affect can be evaluated through survey questions asking participants to evaluate levels of pleasantness and arousal (Russell, 1980), or through the single-item Affect grid (Russell et al., 1989). Positive and negative affect can be evaluated through the Positive and Negative Affect Schedule (PANAS) scales (Watson et al., 1988; Thompson, 2007). The Activation-Deactivation Adjective Check List (AD ACL) measures levels of activation (Thayer, 1967, 1986). Surveys provide a simple and low cost manner of gathering data of affective states. When using surveys in an experiment, they either interrupt participants, or must be used after tasks are finished. This might influence results, either because of the effect of an interruption, or that participants must recall how they felt during a task. Due to the subjective nature of surveys, there might be issues of self-filtering and different interpretations of questions.

2.1.2. Behavioural measurements of affect

Behavioural measurements of affect are typically concerned with measuring components of facial expression (Ekman and Friesen, 1978; Gottman and Krokoff, 1989), pitch of voice (Russell et al., 2003), and body posture and gesturing (Coulson, 2004; Wulvik et al., 2016). Advantages with behavioural measurements of affects is a very fine grained analysis of behaviour by trained experts, partially avoiding self-filtering of results, such as might be the case when answering surveys. Drawbacks are that these analyses are very labour- and time intensive, and that there might be issues of inter-coder reliability.

2.1.3. Physiological measurements of affect

The autonomic nervous system (ANS) is in charge of modulating peripheral functions of the body (Öhman et al., 2000; Mauss and Robinson, 2009). The ANS consists of the sympathetic and parasympathetic system. The sympathetic system is dominant during periods of activation, or "fight or flight", while the parasympathetic system is dominant during resting periods of the body. Changes in affective state are linked to physiological responses through the ANS. These responses can be through heart rate, heart rate variability, breathing rate, pupil dilation, muscle tension, galvanic skin response, body temperature, blood pressure, and brain activity to mention some. Healey and Picard (2005) showed a relation between levels of stress and metrics derived from galvanic skin response and heart rate variability. Baltaci and Gokcay (2016) differentiates affective states from relaxation to stress through pupil dilation and facial temperature. For a more comprehensive overview we refer to Balters and Steinert (2017), Mauss and Robinson (2009) and Levenson (2014). Physiology sensors have the advantage of providing continuous data without interrupting the person being measured, as opposed to subjective measurements through surveys. One limitation is that human physiology is very complex, and it is difficult to control all influencing factors. Another challenge with physiology data is interpreting results. How does e.g. a change in measured voltage between two sensors placed on the chest translate into affect? We recommend reading Balters and Steinert (2017) for a more complete overview.

2.2. Workload

Workload or cognitive load refers to the mental effort imposed on working memory by a particular task. (Sweller, 1988; Paas and Van Merriënboer, 1994; Paas et al., 2003) Cognition is related to our perception, in that perceptual activity, such as thinking, deciding, calculation, remembering, looking, searching increases the perceptual load, thereby the workload (Hart and Staveland, 1988). As the working memory is limited, it can be overloaded by increasing the requirements for perceptual activity. Wierwille and Eggemeier (1993) provide an overview of methods to measure workload. These can be divided into Subjective, performance-based and physiological.

2.2.1. Subjective measurements of workload

The NASA Task Load Index (NASA-TLX), a multi-dimensional scale designed for obtaining workload estimates (Hart and Staveland, 1988; Hart, 2006). NASA-TLX consists of rating six sub-scales, mental demand, physical demand, temporal demand, performance, effort, and frustration, from low to high. Participants filling out the survey are also asked to pairwise compare the six dimensions in terms of how important they are for the performed task. An estimate of total workload is then calculated from the weighted average. Another multi-dimensional scale of subjective workload is the Subjective Workload Assessment Technique (SWAT) (Reid and Nygren, 1988). It uses three levels (low, medium, high) along three dimensions, time load, mental effort load, and psychological stress load, to assess workload.

Overall Workload (Vidulich and Tsang, 1987) is a single scale measurement of subjective workload, ranging from very low to very high. Vidulich and Tsang (1987) show that the single-dimension scale of Overall Workload has higher sensitivity than the multi-dimensional scale of NASA TLX. Hill et al. (1992) showed that both the single-dimension scale of Overall Workload and NASA TLX was superior to SWAT in terms of sensitivity.

2.2.2. Performance based measurements of workload

Performance is expected to decrease with increases in workload through reduction in speed and accuracy (Wierwille and Eggemeier, 1993). Two strategies of evaluating workload through performance are common, primary and secondary task performance. Primary task performance, e.g. steering a ship might be insensitive to variations in workload, due to the operator recruiting extra resources to maintain performance (Hart and Wickens, 1990). Secondary task performance can both be assessed through external tasks and embedded tasks. External tasks are not part of the system being tested, e.g. calculating arbitrary arithmetic, while embedded tasks have a logical connection to the primary task, e.g. communicating via radio on a ship.

2.2.3. Physiological measurements of workload

Changes in physiological states has been shown to correspond with changes in workload (Galy et al., 2012). Common physiological measurements to evaluate workload are heart rate variability (HRV) (McDuff et al., 2014), electroencephalography (EEG) (Wilson and Russell, 2003), pupillary response (Iqbal and Bailey, 2005), and galvanic skin response (GSR) (Nourbakhsh et al., 2012).

3. Ship navigation experiment

An experiment was created to investigate concepts of affect and workload in two different ecologically valid scenarios in the context of large ship navigation. One task concerned steering a large ship on open water. The other task concerned steering a large ship through a busy harbour. These tasks can be described as low and high activity respectively. The aim of the experiment was to identify potential differences in affective state and workload in the different scenarios. The implication of different affective states and levels of workload for the various scenarios is that users could have changing capabilities, and that this should be addressed through the design of systems in the future. For the experiment, we formulate the following research question:

Is there a measurable difference in affective state and workload between low and high activity scenarios in the context of large ship navigation?

3.1. Scenarios

Two ecologically valid scenarios were created in the commercial ship simulator software Ship Simulator Extremes (*Ship Simulator Extremes*, 2010), replicating two typical situations in large ship navigation. Ecologically validity is obtained by the nature of the primary and the secondary task and the nature of the environmental stimuli i.e. sounds. Scenarios describe common activities on board large ships in daily operation. The scenarios and stimuli were developed in cooperation with several ship navigators with long experience as professional navigators.

3.1.1. Ship navigation on open sea – low level of activity

The first scenario was designed to recreate a low-activity situation where the task was to navigate on open sea. This is typically an uneventful task with long periods of time spent monitoring systems. The environment was set to *Dover*, and ship set to *Pride of Rotterdam,* a 215-meter long car ferry. The ship was placed close to the exit of Dover harbour with the front of the ship pointing towards the English Channel. Participants were instructed to steer the ship straight ahead towards Calais, France. The task lasted for 15 minutes, but the duration was unknown to participants. The monotonous sound of a ship engine was added to create a realistic backdrop.

3.1.2. Ship navigation in a busy harbour – high level of activity

The second scenario was designed to simulate a high-activity situation where the task was to navigate a busy harbour under a time constraint with additional secondary in the form of radio communication. The environment was set to *Rotterdam*, and *Pride of Rotterdam* was again used at ship. The participants were instructed to steer through narrow channels to a designated berth for docking. Upon leaving the starting position, a ten-minute timer would start and be displayed in the top left corner of the screen, instructing participants to reach their destination within this time limit. At regular intervals throughout the ten minutes, participants were prompted to answer eight pre-recorded questions via radio from immigration, customs and the ship's main office. These questions were voiced by three different people unfamiliar to the participants. Answers to the questions could be found in two lists provided to the participants, a cargo manifest and a crew list. These were consciously designed to be hard to read, with small letters and lots of superfluous information. Questions were repeated after 90 seconds if no answer had been given, or upon request of the participants. If participants reached the designated berth, a new destination was given. The task was designed in such a way that the final destination would be next to impossible to reach in the available ten minutes.

3.2. Physical environment

The aim of the physical environment was a controlled, static, physical space for conducting the abovementioned ecologically valid scenarios in the context of large ship navigation. A honeycomb cardboard cubicle was built (similar to the one made by Leikanger et al. (2016), equipped with a 27" computer screen mimicking the window view. A keyboard had the numerical pad marked with stickers indicating what ship functionality they controlled, e.g. rudder, thruster, etc. Today, a ship bridge control interface consists of button arrays resembling a keyboard. Additionally, much of monitoring tasks are conducted using information conveyed on a computer screen. Headphones eliminated external noise, ensuring exposure to the sound introduced by the experimenters only, i.e. ship engine noise, radio chatter and the task-specific questions. Effects from changes in external light was controlled by obscuring ambient light and illuminating the cubicle artificially with an LED strip and normal ceiling lights. Additional equipment included a mouse for answering the surveys, two web cameras for recording and monitoring the participant, a Bluetooth antenna hidden close to the devices, lists with information regarding the questions in the second scenario and marking tape indicating the area for placing the left hand. Figure 2 shows the experiment environment.

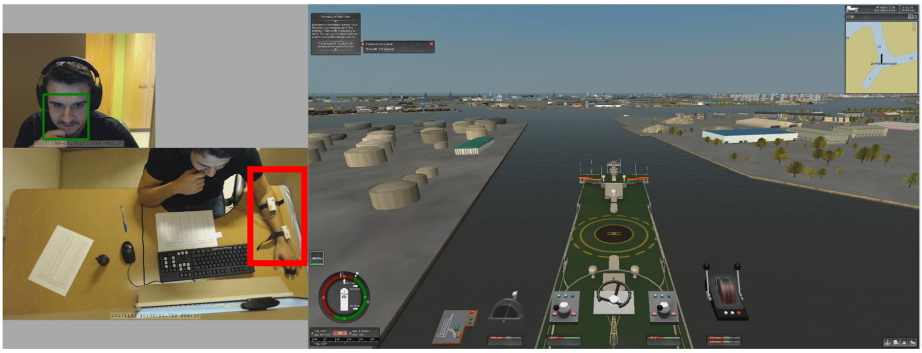


Figure 2. Experiment environment, both physical and virtual. ECG (top) and GSR (bottom) sensors highlighted in red rectangle

3.3. Collecting data from participants

A combination of self-report surveys and physiology sensors were used for data collection in the experiment. In addition, video was recorded to allow in-depth analysis of collected data.

3.3.1. Self-report surveys

To evaluate subjectively experienced affect, participants were asked to evaluate their state of arousal, awakeness, alertness, pleasantness, and stress on scales from 0 to 10. Arousal and pleasantness was taken directly from the Circumplex Model of Affect. Awakeness and alertness were added after pilot studies uncovered that participants had trouble understanding the meaning of arousal to triangulate their meaning. A question of stress was included in the surveys to capture the participants' notion of stress directly, and not only as a combination of arousal and pleasantness.

For self-assessment of workload, the single-dimension Overall Workload scale (Vidulich and Tsang, 1987) and NASA TLX (Hart and Staveland, 1988) was used. Overall Workload was evaluated on a scale from 0 to 10, and the six dimensions of the NASA TLX survey was evaluated on scales from 1 to 7, as well as 15 pairwise comparisons. All survey answers were collected through Google Forms.

3.3.2. Physiology sensors

Two types of physiology data were collected in this experiment, electrocardiography (ECG) and galvanic skin response (GSR). Electrocardiography measures electric potentials over the heart through sensors placed on the skin. The Shimmer3 ECG unit (Shimmersense, 2017a) was used in this experiment, with a sampling rate of 512 Hz. Five sensors were placed on the skin of participants per the instructions provided by Shimmer, with the V_x lead placed on position six. Data collected through ECG is measured in millivolts [mV], and can be translated into variables such as heart rate and heart rate variability. Galvanic skin response (GSR) is a measurement of conductance over the skin. The Shimmer3 GSR+ Unit (Shimmersense, 2017b) was used to measure skin conductivity. Two sensors were connected to the underside of the medial phalanx on the index and middle finger of the left hand. Sampling rate was set to 128 Hz.

3.3.3. Organising stimuli and synchronizing data

iMotions 6.4 (*iMotions*, 2017), a software platform for biometric research was as framework for presenting stimuli and synchronizing data. The sequence of instructions, surveys and simulator tasks were pre-defined in iMotions. Physiology data and video were given a common timestamp from iMotions, syncing data for future analysis.

3.4. Stating the hypotheses

The analysis of results in this paper concerns the change of self-reported affective state and workload. We operationalise the research question stated above into testable hypotheses.

Is there a measurable difference in affective state and workload between high and low activity scenarios in the context of large ship navigation?

3.4.1. Affect hypotheses

Affect is measured by asking participants to evaluate their level of arousal, awakeness, alertness, pleasantness, and stress. This leads to the following five hypotheses:

- **Affect H1:** *There is a significant change in self-reported arousal between low and high activity tasks.*
- **Affect H2:** *There is a significant change in self-reported awakeness between low and high activity tasks.*
- **Affect H3:** *There is a significant change in self-reported alertness between low and high activity tasks.*
- **Affect H4:** *There is a significant change in self-reported pleasantness between low and high activity tasks.*

- **Affect H5:** *There is a significant change in self-reported stress between low and high activity tasks.*

3.4.2. Workload hypotheses

Workload has been evaluated by participants assessing their overall workload on a single scale and through filling out the NASA TLX survey. This leads to the two following hypotheses:

- **Workload H1:** *There is a significant change in overall workload between low and high activity tasks.*
- **Workload H2:** *There is a significant change in TLX workload between low and high activity tasks.*

3.5. Running the experiment

This section aims to display how the experiment was run, and give a detailed description of the data foundation.

3.5.1. Participants

Participants in this experiment came from an engineering background (N=31). Age ranged from 19 to 33 years (24.0 ± 2.74). Out of 31 participants, 18 were male and 13 female. In addition, there were eleven participants were excluded from the analysis due to technical errors and failure to follow instructions. In the invitation to the experiment, participants were asked to participate in a study concerning "Ship Manoeuvring Behaviour". They were asked to wear a loose top for convenient connection of physiology sensors.

3.5.2. Experimenters

Two researchers conducted the experiment. The first experimenter would greet, brief, and attach sensors to participants. All interactions were scripted in advance to ensure that every participant was exposed to the same stimuli. The experimenter read all instructions from a manuscript, wore similar clothing (black jeans, light coloured dress shirt, hair pulled back in pony-tail, and no make-up). The second experimenter would sit behind a wall controlling the stimuli. After the experiment finished, the first experimenter debriefed the participant and removed sensors.

3.5.3. Protocol

Participants were greeted, introduced to the experiment, and informed about what kind of data that would be recorded. A consent form was signed by the participant, agreeing to have video, physiology data (electrocardiography and galvanic skin response) and survey answers recorded. Physiology sensors were attached by the experimenter. Participants were then instructed to sit down in front of a computer screen, and place their left hand on the table, making sure their arm was resting comfortably. They were told that instructions may be given both on-screen and through audio. In the case of audio instructions, answers should be given through a radio handset. Usage of the radio handset was explained and demonstrated. Participants were instructed to keep their left hand still throughout the experiment to ensure the quality of GSR data recorded. After instructions were given, the experimenter left the room and joined the second experimenter behind a wall. The computer screen showed a black image with white crosshairs in the middle when participants entered the room. When participants were ready to start, the second experimenter manually started the sequence of stimuli in iMotions. Participants were first presented with neutral stimuli. Participants then filled out a survey on their affective state to serve as a reference baseline. Information about the experiment was given in writing with a white background. They were informed that they would be controlling the ship *Pride of Rotterdam* and execute various tasks. Participants were informed that there were two printed lists, a crew list and cargo manifest, to their right side. These lists should not be used before instructed to do so. Following the initial brief, participants were shown a video giving instructions for how to control the ship with the keyboard. All keys to be used on the keyboard were physically labelled with a short explanatory name. After receiving instructions, participants were informed that the first task would begin and the computer screen switched to the simulator software for the low activity task. The second experimenter manually unpaused the software and gave over control to the participants. After 15 minutes from leaving the harbour in Dover the software would display a loading screen, initialising the high activity task. The second experimenter would manually change the view to the second survey, concerning affective state and workload. After completing the survey, the view was manually switched back to the simulator software, starting the high activity task. The ten-minute timer would start after the ship had started moving, and pre-recorded radio questions were manually played at pre-defined intervals by the second experimenter. After the ten minutes passed, a screen telling participants that they failed their mission (no participants were able to complete the mission, as expected). The third survey was presented to participants, asking about affective state and workload. When completed, participants were prompted to answer background questions, e.g. age, gender, occupation, in a fourth survey. After completing the final survey, they were informed that the experiment was finished, and were thanked for their participation. The first experimenter would walk back to debriefing the participants, thank them for their contribution, remove the sensors and ask them not to share content or details about the experiment to others. All equipment was cleaned and printed lists were replaced after each participant.

4. Survey results

Survey results from the 31 participants completing the experiment was analysed in SPSS Statistics (IBM, 2016) to investigate potential statistical differences in affective states and workload. A total of seven variables were tested for statistically significant change in values on an 11-point scale, between low and high activity tasks. Statistical tests were selected based on the properties of recorded data, i.e. outliers, normal-, and symmetric distributions. Difference in values between the two scenarios is the foundation for the tests. Paired samples t-test was used for normally distributed date without significant outliers. For data violating the assumptions of normal distribution or no significant outliers, the Wilcoxon signed-rank test was used if the data was symmetrically distributed. For non-symmetric distributions, the Sign test was used. The Wilcoxon signed-rank test and the Sign test evaluates median differences as opposed to mean differences in the paired samples t-test. Outliers are defined by SPSS statistics as values more than 1.5 box-lengths from the edge of a box in a box plot. Shapiro-Wilk's test for normal distribution was used to assess whether values were normally distributed, where significance values larger than 0.05 indicates normally distributer variables. Symmetricity of distribution was evaluated visually using histograms. Data are mean ± standard deviation, unless otherwise stated. Table 1 contains descriptive statistics, and Table 2 contains metrics associated with assumptions that decide which statistical tests to use along with the corresponding results. As shown in Table 2, all seven variables are significantly different in the two scenarios.

Table 1. Descriptive statistics

Variable	S1	S1 Median	S2	S2 Median	Diff.	Diff. Median
Arousal	5.61 ± 2.38	6	6.68 ± 2.70	8	1.06 ± 1.75	1
Awakeness	6.48 ± 2.05	7	7.61 ± 2.04	8	1.13 ± 1.09	1
Alertness	6.26 ± 1.95	7	7.42 ± 1.86	8	1.16 ± 1.61	1
Pleasantness	6.35 ± 1.62	7	4.68 ± 1.80	4	-1.68 ± 1.54	-2
Stress	3.23 ± 1.94	3	6.39 ± 2.14	7	3.16 ± 1.88	3
Overall Workload	2.03 ± 1.70	2	8.03 ± 1.78	8	6.00 ± 2.93	7
TLX Workload	2.86 ± 1.23	2.73	5.60 ± 0.74	5.80	2.73 ± 1.37	2.86

Table 2. Testing for statistical difference change in variables between low and high activity scenarios

Variable	Outliers	Shapiro-Wilk's test	Symmetric	95% CI Lower	95% CI Upper	Sig. (2-tailed)
Arousal	Yes	0.063	Yes	0.42	1.71	< 0.01 [a]
Awakeness	No	< 0.01	No	0.73	1.53	< 0.01[c]
Alertness	Yes	0.014	No	0.57	1.75	< 0.01[c]
Pleasantness	No	0.214	Yes	-2.24	-1.11	< 0.01 [a]
Stress	No	0.112	Yes	2.47	3.85	< 0.01 [a]
Overall Workload	Yes	< 0.01	No	4.93	7.07	< 0.01[c]
TLX Workload	No	0.48	Yes	2.23	3.24	< 0.01[a]

a: Paired samples t-test, b: Wilcoxon signed-rank test, c: Sign test

5. Discussion: Interpreting the results and the way forward

Results in the above tests show that there are significant changes in all seven variables. TLX Workload is a weighted sum of the six dimensions: Mental demand, physical demand, temporal demand, performance, effort, and frustration level. With the exception of performance (p=0.69), all dimensions had significantly changes. This might be due to difficulties related to comparing performance in two very different and unfamiliar scenarios. Changes are quite small for the variables of arousal, awakeness and alertness, with mean changes of around one on an eleven-point scale. Variables of pleasantness, stress, overall workload and TLX workload have a larger change (see Table 1). We are not sure whether the differences in magnitude of change is due to real differences, or due to how participants interpret the survey questions. One can speculate e.g. that participants did not have a clear understanding of the concepts of arousal, awakeness, and alertness, or at least had difficulties evaluating them. Pleasantness, stress and workload might be more intuitively understandable for the participants, which might be the reason for the difference in magnitude of change. This finding is interesting, as it contrasts with the fact that Russell (1980) defines stress as a combination of arousal and pleasantness. We know from literature that there is supposed to be a link between physiological data and arousal, e.g. heart rate variability and skin conductance. Further work will include analysing physiological data and comparing results with subjective assessment of affective state and workload, investigating the relationship between the two. One limitation of our study is that participants were sampled from a student population. Results might have been influenced by this fact, due to being unfamiliar with the situation of ship piloting. We believe that the findings that show a difference in affective state and workload between the two scenarios are valid for the context of ship navigation, although the effect size should be verified through testing with professional navigators in more realistic contexts, i.e. professional ship simulators or real ships.

The results nevertheless show that there is a clear difference in affective state and workload in the two scenarios tested in this experiment. Consequently, one should consider distinctly varying affects and workloads from users in varying contexts. This, if translated into product development, GUI, and UI design suggest new design paradigms such as dynamically adaptive interfaces.

References

Baddeley, A.D. (1972), "Selective attention and performance in dangerous environments", *British Journal of Psychology*, Vol. 63 No. 4, pp. 537–546. https://doi.org/10.1111/j.2044-8295.1972.tb01304.x

Baltaci, S. and Gokcay, D. (2016), "Stress Detection in Human–Computer Interaction: Fusion of Pupil Dilation and Facial Temperature Features", *International Journal of Human–Computer Interaction*, Vol. 32 No. 12, pp. 956–966. https://doi.org/10.1080/10447318.2016.1220069

Balters, S. and Steinert, M. (2014), "Decision-making in engineering-a call for affective engineering dimensions in applied engineering design and design sciences", *Proceedings of the 2014 International Conference On*

Innovative Design and Manufacturing (ICIDM 2014), August 13-15, 2014, Montreal, Canada, IEEE, pp. 11–15. https://doi.org/10.1109/IDAM.2014.6912663

Balters, S. and 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*, Vol. 28 No. 7, pp. 1585 - 1607. https://doi.org/10.1007/s10845-015-1145-2

Coulson, M. (2004), "Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence", *Journal of Nonverbal Behavior*, Vol. 28 No. 2, pp. 117–139. https://doi.org/10.1023/B:JONB.0000023655.25550.be

Ekman, P. (1992), "An argument for basic emotions", *Cognition and Emotion*, Vol. 6 No. 3-4, pp. 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P. and Friesen, W.V. (1971), "Constants across cultures in the face and emotion", *Journal of Personality and Social Psychology*, Vol. 17 No. 2, pp. 124-129. https://doi.org/10.1037/h0030377

Ekman, P. and Friesen, W.V. (1978), *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto, California.

Galy, E., Cariou, M. and Mélan, C. (2012), "What is the relationship between mental workload factors and cognitive load types?", *International Journal of Psychophysiology*, Vol. 83 No. 3, pp. 269–275. https://doi.org/10.1016/j.ijpsycho.2011.09.023

Gottman, J.M. and Krokoff, L.J. (1989), "Marital interaction and satisfaction: a longitudinal view", *Journal of Consulting and Clinical Psychology*, Vol. 57 No. 1, pp. 47-52.

Hart, S.G. (2006), "Nasa-Task Load Index (NASA-TLX); 20 Years Later", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* Vol. 50 No. 8, pp. 904–908. https://doi.org/10.1177/154193120605000909

Hart, S.G. and Staveland, L.E. (1988), "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research", *Advances in Psychology*, Vol. 52, pp. 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hart, S.G. and Wickens, C.D. (1990), "Workload Assessment and Prediction", In: Booher, H.R. (Ed.), *Manprint,* Springer, Dordrecht, pp. 257–296. https://doi.org/10.1007/978-94-009-0437-8_9

Healey, J.A. and Picard, R.W. (2005), "Detecting stress during real-world driving tasks using physiological sensors", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 6 No. 2, pp. 156–166. https://doi.org/10.1109/TITS.2005.848368

Hetherington, C., Flin, R. and Mearns, K. (2006), "Safety in shipping: The human element", *Journal of Safety Research*, Vol. 37 No. 4, pp. 401–411. https://doi.org/10.1016/j.jsr.2006.04.007

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklade, A.L. and Christ, R.E. (1992), "Comparison of Four Subjective Workload Rating Scales", *Human Factors*, Vol. 34 No. 4, pp. 429–439. https://doi.org/10.1177/001872089203400405

IBM (2016), *IBM SPSS Statistics for Mac.* [online] IBM, New York, USA. Available at: https://www.ibm.com/products/spss-statistics

iMotions (2017), *iMotions biometric research platform*. [online] iMotions. Available at: https://imotions.com

Iqbal, S.T. and Bailey, B.P. (2005), "Investigating the Effectiveness of Mental Workload As a Predictor of Opportune Moments for Interruption", *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05), Portland, USA, April 2-7, 2005*, ACM, New York, USA, pp. 1489–1492. https://doi.org/10.1145/1056808.1056948

Kahneman, D. and Tversky, A. (1979), "Prospect Theory: An Analysis of Decision under Risk", *Econometrica,* Vol. 47 No. 2, pp. 263–292.

Kahneman, D. and Tversky, A. (1984), "Choices, values, and frames", *American Psychologist*, Vol. 39 No. 4, pp. 341-350. https://doi.org/10.1037/0003-066X.39.4.341

Leikanger, K.K., Balters, S. and Steinert, M. (2016), "Introducing the Wayfaring Approach for the Development of Human Experiments in Interaction Design and Engineering Design Science", *Proceedings of the DESIGN 2016 / 14th International Design Conference, Dubrovnik, Croatia, May 16-19, 2016*, The Design Society, Glasgow, pp. 1751–1762.

Levenson, R.W. (2014), "The Autonomic Nervous System and Emotion", *Emotional Review*, Vol. 6 No. 2, pp. 100–112. https://doi.org/10.1177/1754073913512003

Mauss, I.B. and Robinson, M.D. (2009), "Measures of emotion: A review", *Cognition and Emotion,* Vol. 23 No. 2, pp. 209–237. https://doi.org/10.1080/02699930802204677

McDuff, D., Gontarek, S. and Picard, R. (2014), "Remote measurement of cognitive stress via heart rate variability", *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2014), Chicago, Illinois, August 26-30, 2014*, IEEE, pp. 2957–2960. https://doi.org/10.1109/EMBC.2014.6944243

Nilsson, R., Gärling, T. and Lützhöft, M. (2009), "An experimental simulation study of advanced decision support system for ship navigation", *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 12 No. 3, pp. 188–197. https://doi.org/10.1016/j.trf.2008.12.005

Norros, L. (2004), Acting under uncertainty: The core-task analysis in ecological study of work, VTT Technical Research Centre of Finland, Espoo, Finland.

Nourbakhsh, N., Wang, Y., Chen, F. and Calvo, R.A. (2012), "Using Galvanic Skin Response for Cognitive Load Measurement in Arithmetic and Reading Tasks", *Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI '12), Melbourne, Australia, November 26-30, 2012,* ACM, New York, USA, pp. 420–423. https://doi.org/10.1145/2414536.2414602

Öhman, A., Hamm, A. and Hugdahl, K. (2000), "Cognition and the autonomic nervous system: orienting, anticipation, and conditioning", In: Cacioppo, J.T., Tassinary, L.G. and Berntson, G.G. (Eds.), *Handbook of psychophysiology*, 2nd ed., Cambridge University Press, New York, pp. 533–575.

Paas, F., Tuovinen, J.E., Tabbers, H. and Gerven, P.W.M.V. (2003), "Cognitive Load Measurement as a Means to Advance Cognitive Load Theory", *Educational Psychologist*, Vol. 38 No. 1, pp. 63–71. https://doi.org/10.1207/S15326985EP3801_8

Paas, F.G. and Van Merriënboer, J.G. (1994), "Instructional control of cognitive load in the training of complex cognitive tasks", *Educational Psychology Review,* Vol. 6 No. 4, pp. 351–371. https://doi.org/10.1007/BF02213420

Reid, G.B. and Nygren, T.E. (1988), "The subjective workload assessment technique: A scaling procedure for measuring mental workload", *Advances in Psychology*, Vol. 52, pp. 185–218. https://doi.org/10.1016/S0166-4115(08)62387-0

Rothblum, A.M. (2000), "Human error and marine safety", *National Safety Council Congress and Expo, Orlando, Florida, October 16-18, 2000.*

Russell, J.A. (1980), "A circumplex model of affect", *Journal of Personality and Social Psychology*, Vol. 39 No. 6, pp. 1161–1178. https://doi.org/10.1037/h0077714

Russell, J.A. and Barrett, L.F. (1999), "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant", *Journal of Personality and Social Psychology*, Vol. 76 No. 5, pp. 805-819. https://doi.org/10.1037/0022-3514.76.5.805

Russell, J.A., Bachorowski, J.-A. and Fernández-Dols, J.-M. (2003), "Facial and Vocal Expressions of Emotion", *Annual Review of Psychology*, Vol. 54, pp. 329–349. https://doi.org/10.1146/annurev.psych.54.101601.145102

Russell, J.A., Weiss, A. and Mendelsohn, G.A. (1989), "Affect grid: A single-item scale of pleasure and arousal", *Journal of Personality and Social Psychology,* Vol. 57, pp. 493–502. https://doi.org/10.1037/0022-3514.57.3.493

Sanders, M.S. and McCormick, E.J. (1987), *Human factors in engineering and design*, McGraw-Hill.

Shimmersense (2017a), Shimmer3 ECG/EMG Unit, Available at: http://www.shimmersensing.com/products/shimmer3-ecg-sensor

Shimmersense (2017b), Shimmer3 GSR+ Unit, Available at: http://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor

Ship Simulator Extremes (2010), *Ship Simulator Extremes.* [online] ShipSim.com. Available at: https://www.shipsim.com/products/shipsimulatorextremes

Sweller, J. (1988), "Cognitive load during problem solving: Effects on learning", *Cognitive Science*, Vol. 12 No. 2, pp. 257–285. https://doi.org/10.1207/s15516709cog1202_4

Thayer, R.E. (1967), "Measurement of Activation through Self-Report", *Psychological Reports*, Vol. 20 No. 2, pp. 663–678. https://doi.org/10.2466/pr0.1967.20.2.663

Thayer, R.E. (1986), "Activation-Deactivation Adjective Check List: Current Overview and Structural Analysis", *Psychological Reports,* Vol. 58 No. 2, pp. 607–614. https://doi.org/10.2466/pr0.1986.58.2.607

Thompson, E.R. (2007), "Development and Validation of an Internationally Reliable Short-Form of the Positive and Negative Affect Schedule (PANAS)", *Journal of Cross-Cultural Psychology*, Vol. 38 No. 2, pp. 227–242. https://doi.org/10.1177/0022022106297301

Tomkins, S. (1962), *Affect imagery consciousness: Volume 1: The positive affects,* Springer Publishing Company.

Tzannatos, E. (2010), "Human Element and Accidents in Greek Shipping", *The Journal of Navigation*, Vol. 63, pp. 119–127. https://doi.org/10.1017/S0373463309990312

Vidulich, M.A. and Tsang, P.S. (1987), "Absolute Magnitude Estimation and Relative Judgement Approaches to Subjective Workload Assessment", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 31 No. 9, pp. 1057–1061. https://doi.org/10.1177/154193128703100930

Watson, D. and Tellegen, A. (1985), "Toward a consensual structure of mood", *Psychological Bulletin*, Vol. 98 No. 2, pp. 219-235. https://doi.org/10.1037/0033-2909.98.2.219

Watson, D., Clark, L.A. and Tellegen, A. (1988), "Development and validation of brief measures of positive and negative affect: The PANAS scales", *Journal of Personality and Social Psychology*, Vol. 54 No. 6, pp. 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Westman, M. and Eden, D. (1996), "The inverted-U relationship between stress and performance: A field study", *Work Stress*, Vol. 10 No. 2, pp. 165–173. https://doi.org/10.1080/02678379608256795

Wierwille, W.W. and Eggemeier, F.T. (1993), "Recommendations for Mental Workload Measurement in a Test and Evaluation Environment", *Human Factors*, Vol. 35 No. 2, pp. 263–281. https://doi.org/10.1177/001872089303500205

Wilson, G.F. and Russell, C.A. (2003), "Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks", *Human Factors*, Vol. 45 No. 4, pp. 635–644. https://doi.org/10.1518/hfes.45.4.635.27088

Woodson, W.E. and Conover, D.W. (1970), *Human engineering guide for equipment designers*, University of California Press, California.

Wulvik, A., Erichsen, J. and Steinert, M. (2016), "Capturing Body Language in Engineering Design – Tools and Technologies", *Proceedings of the NordDesign 2016, Trondheim, Norway, August 10-12, 2016*, The Design Society, Bristol, pp. 165-174.

Andreas Simskar Wulvik, PhD Student
Norwegian University of Science and Technology - NTNU, Department of Mechanical and Industrial Engineering
Richard Birkelands Veg 2B, 7034 Trondheim, Norway
Email: andreas.wulvik@ntnu.no

Appendix C3: Academic contribution 3

Rørvik, S. B., Auflem, M., Dybvik, H., & Steinert, M. (2021). Perception by Palpation: Development and Testing of a Haptic Ferrogranular Jamming Surface. Frontiers in Robotics and AI, 8, 311. https://doi.org/10.3389/frobt.2021.745234

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 28 September 2021
doi: 10.3389/frobt.2021.745234

Perception by Palpation: Development and Testing of a Haptic Ferrogranular Jamming Surface

Sigurd Bjarne Rørvik, Marius Auflem, Henrikke Dybvik and Martin Steinert*

TrollLABS, Department of Mechanical and Industrial Engineering, Faculty of Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Tactile hands-only training is particularly important for medical palpation. Generally, equipment for palpation training is expensive, static, or provides too few study cases to practice on. We have therefore developed a novel haptic surface concept for palpation training, using ferrogranular jamming. The concept's design consists of a tactile field spanning 260 x 160 mm, and uses ferromagnetic granules to alter shape, position, and hardness of palpable irregularities. Granules are enclosed in a compliant vacuum-sealed chamber connected to a pneumatic system. A variety of geometric shapes (output) can be obtained by manipulating and arranging granules with permanent magnets. The tactile hardness of the palpable output can be controlled by adjusting the chamber's vacuum level. A psychophysical experiment (N = 28) investigated how people interact with the palpable surface and evaluated the proposed concept. Untrained participants characterized irregularities with different position, form, and hardness through palpation, and their performance was evaluated. A baseline (no irregularity) was compared to three irregularity conditions: two circular shapes with different hardness (Hard Lump and Soft Lump), and an Annulus shape. 100% of participants correctly identified an irregularity in the three irregularity conditions, whereas 78.6% correctly identified baseline. Overall agreement between participants was high (κ= 0.723). The Intersection over Union (IoU) for participants sketched outline over the actual shape was IoU *Mdn* = 79.3% for Soft Lump, IoU *Mdn* = 68.8% for Annulus, and IoU *Mdn* = 76.7% for Hard Lump. The distance from actual to drawn center was *Mdn* = 6.4 mm (Soft Lump), *Mdn* = 5.3 mm (Annulus), and Mdn = 7.4 mm (Hard Lump), which are small distances compared to the size of the field. The participants subjectively evaluated Soft Lump to be significantly softer than Hard Lump and Annulus. Moreover, 71% of participants thought they improved their palpation skills throughout the experiment. Together, these results show that the concept can render irregularities with different position, form, and hardness, and that users are able to locate and characterize these through palpation. Participants experienced an improvement in palpation skills throughout the experiment, which indicates the concepts feasibility as a palpation training device.

Keywords: haptic interface, tactile surface, simulation, palpation, granular jamming, tactile perception, ferromagnetic granules

OPEN ACCESS

Edited by:
Carlos Rossa,
Carleton University, Canada

Reviewed by:
Nicolas Herzig,
University of Sussex, United Kingdom
Jay Carriere,
University of Calgary, Canada
Min Li,
Xi'an Jiaotong University, China

****Correspondence:***
Marius Auflem
mariuauf@stud.ntnu.no

Specialty section:
This article was submitted to Haptics, a section of the journal Frontiers in Robotics and AI

Received: *21 July 2021*
Accepted: *09 September 2021*
Published: *28 September 2021*

Citation:
Rørvik SB, Auflem M, Dybvik H and Steinert M (2021) Perception by Palpation: Development and Testing of a Haptic Ferrogranular Jamming Surface. Front. Robot. AI 8:745234. doi: 10.3389/frobt.2021.745234

1 INTRODUCTION

In simulated training environments (i.e., augmented, virtual, and mixed reality), realistic rendering of tactile interactions with the physical world is challenging, yet meaningful. This is because haptic interfaces enabling such tactile interactions must complement (and reflect) the vivid audiovisual feedback provided by the simulation (Woodrum et al., 2006). This combination could yield deeper immersion and thus facilitate the transfer of tactile experiences when transitioning to real-world scenarios. Furthermore, by realistically bridging the physical and digital world, users can develop, improve, and maintain critical psychomotor skills (Lathan et al., 2002; Zhou et al., 2012; Zhao et al., 2020). Hence, haptic interfaces in simulation can enable safe, repetitive, and available training alternatives for various professions that require dexterous hands-on experience (Carruth, 2017; Lelevé et al., 2020).

In a medical context, simulation can help narrow the gap of required clinical experience and mitigate the risk of harming or providing unsatisfactory patient treatment. However, various medical procedures require not only hands-on, but hands-only training. One of these procedures is palpation, which is used to examine a patient through touch. By palpation, diagnosis is based on tactile findings such as irregularities (lumps, fluids, tenderness) and locating pain-points. Unfortunately, common equipment such as wearable tactile devices and kinesthetic devices are less suited in this use-case given their current resolution, Degrees of Freedom (DOF), and tactile limitations (Licona et al., 2020). Consequently, simulated palpation exercises are mainly performed using static case-specific models (phantoms) or mannequins (patient simulators). While these can provide safe and repetitive training conditions, their fixed number of study cases, task-specific functionalities, and limited tactile realism are collectively obstacles for current healthcare training and education.

Haptic interfaces designed for palpation training should enable users to practice locating and describing tactile irregularities, as they would when palpating a real patient. Hence, multiple tactile displays are promising in this context by utilizing technology ranging from pin arrays (Wagner et al., 2002), to shape memory alloy actuators (Taylor et al., 1998) and airborne ultrasound (Iwamoto et al., 2008). However, such solutions are generally expensive, complex in operation and non-continuously available, thus limiting their use and widespread in research and education. Moreover, as these solutions rely on using a matrix of actuators or tactile outputs, it restricts the obtainable resolution, scalability and robustness of such interfaces. Furthermore, compliance and flexibility are often compromised by using rigid mechanisms to achieve haptic feedback. Therefore, attention has been brought to using soft robotics principles for haptic applications, as these can approximate soft body animations and organic behaviors suitable to medical training, among others (Manti et al., 2016).

An interesting area of soft robotics for medical training applications is the use of granular jamming mechanisms for haptic feedback. Granular jamming enables interfaces to alter stiffness and thus simulate compliant objects with variable hardness. This technology has been explored in medical training devices as embedded tactile modules (He et al., 2021), multi-fingered palpation interfaces (Li et al., 2014), and as actuation to enable objects and surfaces to alter shape and hardness for palpation (Stanley et al., 2016; Koehler et al., 2020). While this technology looks promising, current solutions often require complex pneumatic systems, since a matrix of actuated cells or objects is needed. Thus, this could limit the tactile resolution and geometrical freedom of rendered objects. Based on this existing work on granular jamming interfaces, we have developed a simple and low-cost technology utilizing ferromagnetic granulate. Our technology enables the granules to be remotely manipulated in an unjammed state and thus create customized tactile objects. Furthermore, when jammed, the hardness of these objects can be altered by the applied vacuum, i.e., how firmly the granules are packed together in a sealed chamber. In a haptic interface prototype described in **Figure 1**, the ferrogranular jamming principle is used to render palpable irregularities between two compliant layers. The prototype was developed to examine the feasibility and usability of this technology in a tactile display application. Moreover, this technology could be used to challenge the complexity, accessibility and cost of current haptic interfaces.

This work relates to the existing literature on tactile interactions, and more precisely, users' tactile perception of hardness and geometrical shapes. Hence, studies investigating the psychophysical perception of hardness and shapes have been of interest (Tan et al., 1992; Srinivasan and LaMotte, 1995; Bergmann Tiest and Kappers, 2009; Frisoli et al., 2011). However, the use-case of palpable interfaces that requires a perceptual exploration and manipulation is a less explored area with fewer examples (Lederman and Klatzky, 1993; Genecov et al., 2014). As this encourages more research on users' interaction and performance using haptic interfaces, our conceptual prototype has been piloted in a palpation experiment. This experiment investigates whether untrained users can locate and determine the form and hardness of rendered irregularities by palpation. Information of hardness, speed (time used to find irregularity) and accuracy of form and position has been collected, together with users' subjective experience throughout the experiment.

This paper examines using soft-robotics principles to alter the characteristics of a haptic interface for medical diagnostics training. This investigation has resulted in the concept shown in **Figure 1**, which uses granular jamming and ferromagnetic granulate manipulation to achieve various palpable outputs. The concept is used to assess untrained users' ability to locate and characterize the shape and hardness of different irregularities using palpation. Considering this concept for a novel haptic interface and the context of medical palpation training, we try to answer the following research questions in this paper:

i. Can the novel ferrogranular jamming concept be used as a haptic interface for palpation exercises?
ii. How well can untrained users determine the position, form and hardness of irregularities rendered by the haptic interface using palpation techniques?

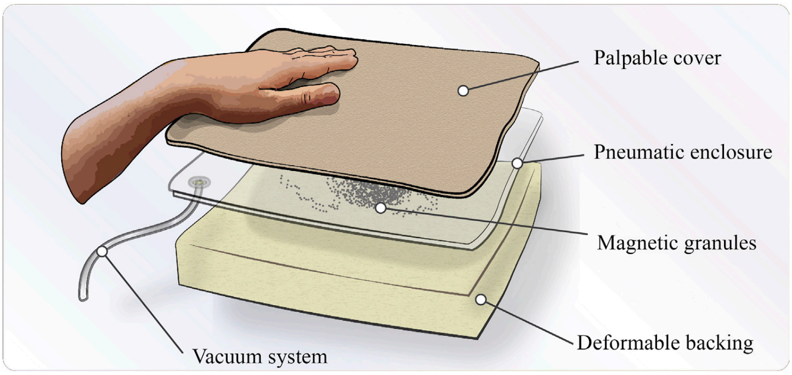


FIGURE 1 | Descriptive illustration of the haptic interface concept.

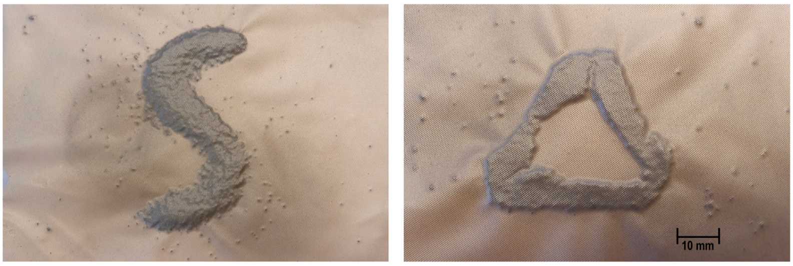


FIGURE 2 | Pictures of two arrangement possibilities.

iii. Did participants think their palpation skills improved during the experiment?

2 MATERIALS: DESIGN OF THE FERROGRANULAR JAMMING INTERFACE

This chapter starts with a short introduction to the ferrogranular jammer. Secondly, the theory of granular jamming and magnetic manipulation is presented. Lastly, the manufacturing of the magnetic granules and chamber is presented before the pneumatic setup.

The prototype was developed to examine the feasibility and usability of a ferrogranular jamming interface in a tactile display for palpation. The novelty of the proposed concept is the introduction of magnetic manipulation of granules in a jamming application. This innovation provides the opportunity to manipulate the granular media inside a compliant vacuum chamber, thus managing the position, form and hardness of the palpable outputs. Some examples are shown in **Figure 2**, where the jammed granulate shapes are visible within the translucent chamber. To act as a deformable and palpable structure the vacuum chamber is sandwiched between a deformable polyurethane (PU) foam backing (60 mm) and a flexible polyethylene (PE) fabric cover (4 mm) (as seen in **Figure 5B**).

2.1 Granular Jamming and Magnetic Manipulation

Granular jamming works by transitioning granular matter from a low-density compliant packing to a high-density rigid packing. This change is done by removing the fluid/medium surrounding the granulate, which produces an external hydrostatic pressure. From this, the granules can behave both like a fluid and a solid. When the granules are in a low-density packing, the intergranular friction is low, resulting in a fluid-like state. Vice versa, when the vacuum level increases, higher intergranular friction results in a jammed and solid-like state. In the jammed state, the granules distributes applied force through the grains so that the group of particles functions as a stiff and compliant material (Cates et al., 1998).

Particle jamming has been a big research topic for engineers and material scientists for the last few decades. The principle of reversibly transitioning the granular media from a fluid-like state to a more rigid state has been seen to be applicable to various domains, such as industrial grippers (Harada et al., 2016; D'Avella et al., 2020), minimally invasive surgery (Jiang et al., 2012) and robotic locomotion (Steltz et al., 2009). Granular jamming is a prevalent type of actuation within soft robotics applications because of two main reasons: 1) considerable stiffness variation with little volume change, and 2) possibility to adjust the stiffness variability area so it can be easily adapted to different soft robotics applications (Fitzgerald et al., 2020).

There has been research on optimizing granules for granular jamming with different aspects; size, shape and volume fraction (Jiang et al., 2012), chamber material (Jiang et al., 2012) and using soft granules (Putzu, Konstantinova, and Althoefer 2019). However, a common feature for these studies is the stasis of the granulate. To the best of our knowledge, there has been no research focusing on the movability of granules in a jamming context. For example, Follmer et al. (2012) reviewed jamming in a user-interface context, where none of the technologies utilized movement of the granules.

Using magnetic fields is an effective way to transport and position magnetic particles in a medium. The most prominent concept of ferromagnetic particles in a fluid is ferrofluid. This colloidal liquid consists of surfactant-coated magnetic particles with a size order of 10 nm suspended in a liquid medium. When the fluid is subjected to a magnetic field, it forms a shape like the magnetic field and acts more like a solid. Generally, ferromagnetic particles are induced by two types of interaction energy: the one between the particles and the magnetic field E^H, and between particles E^M (Cao et al., 2014). Using a magnetic field to manipulate magnetic particles has been used in microfluidic systems, such as magnetorheological fluid in user interfaces (Hook et al., 2009; Jansen et al., 2010) and biological analysis and catalysis (Gijs et al., 2010).

The advantages of using ferromagnetic granules include: 1) Controllability—Ferro-granulate can be arranged numerous ways by designing magnetic fields. 2) Noncontact—Magnetic particles can be remotely manipulated. 3) Precision—Ferromagnetic granules can be placed at a target region with high precision by precisely designing a magnetic field with local maximum field strength at preferred areas (Cao et al., 2014).

2.2 Manufacturing of Magnetic Granulate

Based on the previous research done on granular jamming, manufacturing of ferromagnetic granules to be used in a haptic interface were investigated. A central factor for the granulate in this research is how high interparticle friction yields higher viscosity in the un-jammed state but yields higher hardness when jammed and vice versa. Since moving the granules in the unjammed state is essential, we investigated the granule material and manufacturing methods that produce granules with lower interparticle friction in the unjammed state but still yielding sufficient hardness in the jammed state.

Ground coffee, which Putzu, Konstantinova, and Althoefer (2019) refer to as the gold standard within the field of granular jamming, was evaluated as the most viable option for our case. Ground coffee has been proven to be a successful granulate for jammers that need a large stiffness range (Brown et al., 2010; Cheng et al., 2012). The magnetic coffee ground was produced by mixing fine coffee ground and magnetic paint with a 1:1 volumetric ratio as seen in **Figure 3** (Magnetic undercoat, Lefranc and Bourgeois Déco). After the mixture dried, it was ground to a size of approximately 2 mm using a mortar. Using a crushing technique, instead of grinding, produced less size dispersion of the granulate. Granules with a 1–2.4 mm size were filtered out with a perforated filter with circular holes (see **Figures 3D,E**). It is advantageous to use homogeneous monodisperse granules to make the output more repeatable (Genecov et al., 2014).

The manipulation of the ferromagnetic granulate using a permanent magnet is presented in **Figure 4**. The same type of spikes can be observed in both ferrofluids and iron shavings when in the presence of a magnetic field.

2.3 Chamber Design

Since the concept of this technology is different from traditional granular jamming, the choice of chamber material was evaluated on having surface friction that enabled the granules to be remotely manipulated inside the sealed chamber. Further, the material needed to be flexible to jam the particles together when a vacuum was applied. Different heat-sealing plastic types were evaluated, and a corrugated polyvinylchloride (PVC) film (0.2 mm for vacuum sealing applications) was deemed the most viable due to its flexibility and least warping lines. With the corrugated pattern, we avoided self-sealing as this was a problem with other materials.

2.4 Pneumatic Setup

The pneumatic setup for the ferrogranular jamming concept is shown in **Figure 5**. The chamber is connected to the rest of the pneumatic system through a filter (**Figure 5D**). The 12 V vacuum pump (D2028B, SparkFun Electronics) delivers a vacuum level down to −0.54 bar. Next, a manometer is connected to measure the vacuum level. The vacuum pump is controlled using a speed controller. The chamber was made using an Impulse Heat Sealer (Audion Elektro Sealboy 235). A 3D-printed nozzle connects the chamber to the rest of the system, as seen in **Figure 5E**. Together with butyl vacuum sealant tape, it ensures minimal leakage at the inlet. A ball valve connects the system to atmospheric pressure when open.

3 METHOD: EXPERIMENT

A psychophysical experiment was designed to evaluate the functional abilities of the proposed concept by evaluating the user's performance in locating and characterizing rendered irregularities. The experiment encompassed a palpation task, where qualitative and quantitative data were gathered on both participant performance and prototype reliability.

3.1 Experimental Test Setup

The pneumatic system presented in 2.4 was integrated into the test cabinet shown in **Figure 6A**. A camera is fixed above the

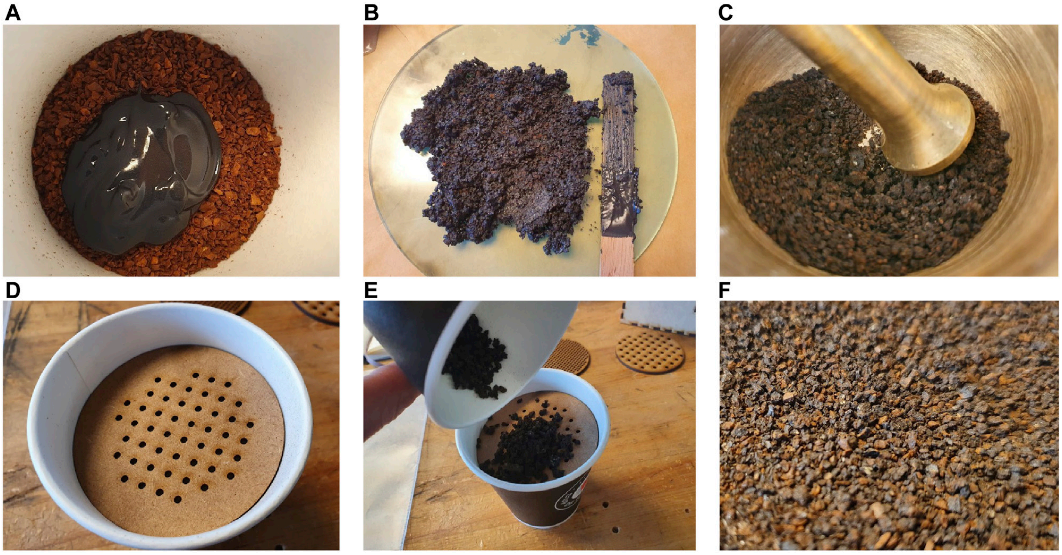


FIGURE 3 | Manufacturing of magnetic granulate **(A)** 1:1 mixing ratio of coffee ground and magnetic paint **(B)** Consistency of the mixture **(C)** Grounding using mortar **(D)** and **(E)** Filtering **(F)** Finished result.

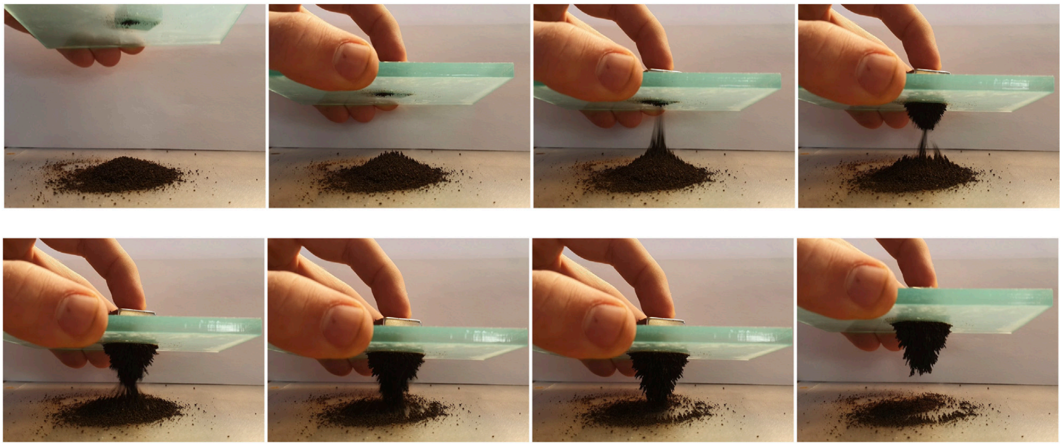


FIGURE 4 | Manipulation of the magnetic coffee ground using a permanent magnet.

haptic interface. The cabinet walls ensure no bias from visual perception during the transition between conditions and provides a consistent working environment. In addition, an overhead LED panel eases picture processing by ensuring consistent lighting. The two different geometrical shapes were created with two arrangements of permanent neodymium magnets, as seen in **Figure 6B**. These magnets were held above the vacuum chamber, arranging the granules in the desired shape, before applying the vacuum. When vacuum was applied, the magnets could be removed and the granulate remained jammed in place. To alter the shape, or remove it, the vacuum was released, before the granules were manually dispersed, rearranged, or moved out of the palpable field. The structural parts of the test rig are laser-cut MDF. The palpable field (260 × 160 mm) is seen as the pink

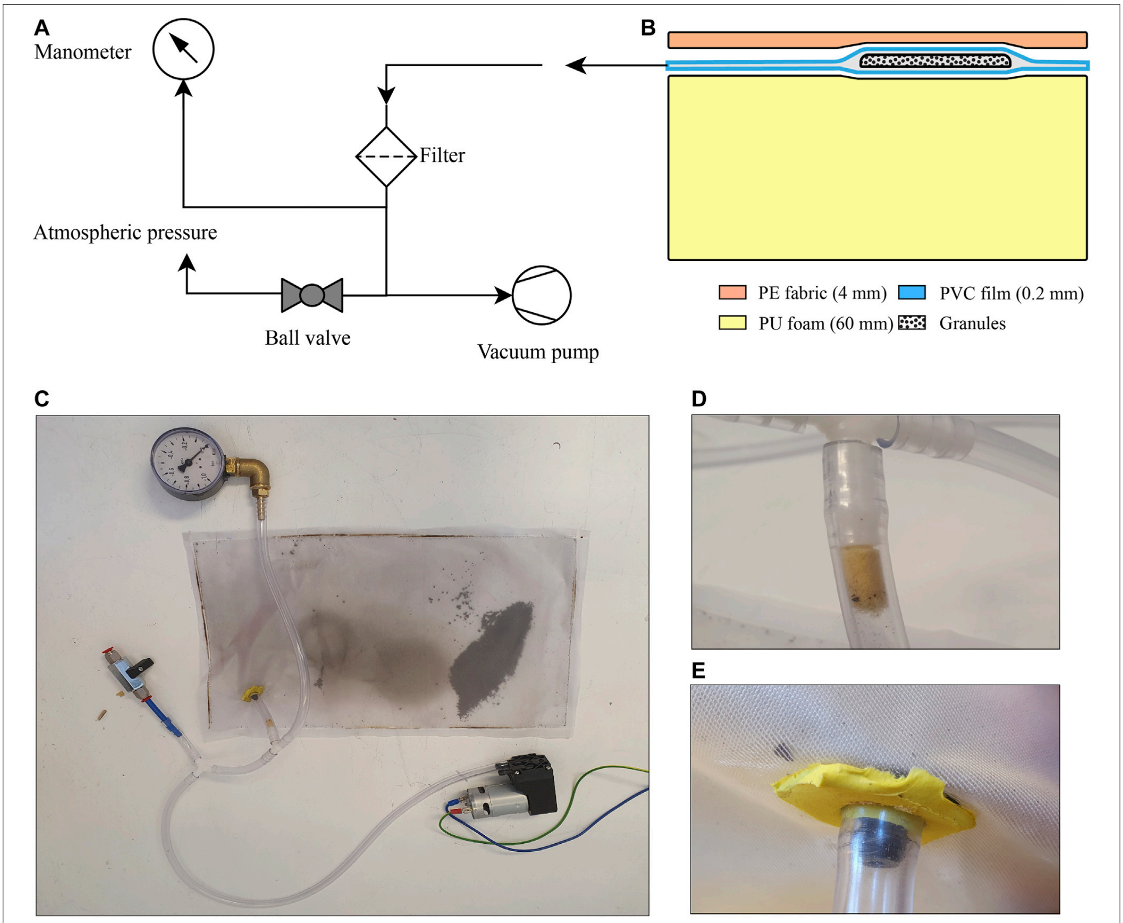


FIGURE 5 | (A) Schematic presentation of the pneumatic setup **(B)** Palpation interface with layer material and thickness **(C)** The pneumatic setup **(D)** Filter **(E)** Inlet seal for vacuum chamber.

area in **Figure 6A**. We used 12 g of filtered ferromagnetic granulate in the chamber.

3.2 Experiment Design

All participants repeated the palpation task four times, under four different conditions. The irregularity could differ in hardness, position and form. The four conditions were as follows:

- C1: Baseline. No irregularity in the palpation field.
- C2: Annulus. Annular-shaped irregularity rendered with the magnet configuration seen in **Figure 6B**. Vacuum level: −0.4 to −0.6 bar, whereas −1 bar is a complete vacuum. Located in the lower left part of the field. Approximately 82 mm outer diameter and 29 mm inner diameter with area M = 4,915 mm^2 SD = 371 mm^2 SE = 70 mm^2.
- C3: Hard Lump. A circular-shaped irregularity rendered with the magnet configuration seen in **Figure 5B**. Vacuum level: 0.4 to −0.6 bar. Located in the top right part of the field. Approximately 100 mm diameter with area M = 7,912 mm^2 SD = 474 mm^2 SE = 89 mm^2.
- C4: Soft Lump. A circular-shaped irregularity rendered with the magnet configuration seen in **Figure 5C**. Vacuum level: 0.1 bar. Located in the top right part of the field. Approximately 100 mm diameter with area M = 8,094 mm^2 SD = 641 mm^2 SE = 121 mm^2.

The sequence of the testing conditions was randomized to avoid potential learning or order effects. The order of conditions was also balanced, i.e., they appear the same number of times in each procedure step.

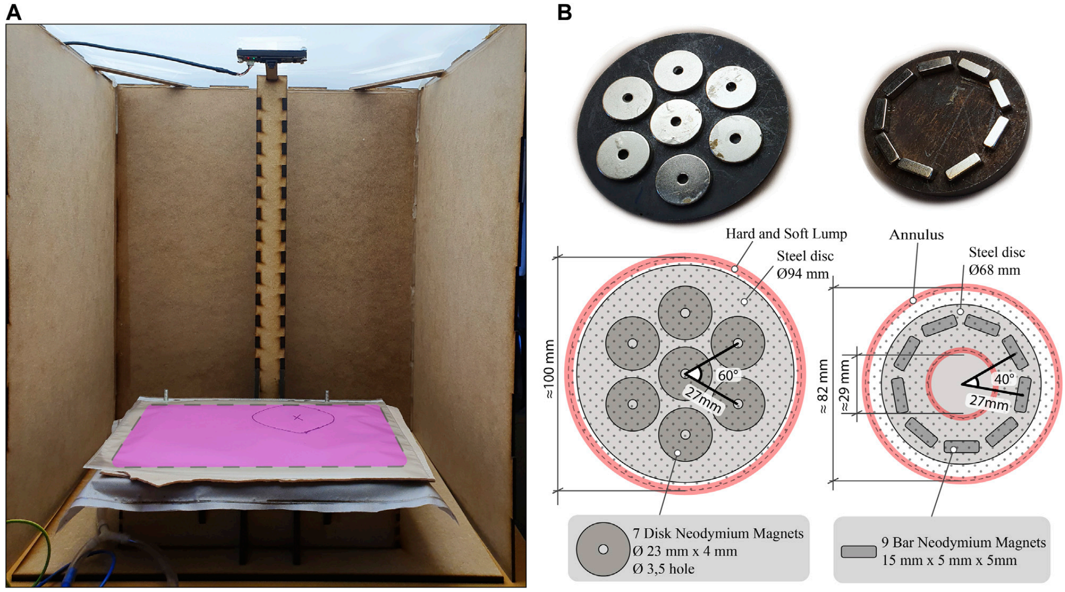


FIGURE 6 | (A) Test rig with camera and cabinet setup (pink area is the palpable field) **(B)** Magnet arrangement with angular and radial distance shown. Approximated outlines for the generated outputs are also illustrated with measurements. (Left) Circular shape for Hard and Soft Lump (Right) Annulus shape with the hollow center.

3.2.1 Participants

N = 28 healthy engineering students were recruited to participate (21 male (75%) and 7 female (25%)). Twenty-seven participants were in the 21–29 years range and one participant in the 18–20 years range. None of the participants were trained in the test or had any relevant knowledge about the technology before participation. Participation was voluntary, and all gave informed consent to be part of the study.

3.2.2 Experimental Procedure and Data Collection

The experimental procedure can be seen in **Figure 7**. After signing a consent form, the participants filled out a demographic questionnaire. Durometer and manometer readings and pictures of the granulate were sampled before the participant was seated in front of the test setup. The hardness of the irregularities was measured with a commercially available Shore durometer (Shenzhen Gairan Tech Co., X.F Type 00), following the requirements described in ISO 48–4:2018. A minimum of three measurements at different positions on the flat parts of the irregularity was performed. After objective data was collected, participants were instructed regarding the proceedings of the experiment. First, participants were told to palpate for a potential irregularity and say stop when they had control of the position and form. The participants did not get any instructions regarding technique to be used, other than using their hands to explore and feel for any irregularities in the field. We measured the time the participant used to find position and form of the irregularity. After completing each palpation, participants were asked to draw the contours of a possible irregularity on a sheet placed above the palpation field. More specifically, they were told to draw the outside and possible inside contours and put an x inside the area enclosing the proposedly identified irregularity (see **Figure 8B**). Pilot experiments showed that this instruction facilitated the participants who found an inside contour to also draw it, instead of drawing the outer contour only. Drawing data were captured using the camera.

To evaluate the hardness of the irregularity, a sampled selection of objects of varying hardness was used. These samples were numbered from 1 to 5 and had a Shore hardness of 00–20, 00–35, 00–55, 00–65, and 00–90, from soft to harder. The objects were presented similarly to the test setup using the same deformable backing and palpable cover as the palpation field. Thus, the participant could palpate the irregularity when doing the hardness test.

After each condition, participants reported their degree of agreement to a series of statements using a Likert Scale from 1 (Totally disagree) to 5 (Totally agree). The statements were: 1) It was hard to find the irregularity. 2) I am confident that I found the position and shape of the irregularity. 3) The irregularity had a constant/homogeneous hardness. To get a measure of a potential learning effect occurring during the experiments, participants also evaluated the statement: 4) I became better at finding the irregularity during the experiment., after completing the experiment.

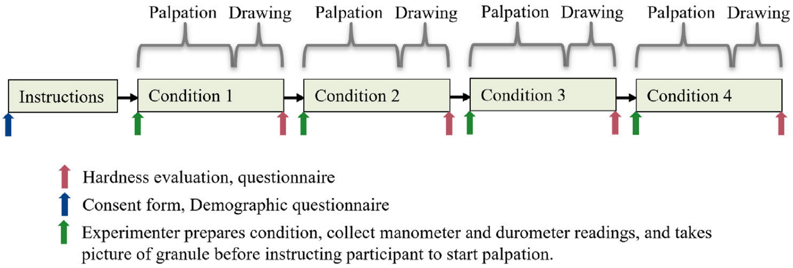

Palpation
Drawing
Palpation
Drawing
Palpation
Drawing
Palpation
Drawing
Instructions
Condition 1
Condition 2
Condition 3
Condition 4
Hardness evaluation, questionnaire
Consent form, Demographic questionnaire
Experimenter prepares condition, collect manometer and durometer readings, and takes picture of granule before instructing participant to start palpation.

FIGURE 7 | Experimental procedure timeline. The order of the conditions was randomized.

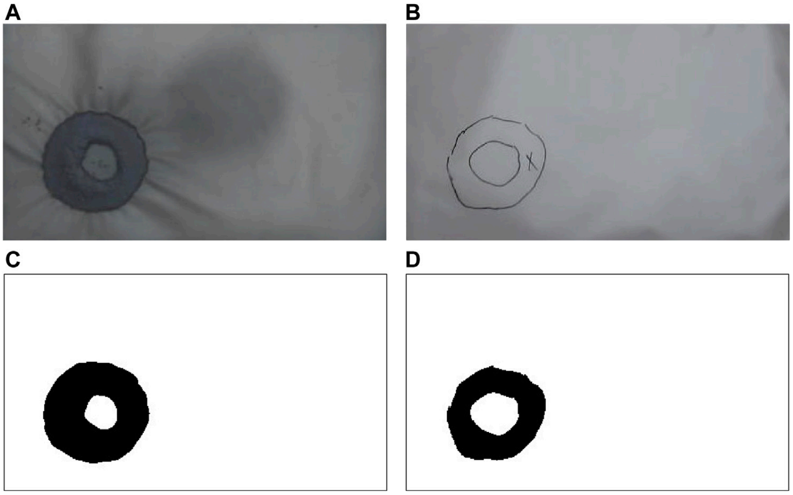


FIGURE 8 | Unprocessed images of **(A)** Granulate and **(B)** Drawing. Binarized pictures of **(C)** Granulate and **(D)** Drawing.

3.3 Data Analysis

The data was collected throughout the experiment to answer the study's research questions. Thus, experiment pictures were processed into binarized matrices that yielded objective data points describing irregularities' and drawings' respective positions and geometrical form. These data, together with the questionnaire and objective measurements, were statistically analyzed for reliability and differences between variables with SPSS Statistics (IBM SPSS Statistics 27, 2020).

3.3.1 Picture Processing

Data about position and form was collected through images. The images of the granulate and drawings were then processed and analyzed using package OpenCV 4.5.1 in Python 3.0. The capturing code also took photos of the manometer during each test. The images were blurred before grayscaling and binarizing to remove noise. An adaptive Gaussian threshold was used on the pictures of the drawings to improve accuracy. The binarized results are shown in **Figures 8C,D**.

3.3.1.1 Distance From Center to Center

The center point distance between granulate and drawings were calculated by finding the center of mass for both the granulate and the drawings using cv2.moments in Python. Then, the Euclidean distance (ΔD) was calculated between the two coordinates, using **Eq. 1**. x_1 and y_1 representing the coordinates for the granulate, while x_2 and y_2 representing the drawing.

$$\Delta D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}$$

3.3.1.2 Intersection Over Union

Intersection over Union (IoU) was used to evaluate the form. First, matrices of the intersection and union of the two binarized pictures were calculated using Python. Then, the number of black

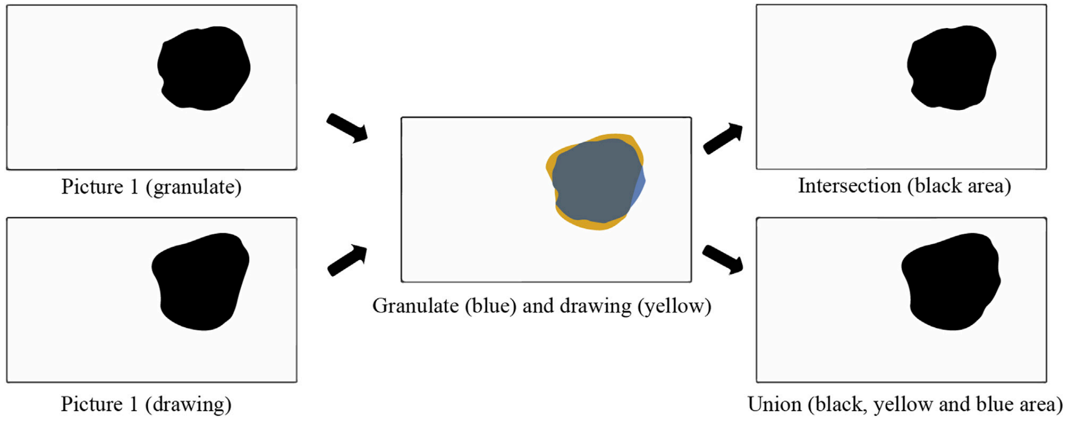

Picture 1 (granulate)
Picture 1 (drawing)
Granulate (blue) and drawing (yellow)
Intersection (black area)
Union (black, yellow and blue area)

FIGURE 9 | Visual presentation of intersection and union.

pixels (pixels with value 0) in the intersection was divided by the number of black pixels in the union, using **Eq. 2**. A visual representation of intersection and union is shown in **Figure 9**.

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \tag{2}$$

3.4 Statistical Tests

To assess reliability, Fleiss' kappa was ran to determine if there was an agreement between participants' judgment of whether there was an irregularity or not (Lump or No Lump) in the four conditions. Fleiss' kappa does not assume that the raters are identical for each condition (which is the case here), but this is the only test we know of that assesses the case when there are multiple raters. Therefore, we report this test along with the frequency. One-way repeated measures ANOVAs were used to investigate differences between conditions for continuous variables. Those were: IoU, hardness (durometer reading) and vacuum level (manometer reading). Assumptions regarding no outliers, normality, and sphericity were inspected with boxplots, histograms and Normal Q-Q Plots, and Mauchly's Test of Sphericity. Violations of the outlier assumption were not removed since it only applied to Durometer and Manometer readings, which were used to corroborate that the conditions Hard Lump and Soft Lump differed in terms of hardness. In addition, a Friedman test was also conducted to ensure similar differences. A Greenhouse-Geisser correction was applied in the case of violating sphericity (Wickens and Keppel, 2004; Field, 2018). A Friedman test was used to investigate differences between conditions for discontinuous variables (the remaining variables), and in the case of more severe violations to ANOVA's assumptions. Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons for both ANOVA and Friedman. Some variables produced a statistically significant Friedman test, but without any significant pairwise comparisons. One reason might be the conservative nature of the multiple comparisons correction. An additional approach, multiple Wilcoxon signed-rank tests, was therefore used to follow up the Friedman tests. We deemed it acceptable to be less conservative since it is the first investigation of an early-stage prototype, and it was important to gain an understanding of where potential differences were. The Wilcoxon signed-rank tests was also used to obtain a z-score, used to estimate effect size (r) (Rosenthal, 1986; Field, 2018). For the ANOVAs, the sample effect size partial eta squared (η^2), and population effect size partial omega squared (ω^2) (Rosenthal, 1986) are reported. The significance level $p < 0.05$ was chosen for highly significant differences. p–values ≤ 0.10 were considered as interesting effects, again due to the experiment involving human participants evaluating an early-stage prototype. We believe a 10% probability for Type 1 error is acceptable in this case.

4 RESULTS

Both objective and subjective data points were gathered throughout all four conditions described in 3.2.2. Each condition focused on localizing and characterizing a potential irregularity based on position, form and hardness. Additional descriptive statistics can be found in **Supplementary Material**.

4.1 Lump or No Lump: How Many Found an Irregularity?

In all three conditions with an irregularity (Annulus, Hard Lump and Soft Lump), all participants found an irregularity (100%

agreement). In Baseline condition six participants (21.4%) found an irregularity, despite there not being one. The remaining 22 participants (78.6%) failed to find an irregularity. Fleiss' kappa determined if there was an agreement between participants' judgment of whether there was an irregularity or not (Lump or No Lump) in the four conditions. The agreement between participants' judgements was statistically significant with κ= 0.723, 95% CI [0.722, 0.725], $p < 0.001$. The individual kappa's for Lump and No Lump categories were also κ= 0.723, 95% CI [0.722, 0.725], $p < 0.001$. This statistic is the proportion of agreement over and above chance agreement, with 0 being no agreement and 1 being perfect agreement. An agreement of 0.723 can be classified as a good agreement (Landis and Koch, 1977).

As stated, six out of 28 participants found an irregularity in the baseline condition. Of these six, three participants drew contours with areas of 22, 64 and 147 mm^2, which are small compared to the actual size of the irregularities. They are similar to granular remnants, which means they could be discarded as an error in the setup. Other participants commented on particle-sized irregularities in the Baseline condition but decided that they were not of sufficient size to be an actual irregularity. Removing these three participants results in three participants (12.0%) finding an irregularity in the Baseline condition, whereas 22 participants (88%) did not find an irregularity. Fleiss' kappa was ran again with these three participants removed to investigate the magnitude of the potential error from the setup. The agreement between the remaining 25 participants was statistically significant with κ= 0.840, 95% CI [0.783, 0.896], $p < 0.001$. The individual kappa's for Lump and No Lump categories were also κ= 0.840, 95% CI [0.783, 0.896], $p < 0.001$.

22 of 28 (78.57%) of the participants found the inner circle. In the two irregularity conditions 50 of 56 (89.29%) drawings were filled circles without any inner contour.

In summary, all participants agreed that there was an irregularity present in all irregularity conditions. Despite a few participants finding an irregularity where there was none, the overall agreement between participants was high.

4.2 Time

The users were not instructed to be as fast as possible but rather spend enough time to be sure of position and form of the irregularity. Therefore, the time represents the procedure time needed to find position and form of the irregularity to the best of the participant's ability.

Time was statistically significantly different in the four conditions, $\chi^2(3) = 29.460$, $p < 0.001$ as shown in **Figure 10**. Post hoc analysis revealed significant differences in Time from Baseline (Mdn = 37.0 s), 95% CI [25.0, 61.0] to Annulus (Mdn = 13.50 s), 95% CI [7.0, 21.0] ($p < 0.001$) and Soft Lump (Mdn = 16.00), 95% CI [11.0, 40.0] ($p = 0.001$) condition. It took longer to determine that there was no irregularity in Baseline condition, compared to finding it in Annulus and Soft Lump condition. The contrast comparing Annulus to Hard Lump (Mdn = 17.50), achieved a significance level $p = 0.050$ and effect size r = −0.50, and Hard Lump to Baseline had a significance level of $p = 0.067$ and effect size r = 0.48. We interpret this to be a notable difference. There was no significant difference between Annulus and Soft Lump (p = 0.80, r = −0.28) and Soft Lump and Hard Lump (p = 1.00, r = 0.22).

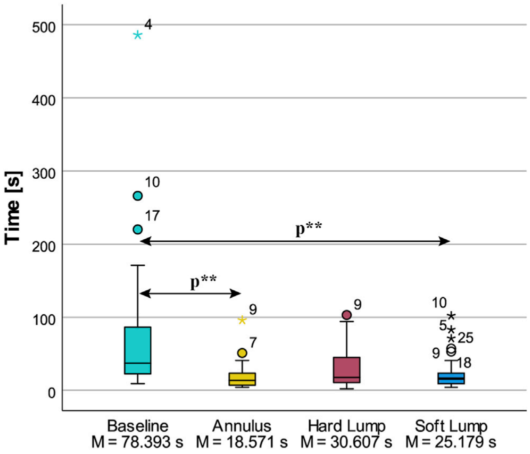


FIGURE 10 | Descriptive statistics of procedure time to find the position and form of the irregularities. Statistically significant differences at $p < 0.05$ are indicated by p^{**}.

4.3 Position: Distance From Center to Center

The distance between the center of the irregularity to the center of the participants' drawing was statistically significantly different in the three irregularity conditions, χ^2 (2) = 16.357, $p < 0.001$. Post hoc analysis revealed statistically significant differences in center distance from Annulus (Mdn = 5.2920 mm), 95% CI [2.978, 7.097], to Hard Lump (Mdn = 7.4366 mm), 95% CI [6.331, 12.217] ($p < 0.001$), and from Soft Lump (Mdn = 6.3908 mm), 95% CI [3.836, 9.654], to Hard Lump condition (*Mdn* = 7.4366 mm) (p = 0.01). There was no significant difference in center distance between Annulus and Soft Lump (p = 1). We also observe that there was a greater spread in the Hard Lump condition. These results are plotted in **Figure 11A**.

4.4 Form: IoU

IoU was statistically significantly different in the three conditions, χ^2 (2) = 12.071, p = 0.002. Post hoc pairwise comparisons with a Bonferroni correction for multiple comparisons yielded one significant difference between Soft Lump (Mdn = 0.793), 95% CI [0.728, 0.825], and Annulus (Mdn = 0.688), 95% CI [0.579, 0.735], p = 0.002, and a corrected p = 0.247 for both the Soft Lump vs Hard Lump (Mdn = 0.767), 95% CI [0.568, 0.754] comparison, and Hard Lump vs Annulus comparison (uncorrected p-value was p = 0.082). Post hoc Wilcoxon tests revealed a statistically significant difference between Annulus (Mdn = 0.688) and Soft Lump (Mdn = 0.793), T = 316.00, p = 0.010, r = 0.49, and a significant difference between Hard Lump (Mdn = 0.767), 95%

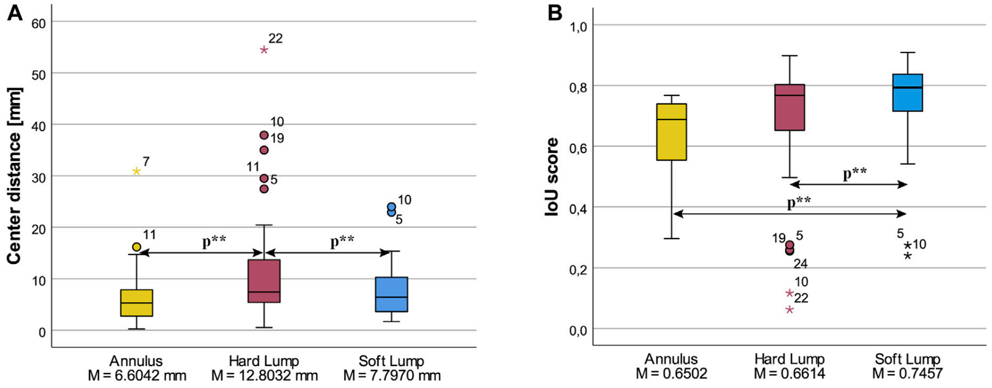


FIGURE 11 | Box plots of **(A)** Center distance of granulate and drawing, and **(B)** Intersection over Union, for the three lump conditions. Statistically significant differences at $p < 0.05$ are indicated by *p***.

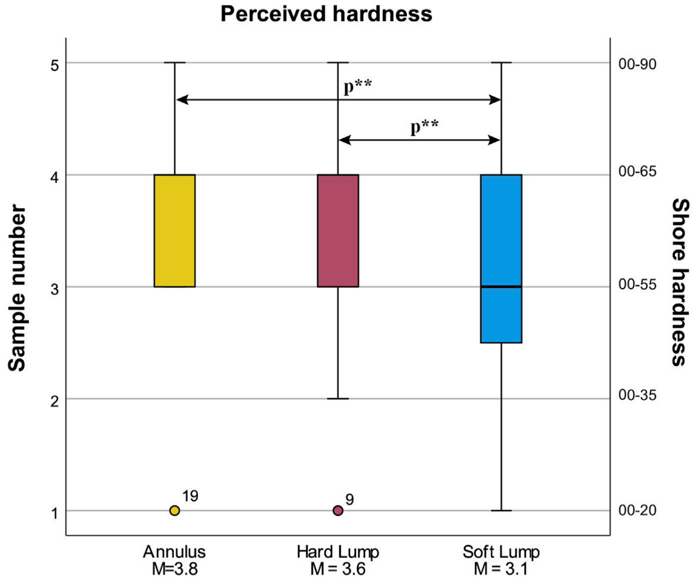


FIGURE 12 | Box plot of perceived hardness (subjective) for the three lump conditions. Statistically significant differences at $p < 0.05$ are indicated by *p***.

CI [0.703, 0.0794], and Soft Lump (Mdn = 0.793), $T = 304.00$, $p = 0.\ 021$, $r = 0.43$. There was no difference between Annulus and Hard Lump, $T = 263.00$, $p = 0.\ 021$, $r = 0.26$ as plotted in **Figure 11B**.

4.5 Hardness

We compared perceived hardness, objective hardness measurements, and vacuum levels of the irregularity conditions.

4.5.1 Perceived Hardness

Perceived hardness was statistically significantly different in the three conditions, χ^2 (2) = 9.129, $p = 0.010$. Post hoc pairwise comparisons with a Bonferroni correction for multiple comparisons yielded one significant difference between Soft Lump (Mdn = 3) and Annulus (Mdn = 4), $p = 0.033$, and a corrected $p = 0.184$ for the Soft Lump and Hard Lump comparison (uncorrected p-value was $p = 0.061$). Post hoc Wilcoxon tests revealed a statistically significant difference

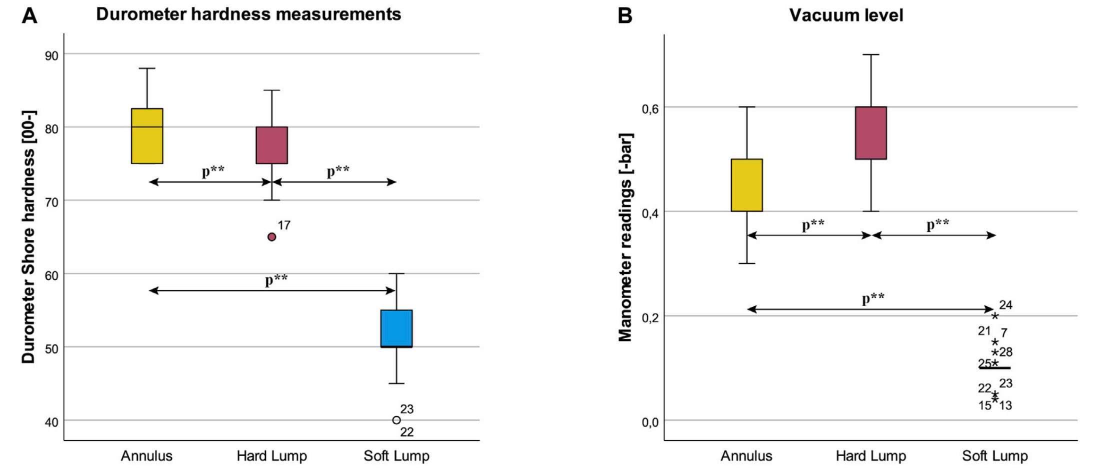


FIGURE 13 | Box plots of **(A)** Durometer hardness measurements and **(B)** vacuum level for the three lump conditions. Statistically significant differences at $p < 0.05$ are indicated by p^{**}.

between Soft Lump (Mdn = 3) and Hard Lump (Mdn = 4), T = 132.00, p = 0. 032, r = −0.41, and a significant difference between Annulus (Mdn = 4) and Soft Lump (Mdn = 3), T = 55.00, p = 0.015, r = −0.46. There was no difference between Annulus and Hard Lump, T = 87.00, p = 0. 474, r = −0.14. These results are as expected. Participants perceived the hardness of Soft Lump to be less than that of both Hard Lump and Annulus which is shown in **Figure 12**.

4.5.2 Durometric Measurements

There were 3 outliers as assessed by boxplot in **Figure 13A**. By visual inspection, the data was approximately normally distributed. The assumption of sphericity was violated, as assessed by Mauchly's test of sphericity, $\chi^2(2)$ = 6.825, p = 0.033. Therefore, a Greenhouse-Geisser correction was applied (ε = 0. 812). Results was statistically significant different in the three conditions F (1.625, 43.872) = 278.699, p < 0.001, η^2 = 0.912, ω^2 = 0.869. Durometric readings were: Annulus (M = 79.96), Hard Lump (M = 76.79), Soft Lump (M = 51.25). Post hoc analysis with a Bonferroni correction yielded statistically significantly difference between Annulus and Hard Lump (M = 3.179, 95% CI [0.63, 5.72], p = 0.011), between Annulus and Soft Lump (M = 0.28.714, 95% CI [24.71, 32.72], p < 0.001), and between Hard Lump and Soft Lump (M = 25.536, 95% CI [22.04, 0.29.03], p < 0.001).

4.5.3 Manometer

There were several outliers as assessed by boxplot in **Figure 13B**. The data was approximately normally distributed by visual inspection. The assumption of sphericity was violated, as assessed by Mauchly's test of sphericity, $\chi^2(2)$ = 13.713, p = 0.001. Therefore, a Greenhouse-Geisser correction was applied (ε = 0. 709). Manometer was statistically significant different in the three conditions F (1.419, 38.301) = 228.636, p <0.001, η^2 = 0.894, ω^2 = 0.844. Manometer readings were: Annulus (M = 0.431), Hard Lump (M = 0.536), Soft Lump (M = 0.101). Post hoc analysis with a Bonferroni adjustment was statistically significantly different between Annulus and Hard Lump (M = −0.105, 95% CI [−0.17, −0.04], p = 0.002), between Annulus and Soft Lump (M = 0.330, 95% CI [0.29, 0.368], p < 0.001), and between Hard Lump and Soft Lump (M = 0.435, 95% CI [0.38, 0.49], p <0.001).

4.6 Questionnaire

Participants completed the questionnaire in the three irregularity conditions.

4.6.1 How Hard Was It to Find the Position?

Participants' evaluation of how hard it was to find the irregularity was statistically significantly different in the three conditions, $\chi^2(2)$ = 7.423, p = 0.024. Post hoc pairwise comparisons with a Bonferroni correction for multiple comparisons yielded no significant differences. Post hoc Wilcoxon tests revealed a statistically significant difference between Soft Lump (Mdn = 1) and Hard Lump (Mdn = 1), T = 63.00, p = 0.006, r = −0.52. There were no significant differences between Annulus (Mdn = 1) and Soft Lump (Mdn = 1), T = 27.00, p = 0.957, r = −0.01, or between Annulus and Hard Lump, T = 89.00, p = 0.096, r = −0.32. Participants found it hardest to locate the irregularity in the Hard Lump (see **Figure 14A**).

4.6.2 Confidence in Finding Position and Shape of the Irregularity

Participants' confidence in finding position and shape of the irregularity was statistically significantly different in the three irregularity conditions, $\chi^2(2)$ = 8.926, p = 0.012. Post hoc pairwise comparisons with a Bonferroni correction for multiple

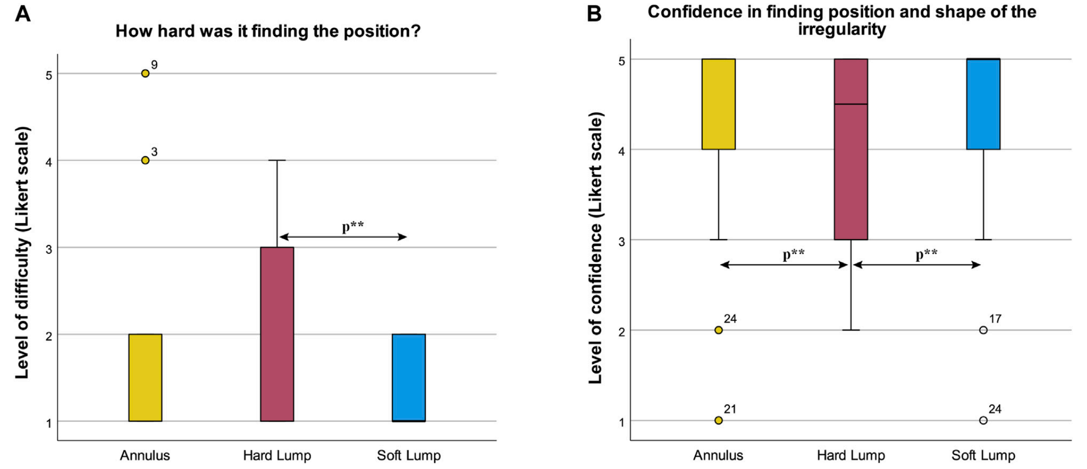


FIGURE 14 | Box plot of results from the questionnaire **(A)** How hard was it finding the position, and **(B)** Self-assessment of confidence in the accuracy of their drawing.

comparisons yielded no significant differences. Post hoc Wilcoxon tests revealed a statistically significant decrease in confidence from Annulus (Mdn = 5) to Hard Lump (Mdn = 4.5), $T = 12.00$, $p = 0.032$, r = –0.40, and a significant decrease in confidence from Soft Lump (Mdn = 5) to Hard Lump (Mdn = 4.5), $T = 13.50$, $p = 0.038$, $r = -0.39$. There was no significant difference between Annulus and Soft Lump, $T = 31.00$, $p = 0.276$, $r = -0.21$. Participants were most confident in finding the position and shape of the irregularity in the Annulus and Soft Lump condition and less confident in finding the irregularity's position and shape in the Hard Lump condition (see **Figure 14B**).

4.6.3 Homogeneous Hardness

Participants' evaluation of whether the irregularity had a constant/homogeneous hardness was not significantly different in the three irregularity conditions, χ^2 (2) = 3.410, $p = 0.182$.

4.6.4 Self-Assessed Improvement in Palpation

After completing the experiment, participants evaluated whether they thought they improved their palpation skills throughout the experiment. 20 participants (71.4%) thought they improved, 3 disagreed (10.7%), and 5 were neutral (17.9%).

5 DISCUSSION

Introducing ferromagnetic granules in a jamming haptic interface has the quality to be a promising solution to produce larger tactile displays cheaply with high accuracy. Palpation trainers need to be robust, safe and have a high level of repeatability. Using adaptable palpation trainers increases the number of study cases and task-specific functionalities the trainer can accomplish. Thus, we think our concept can be taken further for use in a medical training equipment environment. Before that, however, there is a need for further development of the technology and contextual testing.

When comparing our data with relevant research (as mentioned in the Introduction), we have good results for people's perception of both hardness and position. Bergmann Tiest and Kappers (2009) states that users are pretty good at determining hardness. Frisoli et al. (2011) states that cutaneous sensor modality is not affected by size, but kinesthetic performance is reduced with smaller-sized objects. To our knowledge, there is a lack of research on people's perceptual exploration and characterization. Thus, this study could add to the body of knowledge concerning this aspect of both machine interaction and human tactile perception. The following section discusses the objective and subjective results gathered and how they answer our research questions. Further, an evaluation of the participant sampling is presented before we discuss the limitations and outlook of the study.

5.1 Interpreting Results

We defined that our haptic interface should be able to change the position, form and hardness of an irregularity recognizable by palpation. We chose a circle and an annulus as our two shapes to evaluate if people could locate and characterize them. A major part of the participants could differentiate between the circular lumps and the Annulus (78.57% recognized the inner circle of the Annulus, and 89.29% drew the circular lumps with no inner circle). Furthermore, all the participants found an irregularity in all conditions that had an irregularity. For the Baseline condition, six participants found a false positive. When determining the participants' ability to describe the form, we used Intersection over Union (IoU). The median was promising for all three irregularity conditions, with the highest score for Soft Lump (0.793). From these three observations, we could conclude that the overall agreement between participants was great for form. Thus, our concept can manipulate the granules into different shapes that laypersons can distinguish by palpation.

However, an interesting result is that the IoU was significantly lower for Annulus than for Hard Lump and Soft Lump, while Annulus scored best at center point distance. A more logical assumption would be that IoU and center point distance is inversely proportional. There could be at least two reasons why we get a lower IoU score for Annulus. Initially, we observed from the participants contouring the Annulus that they struggled to get the size and position of the inner circle right. Due to how we calculated IoU, a wrong positioned hole yielded a more considerable difference in IoU than a similar error in outer contour. Also, the same error in the center point difference for Lump and Annulus gives a more significant change in IoU for Annulus because of the inner circle.

The results show a statistical difference for both objectively measured and perceived hardness between Soft lump and Hard Lump. Furthermore, when performing Wilcoxon tests on perceived hardness, there was a significantly lower value of the Soft Lump than the Hard Lump and Annulus condition. Thus, we have shown that participants can distinguish between hard and soft objects that the prototype produces, which is essential for palpation tasks, whereas characterizing the physical attributes such as form and hardness of identified irregularities is essential.

Considering how participants conducted the palpation tasks, time spent is of interest. The five most extended procedure times were on baseline condition, and three of them had baseline as the first condition. For example, participant No. 4 stated that there was no irregularity after 90 s before spending six more minutes to palpate before finding a false positive. No. 21 expressed hesitancy after 1 minute, and then spent two more minutes palpating before concluding with a true negative. No. 17 expressed insecurity before spending 2–3 more minutes searching, ending with a true negative. From the respective participant's confidence data, the participant with the false positive (no. 4) answer a four on the Likert scale, i.e., partially agreeing that they were sure they found the correct position and form of the irregularity. Also, all three participants had good results in all irregularity conditions. This connection could mean that some people struggle to trust their sense of touch. The baseline condition presents ambiguity as there is nothing to palpate, and we believe having it as the first condition increased ambiguity and thus uncertainty in participants who probably expected an irregularity.

Another aspect is the repeatability of our testing equipment. We tried to develop a haptic interface that can alter and maintain hardness, position and form of palpable outputs with high repeatability. However, while prototyping the granular jammer, it became apparent that repeatedly creating geometries with identical shape and hardness was challenging. While the outline for the shapes varied for each sample as a result of the manual setup and granule dispersion, the gathered images showed only a small deviation of rendered area for each irregularity condition. The durometric measurements did, however, show a wide hardness range within a prepared condition. This inhomogeneous hardness from granular jamming is similar to the findings of Genecov, Stanley, and Okamura (2014). We thought of two reasons for this, firstly, how the hardness is highly dependent on how the granules interlock or position themselves across the irregularity. Secondly, because the arrangement of granules was made manually, there was an unavoidable variation in the produced output geometries and thus granulate concentration across the area.

Considering the pressure readings for the setup, Soft Lump had more outliers as a result of the vacuum level being manually set (and adjusted). Compared to Hard Lump and Annulus which had a hard stop, governed by the maximum vacuum the setup could provide. Given these being different geometries thus yielding different volumes to drain the air from, this could cause the difference in obtained vacuum level. However, our results from the perceived hardness showed no significant difference between Annulus and Hard Lump. As this being a first prototype, challenges concerning repeatability is expected and the overall results show great promise for this concept to be improved further to address these limitations.

When looking at the questionnaire data, confidence was highest in finding the position and shape of the Annulus and Soft Lump condition. We expected that the Hard Lump would be the easiest to find due to a sharper edge and thus greater difference between the Hard Lump and the palpable area. However, this was not the case. In retrospect we suspect this was due to the increased vacuum level instantly jamming the granules not allowing them to evenly distribute and conform to a smooth shape. This could cause the edges of the Hard Lump to be more jagged/rivet than the edge of the Soft Lump, which had a more circular shape in comparison. We therefore believe this might have made participants more uncertain of where the edge was. Moreover, since the edge of the Hard Lump varied more compared to a circle, this may have contributed to that it was more difficult to find the center of it.

Participants reported they got better at finding the irregularity throughout the experiment, meaning it could be used as a training device for palpation exercises. However, a reported high level of confidence and low level of difficulty for each condition could mean the task being too easy to perform, not leaving much room for progression and learning. Given the ability to find and characterize the irregularities sufficiently, and having a high confidence in doing so, further steps should be made to tailor level of difficulty to specific scenarios and investigating the use of the device in a medical context. The concept has been experimentally, shown to facilitate users' palpation skills by speed, location, shape, and hardness differentiation of palpable findings. Other learning objectives could involve motoric technique and following procedural algorithms, which should be explored in further development of the concept.

5.2 Participant Sampling

In this experiment, the 28 participants were all engineering students who did not have any previous experience with palpation as a medical examination technique. The participants did not get any technical or strategic instructions, meaning their palpation approach would be different to a medical professional. Therefore, we have shown that our concept works for presenting generic geometric shapes for laypersons, which is a promising result considering this a training device. Moreover, having participants with prior medical experience, would thus require a higher level of difficulty. A sample size of 28 was adequate compared to similar studies (Gerling et al., 2003; Asgar-Deen et al., 2020). Also, as we got statistically significant differences between relevant data points, such as the hardness of Soft Lump and Hard Lump, more participants would most likely not produce other results. However, a higher sample size would reduce the possibility of an accepted hypothesis being incorrect.

5.3 Limitations and Further Work

In this research, sensing, automation or participant feedback has not been addressed nor implemented in the palpation concept. The prototype and subsequently the experiments with the prototype are not tied to a medical context. Instead, it explores some of the capabilities and extreme conditions the haptic device can output. Hence, it has not been within the scope of this research to model, synthesize, or simulate physiological attributes for palpation. Nevertheless, we lack to prove that our haptic interface is helpful in medical training because of a simplified experiment focusing on planar perceptual exploration. Therefore, in further development, more levels of difficulty, complex geometries, hardness profiles, locations, and dynamic abilities should be explored. As palpation tasks seldom concerns irregularities in one plane, investigating multiple jamming layers, or simulating depth of palpation by dynamic stiffness control should be investigated. A positive backpressure in combination with ferrogranular jamming could yield 3D-shapes with high tactile resolution and geometrical freedom (Koehler et al., 2020; He et al., 2021). In further work, we seek to test the concept with users who can provide feedback and evaluation on a medical basis. This could reveal hitherto unexplored concept potentials and critical functions to pursue.

6 CONCLUSION

This work has described the development and testing of a novel haptic interface concept that uses ferrogranular jamming. This concept was developed as a compliant simulation interface for medical palpation training, with the ability to simulate geometrical objects of various shapes and tactile properties. The concept was tested by having 28 untrained participants perform a set of structured palpation tasks in an experiment. The experiment consisted of four conditions, one baseline and three containing a palpable irregularity. These conditions were chosen to evaluate if the interface could produce various shapes and hardness levels, while also investigating participants' palpation skills. Given the results of the experiment, we conclude that the concept can create palpable objects with variable hardness by adjusting the jamming vacuum. Laypersons can distinguish these objects by palpation, both by the hardness, location, and shape of objects with good accuracy. Thus, this study also provides insights on peoples' perceptual abilities in explorative palpation. It shows the ability to locate and characterize palpable objects of varying shape and hardness in a satisfactory manner. Further, the results show that, the task was not considered very challenging. This combined with participants reported high level of confidence in performance, indicates that increased difficulty might be required to ensure room for improvement and learning. However, as participants also reported improvements in their palpation skills during the experiment, the technology looks promising to be further developed for medical training applications.

Considering this being an early conceptual prototype, this study revealed opportunities and challenges yet to be addressed. In further work, we want to explore whether the interface can be used as a palpation tool in medical simulation by qualitative testing with expert users. This will require palpable objects where both hardness, shape, and difficulty are tailored to the medical scenarios we want to simulate. Other technical aspects of the ferrogranular jamming concept we want to explore are sensing and feedback, automation, and dynamic and responsive tactile abilities. Collectively, this could improve the experience of using this technology in simulation-based medical training.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article is available in the following repository. SBR; MA; HD; and MS, 2021, Replication Data for: Perception by Palpation: Development and Testing of a Haptic Ferrogranular Jamming Surface, https://doi.org/10.18710/OCMXVP, DataverseNO, V1.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed ideas and concepts throughout the development of the device. SR did the technical work and execution of the experiment. MS provided laboratory and testing facility. All authors contributed and agreed to the final wording of the submission. HD conducted the statistical analysis. MA contributed to the development of the concept and design of the experiment.

FUNDING

This research is supported by the Research Council of Norway through its industrial Ph.D. funding scheme, project number 290404.

ACKNOWLEDGMENTS

Thanks to Sindre Wold Eikevåg for the experiment procedure timeline figure and thanks to Håvard Vestad for creative ideas concerning the ferrogranular jamming concept.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.745234/full#supplementary-material.

REFERENCES

Asgar-Deen, D., Carriere, J., Wiebe, E., Peiris, L., Duha, A., and Tavakoli, M. (2020). Augmented Reality Guided Needle Biopsy of Soft Tissue: A Pilot Study. *Front. Robot. AI* 7, 72. doi:10.3389/frobt.2020.00072

Bergmann Tiest, W. M., and Kappers, A. (2009). Cues for Haptic Perception of Compliance. *IEEE Trans. Haptics* 2, 189–199. doi:10.1109/TOH.2009.16

Brown, E., Rodenberg, N., Amend, J., Mozeika, A., Steltz, E., Zakin, M. R., et al. (2010). Universal Robotic Gripper Based on the Jamming of Granular Material. *Proc. Natl. Acad. Sci.* 107, 18809–18814. doi:10.1073/pnas.1003250107

Cao, Q., Han, X., and Li, L. (2014). Configurations and Control of Magnetic fields for Manipulating Magnetic Particles in Microfluidic Applications: Magnet Systems and Manipulation Mechanisms. *Lab. Chip* 14, 2762–2777. doi:10.1039/c4lc00367e

Carruth, D. W. (2017). “Virtual Reality for Education and Workforce Training,” in 2017 15th International Conference on Emerging eLearning Technologies and Applications (ICETA) (IEEE), 1–6. doi:10.1109/ICETA.2017.8102472

Cates, M. E., Wittmer, J. P., Bouchaud, J.-P., and Claudin, P. (1998). Jamming, Force Chains, and Fragile Matter. *Phys. Rev. Lett.* 81, 1841–1844. doi:10.1103/PhysRevLett.81.1841

Cheng, N. G., Lobovsky, M. B., Keating, S. J., Setapen, A. M., Gero, K. I., Hosoi, A. E., et al. (2012). “Design and Analysis of a Robust, Low-Cost, Highly Articulated Manipulator Enabled by Jamming of Granular media,” in 2012 IEEE International Conference On Robotics And Automation (IEEE), 4328–4333. doi:10.1109/ICRA.2012.6225373

D’Avella, S., Tripicchio, P., and Avizzano, C. A. (2020). A Study on Picking Objects in Cluttered Environments: Exploiting Depth Features for a Custom Low-Cost Universal Jamming Gripper. *Robotics and Computer-Integrated Manufacturing* 63, 101888. doi:10.1016/j.rcim.2019.101888

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*. Los Angeles: SAGE.

Fitzgerald, S. G., Delaney, G. W., and Howard, D. (2020). A Review of Jamming Actuation in Soft Robotics. *Actuators* 9, 104. doi:10.3390/act9040104

Follmer, S., Leithinger, D., Olwal, A., Cheng, N., and Ishii, H. (2012). “Jamming User Interfaces,” in Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology UIST ’12 (New York, NY, USA: Association for Computing Machinery), 519–528. doi:10.1145/2380116.2380181

Frisoli, A., Solazzi, M., Reiner, M., and Bergamasco, M. (2011). The Contribution of Cutaneous and Kinesthetic Sensory Modalities in Haptic Perception of Orientation. *Brain Res. Bull.* 85, 260–266. doi:10.1016/j.brainresbull.2010.11.011

Genecov, A. M., Stanley, A. A., and Okamura, A. M. (2014). “Perception of a Haptic Jamming Display: Just Noticeable Differences in Stiffness and Geometry,” in 2014 IEEE Haptics Symposium (HAPTICS) (IEEE), 333–338. doi:10.1109/HAPTICS.2014.6775477

Gerling, G. J., Weissman, A. M., Thomas, G. W., and Dove, E. L. (2003). Effectiveness of a Dynamic Breast Examination Training Model to Improve Clinical Breast Examination (CBE) Skills. *Cancer Detect. Prev.* 27, 451–456. doi:10.1016/j.cdp.2003.09.008

Gijs, M. A. M., Lacharme, F., and Lehmann, U. (2010). Microfluidic Applications of Magnetic Particles for Biological Analysis and Catalysis. *Chem. Rev.* 110, 1518–1563. doi:10.1021/cr9001929

Harada, K., Nagata, K., Rojas, J., Ramirez-Alpizar, I. G., Wan, W., Onda, H., et al. (2016). Proposal of a Shape Adaptive Gripper for Robotic Assembly Tasks. *Adv. Robotics* 30, 1186–1198. doi:10.1080/01691864.2016.1209431

He, L., Herzig, N., Lusignan, S. d., Scimeca, L., Maiolino, P., Iida, F., et al. (2021). An Abdominal Phantom with Tunable Stiffness Nodules and Force Sensing Capability for Palpation Training. *IEEE Trans. Robot.* 37, 1051–1064. doi:10.1109/TRO.2020.3043717

Hook, J., Taylor, S., Butler, A., Villar, N., and Izadi, S. (2009). “A Reconfigurable Ferromagnetic Input Device,” in Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology UIST ’09 (New York, NY, USA: Association for Computing Machinery), 51–54. doi:10.1145/1622176.1622186

IBM SPSS Statistics 27 (2020). *Downloading IBM SPSS Statistics 27*. NY: IBM Corp..

Iwamoto, T., Tatezono, M., and Shinoda, H. (2008). “Non-contact Method for Producing Tactile Sensation Using Airborne Ultrasound,” in *Haptics: Perception, Devices and Scenarios Lecture Notes in Computer Science*. Editor M. Ferre (Berlin, Heidelberg: Springer), 504–513. doi:10.1007/978-3-540-69057-3_64

Jansen, Y., Karrer, T., and Borchers, J. (2010). “MudPad,” in ACM International Conference on Interactive Tabletops and Surfaces ITS ’10 (New York, NY, USA: Association for Computing Machinery), 11–14. doi:10.1145/1936652.1936655

Jiang, A., Xynogalas, G., Dasgupta, P., Althoefer, K., and Nanayakkara, T. (2012). “Design of a Variable Stiffness Flexible Manipulator with Composite Granular Jamming and Membrane Coupling,” in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IEEE), 2922–2927. doi:10.1109/iros.2012.6385696

Koehler, M., Usevitch, N. S., and Okamura, A. M. (2020). Model-Based Design of a Soft 3-D Haptic Shape Display. *IEEE Trans. Robot.* 36, 613–628. doi:10.1109/TRO.2020.2980114

Landis, J. R., and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 159–174. doi:10.2307/2529310

Lathan, C. E., Tracey, M. R., Sebrechts, M. M., Clawson, D. M., and Higgins, G. A. (2002). “Using Virtual Environments as Training Simulators: Measuring Transfer,” in *Handbook of Virtual Environments* (Boca Raton, FL: CRC Press).

Lederman, S. J., and Klatzky, R. L. (1993). Extracting Object Properties through Haptic Exploration. *Acta Psychologica* 84, 29–40. doi:10.1016/0001-6918(93)90070-8

Lelevé, A., McDaniel, T., and Rossa, C. (2020). Haptic Training Simulation. *Front. Virtual Real.* 1, 2. doi:10.3389/frvir.2020.00003

Li, M., Ranzani, T., Sareh, S., Seneviratne, L. D., Dasgupta, P., Wurdemann, H. A., et al. (2014). Multi-fingered Haptic Palpation Utilizing Granular Jamming Stiffness Feedback Actuators. *Smart Mater. Struct.* 23, 095007. doi:10.1088/0964-1726/23/9/095007

Licona, A. R., Liu, F., Pinzon, D., Torabi, A., Boulanger, P., Lelevé, A., et al. (2020). “Applications of Haptics in Medicine,” in *Haptic Interfaces for Accessibility, Health, and Enhanced Quality of Life*. Editors T. McDaniel and S. Panchanathan (Cham: Springer International Publishing), 183–214. doi:10.1007/978-3-030-34230-2_7

Manti, M., Cacucciolo, V., and Cianchetti, M. (2016). Stiffening in Soft Robotics: A Review of the State of the Art. *IEEE Robot. Automat. Mag.* 23, 93–106. doi:10.1109/MRA.2016.2582718

Putzu, F., Konstantinova, J., and Althoefer, K. (2019). “Soft Particles for Granular Jamming,” in Annual Conference towards Autonomous Robotic Systems (Springer), 65–74. doi:10.1007/978-3-030-25332-5_6

Rosenthal, R. (1986). “Meta-Analytic Procedures for Social Science Research,” in *Educational Researcher* (Beverly Hills: Sage Publications), 14818–14820. doi:10.3102/0013189x015008018

Srinivasan, M. A., and LaMotte, R. H. (1995). Tactual Discrimination of Softness. *J. Neurophysiol.* 73, 88–101. doi:10.1152/jn.1995.73.1.88

Stanley, A. A., Hata, K., and Okamura, A. M. (2016). “Closed-loop Shape Control of a Haptic Jamming Deformable Surface,” in 2016 IEEE International Conference on Robotics and Automation (ICRA) (IEEE), 2718–2724. doi:10.1109/icra.2016.7487433

Steltz, E., Mozeika, A., Rodenberg, N., Brown, E., and Jaeger, H. M. (2009). “JSEL: Jamming Skin Enabled Locomotion,” in 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IEEE), 5672–5677. doi:10.1109/IROS.2009.5354790

Tan, H. Z., Pang, X. D., and Durlach, N. I. (1992). Manual Resolution of Length, Force, and Compliance. *Adv. Robotics* 42, 13–18. doi:10.5802/aif.1307

Taylor, P. M., Moser, A., and Creed, A. (1998). A Sixty-Four Element Tactile Display Using Shape Memory alloy Wires. *Displays* 18, 163–168. doi:10.1016/S0141-9382(98)00017-1

Wagner, C. R., Lederman, S. J., and Howe, R. D. (2002). "A Tactile Shape Display Using RC Servomotors," in Proceedings 10th Symposium on Haptic Interfaces for Virtual Environment and Tele Operator Systems. HAPTICS 2002 (IEEE), 354–355.

Wickens, T. D., and Keppel, G. (2004). *Design and Analysis: A Researcher's Handbook*. Upper Saddle River, NJ: Pearson Prentice-Hall.

Woodrum, D. T., Andreatta, P. B., Yellamanchilli, R. K., Feryus, L., Gauger, P. G., and Minter, R. M. (2006). Construct Validity of the LapSim Laparoscopic Surgical Simulator. *Am. J. Surg.* 191, 28–32. doi:10.1016/j.amjsurg.2005.10.018

Zhao, X., Zhu, Z., Cong, Y., Zhao, Y., Zhang, Y., and Wang, D. (2020). Haptic Rendering of Diverse Tool-Tissue Contact Constraints during Dental Implantation Procedures. *Front. Robot. AI* 7, 1. doi:10.3389/frobt.2020.00035

Zhou, M., Tse, S., Derevianko, A., Jones, D. B., Schwaitzberg, S. D., and Cao, C. G. L. (2012). Effect of Haptic Feedback in Laparoscopic Surgery Skill Acquisition. *Surg. Endosc.* 26, 1128–1134. doi:10.1007/s00464-011-2011-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Appendix C4: Academic contribution 4

Dybvik, H., Abelson, F.G., Aalto, P., Goucher-Lambert, K., Steinert, M. (2023). Inspirational Stimuli Attain Visual Allocation: Examining Design Ideation with Eye-Tracking. In: Gero, J.S. (eds) Design Computing and Cognition'22. DCC 2022. Springer, Cham. https://doi.org/10.1007/978-3-031-20418-0_28

Inspirational Stimuli Attain Visual Allocation: Examining Design Ideation with Eye-Tracking

Henrikke Dybvik, Filip G. Abelson, Pasi Aalto, Kosa Goucher-Lambert, and Martin Steinert

Inspirational stimuli may be used to support the design process. This article aims to elicit new insights on the impact of inspirational stimuli on design ideation with eye-tracking technology. We replicated a design ideation experiment's methodology [1] but collected eye-tracking data and a think aloud protocol. Preliminary results of eye-tracking data demonstrate that inspirational stimuli influence participants' eye movements and visual allocation. Specifically, participants examine inspirational words significantly more than neutral words—and participants examine design problem statements significantly more in absence of inspirational stimuli. We also observe distinct individual visual search strategies. Experimental procedure, data, and code are openly available to facilitate further replication efforts.

Introduction

Visual stimuli affect designers during concept generation [2], facilitating [1] or hampering [2, 3] design ideation depending on the visual stimuli's type and timing. To support design processes, visual stimuli may e.g., serve as "inspiration", which is often sought by designers. If visual stimuli are "inspirational," how does it affect designers' visual allocation? Will designers devote more visual attention to inspirational stimuli?

This work uses eye-tracking technology to obtain further insight into visual stimuli's effect on idea generation by replicating the task, stimuli, and experimental procedure of a design ideation experiment [1] (referred to as "the original study"

H. Dybvik (✉) · F. G. Abelson · P. Aalto · M. Steinert
Norwegian University of Science and Technology, Trondheim, Norway
e-mail: henrikke.dybvik@ntnu.no

K. Goucher-Lambert
University of California, Berkeley, USA

J. S Gero (ed.), *Design Computing and Cognition'22*,
https://doi.org/10.1007/978-3-031-20418-0_28

throughout this article) that investigated whether inspirational stimuli of varying analogical distance to the problem space influence design concept generation.

The original study tasked participants to generate concepts for open-ended design problems and found that inspirational stimuli increased idea fluency [1]. Compared to control stimuli, inspirational stimuli both nearer and farther from the problem space facilitated participants to produce more ideas, and the effect was most prominent after a period of time. Moreover, functional magnetic resonance imaging (fMRI) data suggested two different activation patterns of brain regions. Two search strategies were coined accordingly: The *inspired internal search* activates brain regions associated with memory retrieval and semantic processing—herein, participants likely recognize the inspirational stimuli as helpful or applicable to the design problem. The *unsuccessful external search* increase activation in brain regions associated with directing attention outwards and visual processing—participants likely continue searching the problem space for an inkling. Control stimuli generally produce unsuccessful external search while near stimuli produce inspired internal search. Far stimuli exhibit features from both strategies.

While inspirational stimuli evoked different brain regions and facilitated participants to generate a greater number of ideas, it remains unknown whether these ideas were different from ideas produced in control conditions and which words were most conducive to the "inspired ideas." These unknowns constitute the research objectives of this work, warranting the collection of two new data sources. To learn whether ideas are different with and without inspirational aid, we added a think aloud protocol to the experimental protocol. To determine the most conducive (or "inspirational") words and investigate if they impact the visual allocation of participants or participants' perception, we used eye-tracking.

This article investigates designers' eye movements during ideation; it addresses the following questions; if visual stimuli are "inspirational," how does it affect visual allocation; will more visual attention be devoted to inspirational stimuli?

This article aims to provide new insights from eye-tracking; it presents a preliminary analysis of eye-tracking data, specifically, the differences in visual allocation between stimuli. The results show that participants examined inspirational words significantly more than control words; and that design problem statements were examined significantly more in the absence of inspirational stimuli. We moreover observe distinct individual visual search strategies. Exhaustive analysis of eye-tracking data and transcription of the think aloud recordings will be presented in future publications.

Background

Scientific advancement relies on the core principles of reliability, repeatability, and ultimately reproducibility, as do experiments. However, more often than one might think, experimental results are not replicated, or they fail to replicate due to a lack of information or other unforeseen factors. In the Open Science Collaboration's replication efforts, 32% of original results yielded insignificant results in combination with new data [4, 5]. This is a part of science's broader problems, further exacerbated by publication bias [5, 6].

Replication is a challenge to design cognition's future, a future that simultaneously holds repeated testing of predictions (or results) as a key opportunity [7]. These are closely related, and when taken together, they actualize replicability. Replication is both an opportunity to- and necessary for improving reproducibility, since irreproducible results may occur even within studies of exemplary quality due to random or systematic error [4].

We advocate for minimizing a potential replication crisis in design research and for providing replication studies with a positive connotation (our impression is that replication studies are frowned upon). Thus, this work replicates previous research's methodology [1] while gathering new data sources, and is thus a replication and extension study.

Eye-Tracking Technology

Eye-tracking measures record eye movements and gaze location over time and task [8]. The first record of eye-tracking dates back to 1823, and it was until recent technology advancements an expensive and effortful method. Today, eye-tracking is more affordable and accessible due to video-based eye-trackers [8, 9]. There are two main types of video-based eye-trackers: table and head-mounted configurations [8].

These eye-trackers shine an infrared light at the eye, not visible to humans, and illuminates it. Eye-facing cameras record the infrared light's reflection, which produces a corneal reflection and the pupil center through a bright or dark pupil effect [10]. The corneal reflection appears as a glint on the eye. When the infrared light is aligned with the camera's optical axis, the pupil's reflection is directed towards the camera producing the bright pupil effect. When the infrared light is not aligned with the camera's optical axis, the pupil's reflection is directed away from the camera, thus producing the dark pupil. The gaze position can be calculated by using the location of the corneal reflection and the pupil center.

Eye-Tracking Data

Eye-tracking data are time-series data sampled at a given frequency yielding the gaze position [8]. When the gaze rests (or fixates) at the same target for a period of time, these gaze points can be aggregated into a *fixation*. Fixations consist, therefore, of both a duration and gaze position. Fixation lengths vary and are usually within the range of 180–330 ms [11]. The rapid eye movements between fixations, occurring when scanning the visual space and moving the eyes, are called saccades. Herein, the visual input is suppressed [8, 11]. Some eye-trackers can also measure pupil dilation.

Eye-Tracking in Design Ideation Research

Design research uses eye-tracking to investigate visual reasoning in design activities [12]. Similar design ideation tasks with eye-tracking have explored differences between beginning and advanced design students during idea generation using stimuli of varying distance from the problem space, but used images as stimuli [13]. The Alternative Uses Test (AUT) has been used to investigate the relation between eye movements and idea output (creativity); participants were presented with images of 12 objects, and listed alternative uses of the object (i.e., ideated) for 2 min [14]. Eye-tracking and AUT have also been used to explore differences between designers and engineers in idea generation [15]. We have not found other studies investigating the effects of inspirational word stimuli during design ideation with eye-tracking.

Experimental Method

This experiment differed from the original [1] in using head-mounted eye-tracking technology and the think aloud protocol as an additional task. Here, participants were seated at a desk in front of a monitor, equipped with a conventional computer mouse and keyboard to indicate new ideas and submit questionnaire ratings. Participants in the original experiment lay supine in the fMRI, used a response glove to indicate ideas and provide questionnaire ratings, and did not think aloud.

Participants were tasked to develop as many ideas as possible for 12 different open-ended design problems and instructed to “think aloud” by briefly explaining their idea in a think aloud protocol. Five words were presented along each problem in two blocks for 1 min each, totaling 2 min of ideation time per design problem. The three first words were presented in the first block (called Wordset1), whereas the remanding two words were also presented in the second block (called Wordset2), i.e., the second block displayed all five words. A 1-back memory task was performed between blocks. Participants were exposed to three conditions: Near, Far, and Control. Words near or far from the problem space served as inspirational stimuli in the Near and Far conditions, while the Control condition reused words from the problem statement. See the original paper for an exhaustive description of the task, design problems, and word stimuli [1].

Participants were sequentially assigned to one of three counterbalanced groups of specific problem-condition pairs in the experiment’s repeated measures design.

After each problem, participants rated the words’ usefulness and relevancy, and the developed solutions’ novelty (uniqueness) and quality. The number and timing of generated ideas were collected continuously.

Participants

None of the N = 24 healthy adults (18 male/6 female, 22 right-handed/2 left-handed ages 23–35, mean = 25.8 yrs., SD = 2.9 yrs.,) participating were native English speakers. Since glasses might interfere with the head-mounted eye-tracker, no participant wore glasses, but eight used lenses. We recruited through internal channels and contacts at the Norwegian University of Science and Technology (NTNU). Participants were graduate-level students or higher (minimum 4th year MSc, PhDs) at the Department of Mechanical and Industrial Engineering (MTP) and the Department of Design (ID) to ensure similar educational background as original participants. Monetary compensation was not given.

Experiment Procedure and Calibration

First, participants received general information in Norwegian about the experiment, its procedure, and the task, and gave informed consent. Then, after fitting participants with the eye-tracker, it was 3D calibrated according to manufacturers' "Best Practices" [16]. Thereafter, the experiment commenced by providing information, explaining the design ideation task, and the 1-back again. Finally, participants answered a demographic survey after completing the 1-h experiment.

Hardware

The experiment ran on a conventional desktop computer with a 24 in. monitor, a conventional keyboard and mouse, and a head-mounted eye-tracker from Pupil Labs [17] with binocular setup (cameras on both eyes). See specifications below. Participants were seated in a chair approximately 70 cm from the monitor, see Fig. 1. A microphone was placed on a tripod in front of participants.

Higher accuracy in eye-tracking data may be acquired by using a chin rest. We were interested in areas, words, and patterns as a whole, which means sub-word accuracy was not necessary. We thought a chin rest might restrict participants and/or increase or induce a Hawthorne effect or other expectancy biases. A chin rest was therefore not used.

Hardware specifications:

- Desktop computer: Dell OptiPlex 7050, OS: Windows 10 Education 64-bit, CPU: Intel Core i7-7700 @ 3.60 GHz, RAM: 32 GB
- Monitor: Dell UltraSharp U2412M, Size: 24″ (61 cm), Resolution: 1920 × 1200 pixels, Refresh rate: 60 Hz
- Microphone: Zoom H1 Handy Recorder, fs: 48 kHz, Bit rate: 16 bit, Channels: 1 (mono recording)

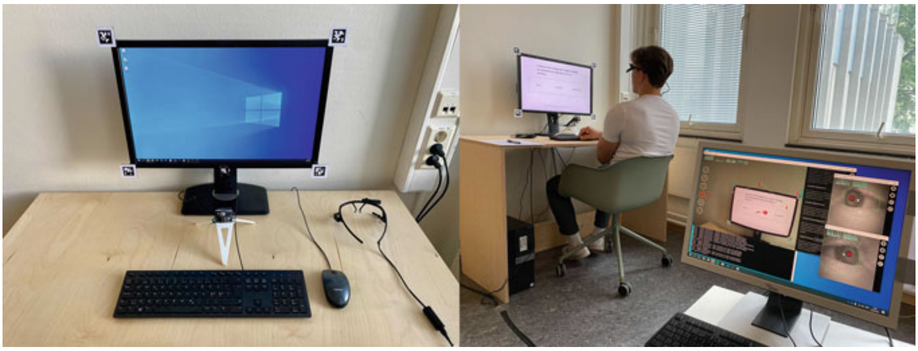


Fig. 1 Experimental setup

- Eye-tracker: Pupil Core, World cam. Resolution: 1280 × 720 pixels, fs: 30 Hz, Field of view: 99 degrees × 53 degrees, Eye cam. Resolution: 192 × 192 pixels, fs: 120 Hz. Gaze accuracy 0.6 degrees, gaze precision 0.02 degrees.

Software

The experiment was recreated in an open-source software, PsychoPy v2021.1.4 [18], with wordsets presented as black text on a white background in font OpenSans. Letter height was set to 5% of the screen's height in PsychoPy, which is 60 pixels on the monitor, or approximately 17 mm, which corresponds well with the fovea's 1.5–2-degree visual field at 70 cm viewing distance [19] and the eye-tracker's 0.6 degree accuracy. Pupil Capture collected and recorded eye-tracking data. To synchronize eye-tracking data, audio data, questionnaire responses, timestamped ideas, and stimuli annotations, we used Pupil Network API. By using this API, we sat Pupil's clock to the global experiment clock in PsychoPy, and thereby ensuring time synchronization of PsychoPy and Pupil Capture. This API was also used to implement automatic data recording, ensuring that Pupil Capture began recording once the PsychoPy experiment was launched. Pupil Player,[1] Pupil Labs' software, exported eye-tracking recordings from Pupil Capture.

Surface Tracking

To recorded participants' gaze relative to the monitor and not only the video frame we used Pupil's Surface Tracker plugin in combination with AprilTags (small binary markers) fastened on the monitor's bezel. We designed and 3D printed custom monitor mounts to ensure no changes in marker setup during the experiment period.

[1] https://docs.pupil-labs.com/core/software/pupil-player/raw-data-exporter.

The planar monitor's surface was mapped out with Pupil's Surface Tracker and the exact size of the monitor was marked in the recording software.

Processing and Analysis of Eye-Tracking Data

To summarize, the following data modalities were recorded: eye-tracking data, audio recordings, number and timing of generated ideas, and subjective ratings via a questionnaire. The two latter have been preliminary analyzed; these results largely corroborated the original and are presented in its entirety elsewhere [20]. This article's scope is a preliminary analysis of eye-tracking data. A comprehensive analysis of all data will be published later.

Data Processing

Exporting Eye-Tracking Data

Raw eye-tracking data were exported to CSV files with Pupil Player and stored in participant-specific folders.

Apart from fixations, all eye-tracking data export without selecting and setting any parameters. Duration and dispersion thresholds must be selected before exporting fixations, since fixations spread out temporally and spatially. Pupil Player uses a dispersion-based algorithm [21] that maximizes the fixation duration within the given parameters and outputs non-overlapping fixations. Pupil Capture calculates the gaze position using the dark pupil effect [17].

Our assumptions: The aim is to obtain an overview of which words were examined, not fixations within the words themselves. We recognized that participants could potentially fixate on a word for several seconds, i.e., a long fixation; we therefore wanted to prevent long fixations from being separated into a series of fixations. On the other hand, participants could also pay little attention to a word, e.g., recognize a control word, not find it interesting or helpful, and thus not spend any more time fixating on it. Such fixations have a short duration, but we want to capture them nevertheless.

The dispersion threshold was set to Pupil Player's maximum of 4.91 degrees. Pupil Labs states that there is no gold standard for setting fixation thresholds.[2] By exporting fixation data with different thresholds, we found that setting maximum duration too low caused a considerable number of fixations passing the threshold, which split long fixations into one or more shorter fixations. We found that we could capture fixations of a wide range of lengths by setting maximum duration to 4000 ms.

[2] See the following section in their "Best Practices": https://docs.pupil-labs.com/core/best-practices/fixation-filter-thresholds.

Although similar ideation research used a lower bound of 150 ms [22], we set the lower fixation bound to 100 ms (see e.g., Wass et al. [23]) to include potential short fixations.

Fixation files with other parameters can be exported since the raw data is publicly available.

Data Concatenation

To ease data handling, a script using the Pandas library [24, 25] iterated across exported files of the same type (e.g., annotations, gaze, fixations) and concatenated the files into one larger file per data type. Assigning participant ID to each row in the concatenated data ensured each row's uniqueness.

Data Analysis: Are Participants Paying More Attention to Inspirational Words?

This article aims, as mentioned, to provide a preliminary analysis of eye-tracking data. We sought "the bigger picture; an overview of which words were examined; and an investigation of whether there are differences in visual allocation for the different wordsets. We hypothesize the following; participants spend significantly more time examining inspirational words (i.e., words presented in Near and Far) than neutral words (i.e., compared to Control). We, therefore, evaluated eye-tracking metrics mostly on an aggregated level.

For data analysis and statistics, we used open-source Python libraries Pandas [24, 25], NumPy [26], SciPy [27], and Pingouin [28]—for visualization methods and plotting we used Seaborn [29] and Matplotlib [30].

Data Quality

The eye-tracking software Pupil Capture appends a confidence score between 0 and 1 for each data point based on the quality of the pupil detection. To ensure high-quality data, we included only data with a confidence score above 0.8, thus discarding data with low confidence scores, e.g., blinking.

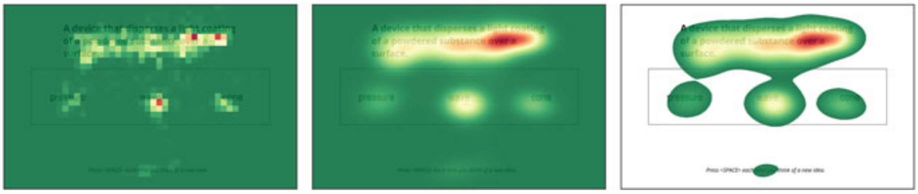


Fig. 2 Heatmap generation process from left to right: (Left) 2D histogram binned to the monitors' pixels (here visualized with large bins for visibility); (Center) Heatmap binned at each pixel with a Gaussian filter; (Right) Lower values filtered out of center heatmap

Heatmaps

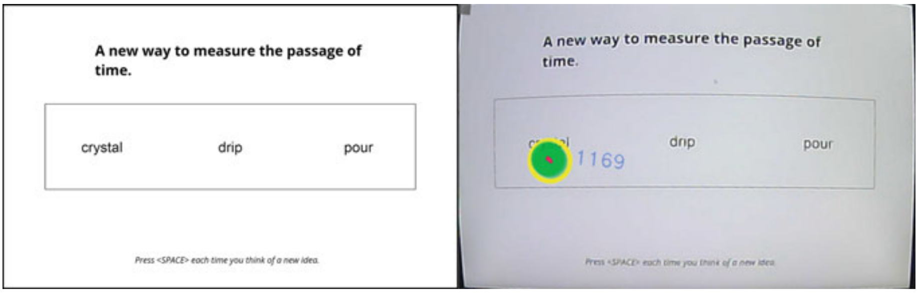
Heatmaps provide a visual overview of gaze positions. First, our custom code implementation made two-dimensional histograms with each gaze data point binned into bins similar to the pixels of the monitor (1920 × 1200 pixels)—then, we smoothened the values with a Gaussian filter. Afterward, to increase visual differences between heatmaps, we filtered out values below a lower bound. The lower bound was the mean of the histograms' non-zero values, divided by 2. Figure 2 illustrates the heatmap generation process. The color indicates the relative gaze distribution from green to red with increasing gaze density.

Fixation Distribution

To obtain descriptive data of fixation distribution to test differences between stimuli, we split the monitor into four areas of interest (AOI) and calculated the ratio of time participants spent examining each area. The AOIs were: *problem*, *words*, *off-screen* and *other*. *Problem* represents design problem statements, and *words* represent wordsets; these AOIs indicate how interesting the words are and how one might draw inspiration from the problem statement itself; they were selected to investigate differences between words and problems. We included *off-screen* since we observed some participants gazing outside the monitor when ideating in the experiment. *Other* represents the remaining parts of the monitor.

The Surface Tracker plugin does not map the monitor perfectly (as seen in Fig. 3) due to distortion from the world camera's fisheye lens. Upon preliminary inspection of heatmaps, we noticed a slight vertical offset relative to the text on the screen. We, therefore, extended the boundaries for the box encompassing the words, particularly for wordset 1, see Fig. 4. This distortion explains offsets when plotting heatmaps and scanpaths over a screenshot.

Fixation distribution data was made by first assigning a label with corresponding AOI to each fixation based on the fixation's position on the monitor. Second, AOI distribution ratio was calculated by summing up fixation duration for each label and



A new way to measure the passage of time.

crystal drip pour

Press <SPACE> each time you think of a new idea.

Fig. 3 (Left) Frame as shown on monitor. (Right) Monitor as mapped out by Pupil software

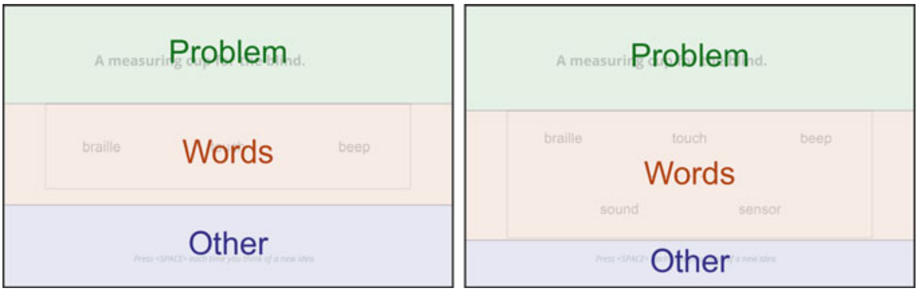


Fig. 4 Areas of interest (AOI) borders defined within the monitor for Wordset1 (left) and Wordset2 (right). Off-screen is outside the monitor

wordset (totaling 100%), and dividing by each label and wordset's fixation duration sum, respectively.

Statistical Analysis of Fixation Distribution Data

Friedman's test (a non-parametric test) assessed differences in fixation distribution between conditions due to violation of ANOVA's normality assumption for several subsets. Wilcoxon signed-rank test with a Bonferroni correction for multiple comparisons assessed pairwise comparisons post hoc. Hedges' g is used as effect size [6]. A significance level of $p < 0.05$ was selected for all tests.

Scanpaths

Scanpaths visualize fixation data in a scatterplot where the dots are connected by lines. A dot indicates a fixation; the dot's size varies with the fixation's duration, i.e., larger dots indicate longer fixations. The lines connecting the fixations indicate saccades. The first and last fixation is indicated by a green and cyan point, respectively. This custom scanpath implementation plots the line between each fixation with chronologically varying opacity from transparent to opaque, meaning that the visualization retains the temporality in the eye-tracking data, whereas heatmaps only aggregates the position of gaze data.

Due to the temporal aspect of scanpaths, they are difficult to compare directly on an aggregated basis, which is possible with heatmaps.

Results

Heatmaps

Aggregated heatmaps for all combinations of conditions and wordsets are presented in Fig. 5. Firstly, we observe a strong tendency of gaze allocation towards the monitor's center for all conditions, which we attribute to the *central fixation bias* [31]; the "marked tendency to fixate the center of the screen when viewing scenes on computer monitors." In other words, the monitor's center is a natural place to rest the gaze when not actively scanning for new visual input.

Despite the *central fixation bias*, there is a clear visual difference between the inspirational stimuli and control words in the time participants spent looking at the different AOIs, as seen in Fig. 5. Gaze is allocated more to the problem statement in control conditions, whereas the gaze distributes more evenly over the entire monitor and more on the wordsets in Near and Far condition.

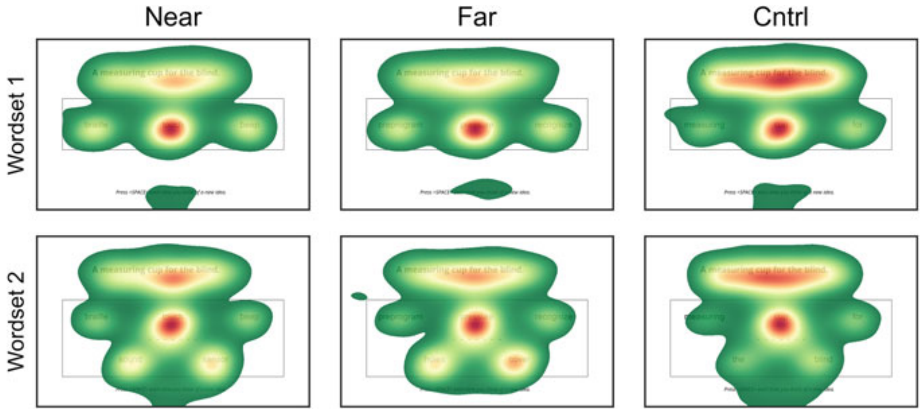


Fig. 5 Aggregated heatmap across all conditions

Table 1 Friedman test of AOI ratio

Wordset	AOI	DOF	χ^2	p
1	Problem	2	7.583	0.023*
	Words	2	18.250	< 0.001**
	Off-screen	2	4.750	0.093
	Other	2	2.333	0.311
2	Problem	2	7.583	0.023*
	Words	2	14.333	0.001**
	Off-screen	2	2.333	0.311
	Other	2	0.083	0.959

* $p < 0.05$, ** $p < 0.01$

Fixation Distribution

Friedmans test evaluated differences in fixation distribution between conditions and wordsets, i.e., objectively testing whether participants spent more or less time in any AOI. Table 1 presents the results, which were significant for AOI *words* and *problem* for both wordsets.

Pairwise comparisons with Wilcoxon tests are presented in Table 2. For AOI *problem,* there were significant differences between Control and Near for Wordset1 and between Control and Far for Wordset2. Moreover, the difference between Control and Near for Wordset2 obtained a $p = 0.051$, close to the significance threshold and thus noteworthy. Participants spent more time examining problem statements in control conditions compared to inspirational conditions.

For AOI *words*, there were significant differences between Control and Near, and Control and Far for both Wordset1 and Wordset2. Participants spent more time examining the inspirational words than control words. These findings align with our heatmap observations: when receiving control stimuli, participants spend time on the problem and less time on the words, compared to receiving inspirational stimuli, in which case participants spend more time on the words and less on the problem. This effect is apparent in Fig. 6 as well.

Discussion

The effect of inspirational words on participants visual allocation was significant. Inspirational words both near and far from the problem space received greater visual attention, i.e., participants spent more time visually fixating on the inspirational words, compared control words, throughout the entire ideation session. Further, participants visually examined the problem statement significantly more in control ideation sessions' second halves (Wordset2) compared to far inspirational ideation,

Table 2 Wilcoxon signed-rank test with Bonferroni correction

WS	AOI	Between		W	p	Corr. p	Hedges' g
1	Problem	Control	Far	79.00	0.044	0.132	0.378
		Control	Near	45.00	0.003	0.008*	0.482
		Far	Near	138.00	0.742	1.000	0.117
	Words	Control	Far	35.00	0.001	0.003**	−0.608
		Control	Near	19.00	<0.001	0.001**	−0.620
		Far	Near	143.00	0.853	1.000	−0.001
	Off-screen	Control	Far	75.00	0.033	0.100	0.311
		Control	Near	92.00	0.100	0.301	0.178
		Far	Near	111.00	0.271	0.814	−0.146
2	Problem	Control	Far	47.00	0.003	0.010*	0.592
		Control	Near	66.00	0.017	0.051^{n}	0.570
		Far	Near	147.00	0.943	1.000	0.000
	Words	Control	Far	22.00	0.000	0.001*	−0.695
		Control	Near	56.00	0.008	0.023*	−0.510
		Far	Near	111.00	0.271	0.814	0.149
	Off-screen	Control	Far	107.00	0.225	0.674	0.256
		Control	Near	148.00	0.966	1.000	0.092
		Far	Near	107.00	0.225	0.674	−0.188

* $p < 0.05$, ** $p < 0.01$, ^{n} noteworthy, AOI *Other* is not included since its Friedman test was insignificant

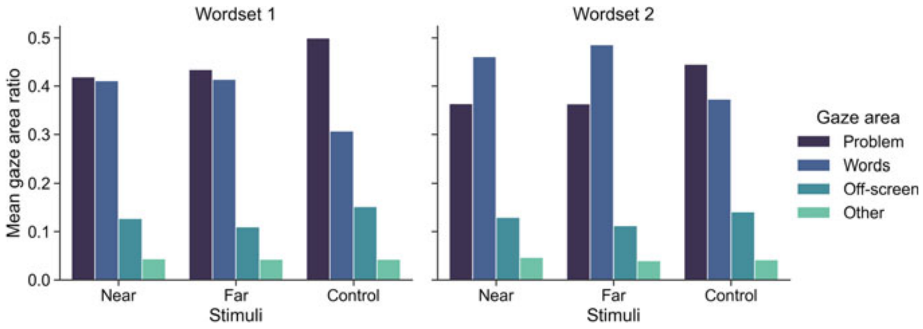


Fig. 6 Fixation distribution of AOIs for all conditions

and significantly more in control ideation sessions' first halves (Wordset1) compared to near inspirational ideation. The difference between control and near inspirational ideation sessions' second half (Wordset2) obtained a $p = 0.051$, close to the significance threshold. It may have turned out significant with a larger or slightly different participant pool. Because its effect size ($g = 0.570$) is comparable to that of the

Control-Far (Wordset2) ($g = 0.592$), we take this as an indication of the effect also occurring in the second half.

To summarize, this preliminary analysis of eye-tracking data yield two main findings/conclusions. One, participants allocate more visual attention (time) to word stimuli when receiving inspirational stimuli of any kind (both near and far from the problem space) compared to neutral (control) words throughout the entire ideation session; two, in the absence of inspirational stimuli (control condition) participants devote more visual attention to problem statements, an effect whose magnitude might depend on the inspirational words' distance to the problem space.

Finding one may be related to the *inspired internal search* from the original study, which suggested that participants found/recognized the inspirational stimuli as helpful or applicable to the design problem. Participants did rate inspirational stimuli as more useful both in the original study [1] as well as in this replication [20], which we suppose is why participants spent more time examining inspirational words than neutral words.

The second finding can be related to the strategy *unsuccessful external search* employed in the absence of inspirational stimuli where, originally, it is suggested that participants continue to search for clues in the design problem space [1]. The eye-tracking data confirm that participants continue trying to use the problem statement as a source of inspiration when they are not provided with any inspirational stimuli.

Scanpaths

The scanpaths presented here visualize a difference in how participants move their gaze around, possibly using different strategies during ideation. We selected an example illustrating participant 3 (in Fig. 7) versus participant 6 (in Fig. 8) for all problems in Wordset2 (both from group C).[3] Participant 3 stays fairly central at all times, exhibiting the central fixation bias to a greater extent than participant 6, who moves vigorously around the visual space, looking for inspiration in almost every stimuli word, to us in a pattern strikingly similar to a hexagon. Although both participants show lacking interest in control stimuli words for problem 10, it appears that individual participants have different search strategies; this will be investigated further in future work. Further conclusions regarding search strategies are therefore not drawn here.

[3] While only presenting a selection in this article, scanpaths were generated for all problems and participants.

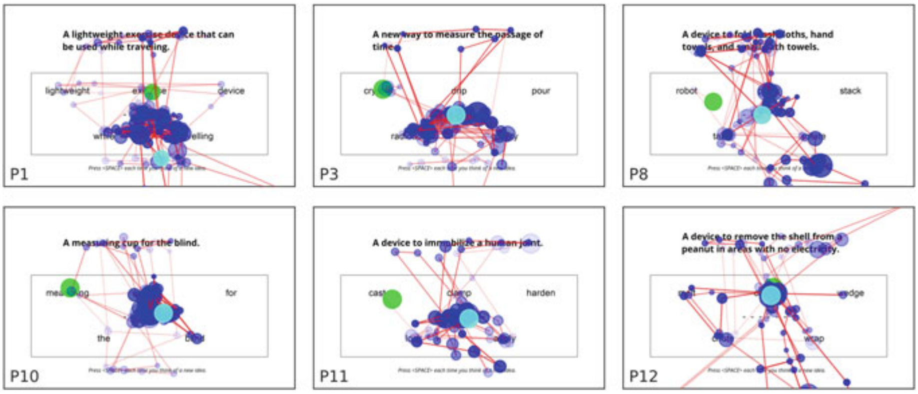


Fig. 7 Scanpaths for participant 3 for selected problems in Wordset2

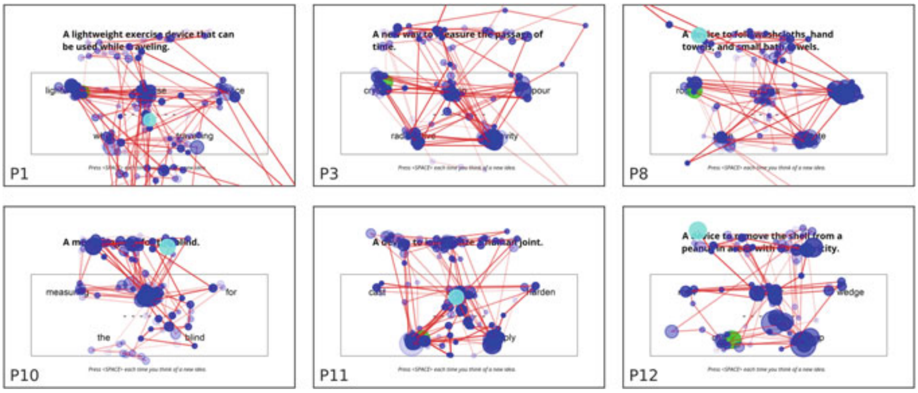


Fig. 8 Scanpaths for participant 6 for selected problems in Wordset2

Limitations

The study is limited by the central fixation bias, which could have been corrected for by randomizing the visual stimuli's position (problem statement and words etc.) on the monitor. However, randomization of positions was not possible since the study employed existing stimuli. Therefore, results are presumably influenced by the central fixation bias with *words* receiving disproportionately greater visual attention than other AOIs, as seen in Fig. 5. If we assume that the central fixation bias says consistent across conditions it is not affecting statistical results; if, however, it varies from condition to condition the statistical results are influenced.

Since this paper presents a preliminary analysis of eye-tracking data only the research objectives are not exhaustively answered; for this an exhaustive analysis is necessary. Further studies—collecting new and additional data modalities, in settings with higher ecological validity or in situ—are necessary to fully understand design ideation, visual- and inspirational stimuli.

Future Work

Future work intends to present an exhaustive joint analysis of eye-tracking data, the think aloud protocol's transcription, the behavioral-, and subjective data measured.

The illustrated scanpaths appear to indicate that there may exist individual visual search strategies amongst participants, although we do not draw any conclusions here. Scanpaths are interesting descriptive data that we will use to inform future analysis. Currently, we do not know how nor if scanpaths will be useful or not, but this will be investigated in future work.

Conclusion

We used eye-tracking and a think aloud protocol in a replication and extension of a design ideation experiment with and without inspirational stimuli [1]. This article provided a preliminary analysis of eye-tracking data and aimed to provide new insights from eye-tracking technology. Results show clear influence from inspirational stimuli on visual allocation; participants examine (or gaze) significantly more on inspirational words than neutral words; in inspirational stimuli's absence, participants examine design problem statements significantly more. Finally, we facilitate further replication with the openly available experimental procedure, data, and code.

Published Code Repository and Data

The code, raw data, and results from this study are publicly available:

- Code repository [32]: https://doi.org/10.5281/zenodo.5130090
- Pre-processed data [33]: https://doi.org/10.18710/PZQC4A
- Raw eye-tracking data [34]: https://doi.org/10.21400/7kq02wjl.

References

1. Goucher-Lambert K, Moss J, Cagan J (2019) A neuroimaging investigation of design ideation with and without inspirational stimuli—understanding the meaning of near and far stimuli. Des Stud 60:1–38. https://doi.org/10.1016/j.destud.2018.07.001

2. Tseng I, Moss J, Cagan J, Kotovsky K (2008) The role of timing and analogical similarity in the stimulation of idea generation in design. Des Stud 29:203–221. https://doi.org/10.1016/j.destud.2008.01.003
3. Jansson DG, Smith SM (1991) Design fixation. Des Stud 12:3–11. https://doi.org/10.1016/0142-694X(91)90003-F
4. Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349. https://doi.org/10.1126/science.aac4716
5. Shrout PE, Rodgers JL (2018) Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. Annu Rev Psychol 69:487–510. https://doi.org/10.1146/annurev-psych-122216-011845
6. Field A (2018) Discovering statistics using IBM SPSS statistics, 5th edn. SAGE Publications, Thousand Oaks, CA
7. Hay L, Cash P, McKilligan S (2020) The future of design cognition analysis. Des Sci 6. https://doi.org/10.1017/dsj.2020.20
8. Carter BT, Luke SG (2020) Best practices in eye tracking research. Int J Psychophysiol 155:49–62. https://doi.org/10.1016/j.ijpsycho.2020.05.010
9. Wade P of VPN, Wade N, Tatler BW, Tatler L in PB (2005) The Moving tablet of the eye: the origins of modern eye movement research. Oxford University Press
10. Duchowski AT (2017) Eye tracking methodology, 3rd edn. Springer International Publishing, Cham
11. Rayner K (2009) Eye movements and attention in reading, scene perception, and visual search. Q J Experiment Psychol 62:1457–1506. https://doi.org/10.1080/17470210902816461
12. Gero JS, Milovanovic J (2020) A framework for studying design thinking through measuring designers' minds, bodies and brains. Des Sci 6. https://doi.org/10.1017/dsj.2020.15
13. Cao J, Xiong Y, Li Y, Liu L, Wang M (2018) Differences between beginning and advanced design students in analogical reasoning during idea generation: evidence from eye movements. Cogn Tech Work 20:505–520. https://doi.org/10.1007/s10111-018-0477-z
14. Kwon E, Ryan JD, Bazylak A, Shu LH (2019) Does visual fixation affect idea fixation? J Mech Des 142. https://doi.org/10.1115/1.4045600
15. Colombo S, Mazza A, Montagna F, Ricci R, Monte OD, Cantamessa M (2020) Neurophysiological evidence in idea generation: differences between designers and engineers. In: Proceedings of the design society: DESIGN conference, vol 1, pp 1415–1424. https://doi.org/10.1017/dsd.2020.161
16. Pupil Labs (2021) best practices—tips for conducting eye tracking experiments with the Pupil Core eye tracking platform. In: Pupil Labs. https://docs.pupil-labs.com. Accessed 27 Apr 2021
17. Kassner M, Patera W, Bulling A (2014) Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: adjunct publication. Association for computing machinery, New York, NY, USA, pp 1151–1160
18. Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK (2019) PsychoPy2: experiments in behavior made easy. Behav Res 51:195–203. https://doi.org/10.3758/s13428-018-01193-y
19. Holmqvist K (2011) Eye tracking: a comprehensive guide to methods and measures. Oxford University Press, Oxford, New York
20. Dybvik H, Abelson F, Aalto P, Goucher-Lambert K, Steinert M (2022) Inspirational stimuli improve idea fluency during ideation: A replication and extension study with eye-tracking. Proc Des Soc 2:861–870. https://doi.org/10.1017/pds.2022.88
21. Salvucci D, Goldberg J (2000) Identifying fixations and saccades in eye-tracking protocols
22. Vendetti MS, Starr A, Johnson EL, Modavi K, Bunge SA (2017) Eye movements reveal optimal strategies for analogical reasoning. Front Psychol 8. https://doi.org/10.3389/fpsyg.2017.00932
23. Wass SV, Smith TJ, Johnson MH (2013) Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. Behav Res 45:229–250. https://doi.org/10.3758/s13428-012-0245-6

24. McKinney W (2010) Data structures for statistical computing in Python. Austin, Texas, pp 56–61
25. Reback J, Jbrockmendel, McKinney W, Van Den Bossche J, Augspurger T, Cloud P, Hawkins S, Gfyoung, Sinhrks, Roeschke M, Klein A, Terji Petersen, Tratner J, She C, Ayd W, Hoefler P, Naveh S, Garcia M, Schendel J, Hayden A, Saxton D, Gorelli ME, Shadrach R, Jancauskas V, McMaster A, Fangchen Li, Battiston P, Skipper Seabold, Attack68, Kaiqi Dong (2021) pandas-dev/pandas: Pandas 1.3.0. Zenodo
26. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE (2020) Array programming with NumPy. Nature 585:357–362. https://doi.org/10.1038/s41586-020-2649-2
27. Virtanen P, Gommers R, SciPy 1.0 Contributors, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2
28. Vallat R (2018) Pingouin: statistics in Python. J Open Sour Softw 3:1026. https://doi.org/10.21105/joss.01026
29. Waskom M (2021) Seaborn: statistical data visualization. JOSS 6:3021. https://doi.org/10.21105/joss.03021
30. Hunter JD (2007) Matplotlib: a 2d graphics environment. Comput Sci Eng 9:90–95. https://doi.org/10.1109/MCSE.2007.55
31. Tatler BW (2007) The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. J Vis 7:4–4. https://doi.org/10.1167/7.14.4
32. Abelson FG (2021) Code repository for design ideation experiment (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5130090
33. Abelson FG, Dybvik H, Steinert M (2021) Dataset for design ideation study. DataverseNO. https://doi.org/10.18710/PZQC4A
34. Abelson FG, Dybvik H, Steinert M (2021) Raw data for design ideation study. https://doi.org/10.21400/7KQ02WJL

C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16

Appendix C5: Academic contribution 5

Dybvik, H., Abelson, F. G., Aalto, P., Goucher-Lambert, K., & Steinert, M. (2022). Inspirational Stimuli Improve Idea Fluency during Ideation: A Replication and Extension Study with Eye-Tracking. Proceedings of the Design Society, 2, 861–870. https://doi.org/10.1017/pds.2022.88

INTERNATIONAL DESIGN CONFERENCE – DESIGN 2022
https://doi.org/10.1017/pds.2022.88

Inspirational Stimuli Improve Idea Fluency during Ideation: A Replication and Extension Study with Eye-Tracking

H. Dybvik [1,✉], F. G. Abelson [1], P. Aalto [1], K. Goucher-Lambert [2] and M. Steinert [1]

[1] Norwegian University of Science and Technology, Norway,
[2] University of California, Berkeley, United States of America

✉ henrikke.dybvik@ntnu.no

Abstract

We replicate a design ideation experiment (Goucher-Lambert et al., 2019) with and without inspirational stimuli and extend data collection sources to eye-tracking and a think aloud protocol to provide new insights into generated ideas. Preliminary results corroborate original findings: inspirational stimuli have an effect on idea output and questionnaire ratings. Near and far inspirational stimuli increased participants' idea fluency over time and were rated more useful than control. We further enable experiment reproducibility and provide publicly available data.

Keywords: design cognition, conceptual design, eye tracking, experimentation, replication

1. Introduction

An important part of science and experiments is reliability, repeatability (or replicability), and reproducibility—however, experiments are not replicated as often as they ought to be or fail to replicate for numerous reasons. Open Science Collaboration's recent effort to conduct 100 replications of systematically sampled psychology results in top-tier journals, produced significant results in 36% of the replication studies, which, compared to 97% significant results in original studies (Open Science Collaboration, 2015), we find shocking. Moreover, 32% of original results were not significant when combined with new data (Open Science Collaboration, 2015; Shrout and Rodgers, 2018). The reproducibility crisis, across scientific fields, is exacerbated by a publication bias towards statistically significant results and reluctance to publish replication studies (Field, 2018; Martin and Clarke, 2017; Shrout and Rodgers, 2018). Even studies of exemplary quality may have irreproducible results due to random or systematic error, replication is therefore not only an opportunity to improve reproducibility—it is necessary (Open Science Collaboration, 2015).

We want to minimize potential replication issues in design research and advocate for providing replication efforts with a positive connotation. To this end, we have replicated the focus and activity (i.e., experimental design and stimuli) of a design ideation study that used neuroscience methods and means of data collection (Goucher-Lambert et al., 2019), and extended the study by changing and adding new sources of data collection. The design ideation study, referred to as "the original study" throughout this article, conducted an experiment were participants lay supine in an fMRI and were tasked to generate ideas for 12 design problems assisted by word stimuli that were inspirational, either near- or far from the solution space, or that served as a control (Goucher-Lambert et al., 2019). The original study explored the impact of inspirational stimuli on design ideation, behavioral-, and neurological processes, and demonstrated that inspirational stimuli both near and far from the problem space increase idea fluency compared to control stimuli. Inspirational stimuli was most beneficial after

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

some time, it enabled participants with a higher idea output over time. Inspirational stimuli had a significant effect on subjective ratings of relevancy and usefulness of the stimuli, but not on quality and novelty of the ideas. fMRI data suggested two search strategies: In the positive strategy—the inspired internal search—participants recognize the inspirational stimuli as applicable to the design problem, and it activates brain regions associated with memory retrieval and semantic processing. The negative strategy—unsuccessful external search—participants continue searching the problem space for an inkling, and it increases activation in brain regions associated with directing attention outwards and visual processing. Control stimuli is consistent with the negative strategy, while near stimuli triggers the positive strategy. Far stimuli, depending on the actual distance from the problem space, exhibits features from both strategies.

The replication and extension study presented in this article originated from us possessing eye-tracking technology while simultaneously contemplating the nature of the original study's stimuli; specifically, when observing the different words within each stimuli we wondered whether any of them were more or less "inspirational" and whether participants paid more attention to them. Eye-tracking may provide insights into visual allocation across various stimuli, revealing potentially subconscious behavior during ideation. The present study was solely driven by a desire to investigate these questions with eye-tracking technology; it assumes the fMRI results' validity since it does not have access to an fMRI for verification purposes. Further, since original participants ideated silently there is no record of which ideas were produced, thus, it is unknown whether the resulting ideas were different; we wanted to address this as well, by adding a think aloud protocol. The eye-tracking and think aloud combination is particularly interesting to us since it could reveal exactly which words were used when producing specific ideas. An fMRI environment is restrictive, does not allow for a think aloud protocol without affecting data quality, and share few commonalities with practitioners' ideation. By moving the experimental design and task from the fMRI context to a conventional office-with-desk context we obtain a more realistic experimental context, and it becomes interesting to investigate if the number of generated ideas and participants subjective ratings hold true across contexts. We replicate the original experimental design and stimuli, change context from an fMRI to a conventional desk, and extend by adding eye-tracking and think aloud protocol as means of data collection.

This article describes the experiment briefly, including replicated- and new content, and adaptations required to include the new data collection sources properly. We further present preliminary results from analyzing the number of generated ideas and participants' subjective ratings of ideas, which largely corroborate original results. Eye-tracking data and recording of the think aloud protocol is currently under analysis and will be presented in future publications.

2. Background

2.1. Similar research

Eye-tracking as a research tool has gained popularity across several research fields the past 20 years (Carter and Luke, 2020). In design research eye-tracking is listed among the tools for studying design physiology and that it “gives insight into visual reasoning during a design task” (Gero and Milovanovic, 2020). A substantial amount of existing research uses eye-tracking as a tool to explore engineering and product design using image stimuli. A search for similar eye-tracking research was performed by querying "design ideation" AND "eye-tracking" in Google Scholar, and querying "eye-tracking" amongst the 48 citations of Goucher-Lambert et al. (2019). An extensive review of existing design research using neurophysiological and biometric measures, thus including eye-tracking, was also used to search for similar work (Borgianni and Maccioni, 2020). This search did not find any study using eye-tracking to explore the effects of inspirational word stimuli in design ideation, but examples of eye-tracking for other or similar design ideation tasks were found. Cao et al. (2018) also uses stimuli of varying distance from the problem space to explore difference between beginning and advanced design students during idea generation, but uses images as stimuli. Kwon et al. (2019) looked into the relation of eye movements and idea output (creativity) to an “alternative uses test (AUT)”, where participants are presented with 12 object images and get 2 minutes per object to name alternative uses of the object.

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

This study shares similarities with the idea generation study by Colombo et al. (2020) in which AUT and eye-tracking explore the differences between designers and engineers.

2.2. Eye-tracking technology

Eye-tracking is a measure of eye movements, and thus gaze location over time (Carter and Luke, 2020). Recording of eye movements dates back to 1823, and the technology have seen vast improvements in recent years using video-based eye-trackers making it more affordable and accessible (Carter and Luke, 2020; Wade et al., 2005). There are two main types of video-based eye-trackers: table and head-mounted configurations (Carter and Luke, 2020). Head-mounted eye-trackers work by shining infrared light at the eye, and illuminating it without being visible to humans, resulting in a corneal reflection and bright pupil effect (Duchowski, 2017). The corneal reflection appears as a glint on the eye, and the bright pupil effect are both caused by the reflection of the infrared light and are recorded using eye-facing cameras. By using the location of the corneal reflection and the pupil center, software can calculate the gaze position after device calibration. Eye-tracking data are time series sampled at a given frequency yielding the gaze position (Carter and Luke, 2020). When the eyes fixate on a target over a period of time the gaze points can be aggregated into a fixation. Fixation length vary and are usually within the range of 180-330 milliseconds (Rayner, 2009). The rapid eye movements between fixations, happening while scanning the visual space and moving the eyes, are called saccades and during these the visual input is suppressed (Carter and Luke, 2020; Rayner, 2009).

3. Method

3.1. Experimental design and setup

3.1.1. Ideation task within a repeated measures experimental design

The task, and thus problems and words used in this experiment are the exact same as in Goucher-Lambert et al. (2019). Participants were tasked to develop as many ideas as possible for 12 different design problems and instructed to "thinking aloud" by briefly explaining their idea in a think aloud protocol. Each new idea was indicated by pressing the space bar. Five words were presented along with each problem. Reused words from the problem statement was presented in the Control condition. Words near or far from the problem space was used as inspirational stimuli in the Near and Far condition respectively. The 2 minutes ideation time per problem was divided into two blocks of 1 minute. The first block called Wordset1 displayed the three first words. The second block called Wordset2 displayed all five words. The 1-back memory task was performed between blocks. Participants completed a questionnaire—rating the usefulness and relevancy of the words presented, and the novelty (uniqueness) and quality of the solutions developed on a scale from 1 to 5—after each problem. The experiment follows a repeated measures design assigning participants to one of three counterbalanced groups of specific problem-condition pairs. The full experimental procedure with routines' timing is visualized in Figure 1. The fMRI-specific fixation cross routine, indicated by "+", was kept to retain the original study's temporality, and had a random duration between 0.5 and 4 seconds.

3.1.2. Differences from original study

The main difference between the original study and this study was the use of eye-tracking technology. Originally, participants lay supine in an fMRI viewing stimuli on a monitor through a look out mirror attached to the head mounted coil. By using a response glove strapped to their right hand, participants could indicate new ideas with their index finger and provide questionnaire ratings with all five fingers. In this experiment participants sat in a chair in front of a monitor, equipped with a conventional computer mouse and keyboard to indicate new ideas and submit questionnaire ratings.

Participants were additionally tasked to think aloud which may impact the number of ideas produced as it may require more time to articulate an idea compared to only thinking of it.

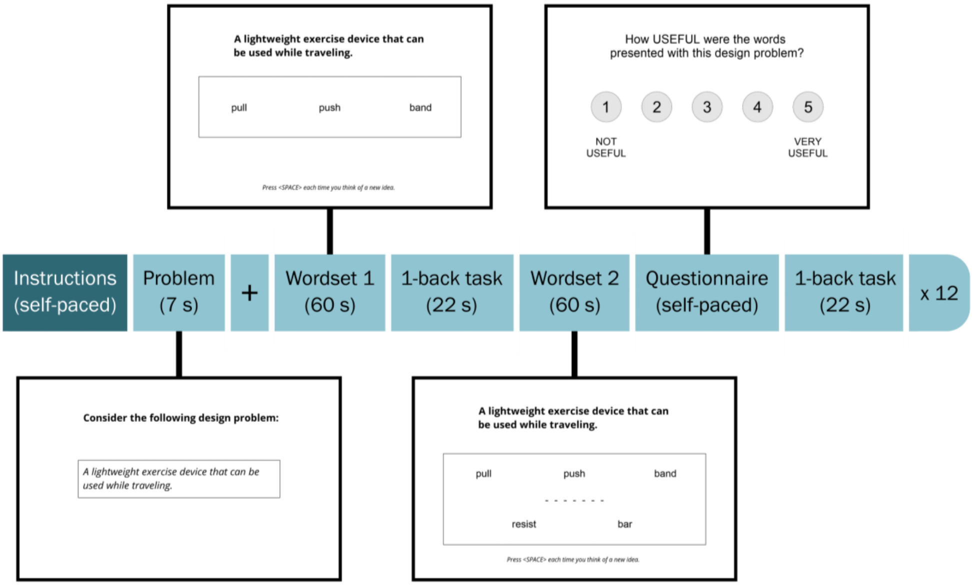
https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

3.2. Hardware

The experiment was run on a conventional desktop computer with a 24 inch monitor along with the head-mounted eye-tracker from Pupil Labs (Kassner et al., 2014) with binocular setup (cameras on both eyes). Participants were placed in a chair approximately 70 cm from the monitor, see Figure 2. We weren't interested in sub-word accuracy, but rather areas, words, and patterns as a whole meaning that a higher accuracy obtained by a chin rest wasn't necessary. We believed a chin rest would feel restricting for participants during ideation, perhaps also increasing a Hawthorne effect or other expectancy biases, and thus chose to not use one. A USB-connected microphone was placed on a tripod in front of the participant to obtain high quality audio recordings. A conventional keyboard and mouse were used. Additional hardware specifications are listed below.

Hardware specifications:

- Desktop computer: Dell OptiPlex 7050, OS: Windows 10 Education 64-bit, CPU: Intel Core i7-7700 @ 3.60GHz, RAM: 32 GB
- Monitor: Dell UltraSharp U2412M, Size: 24” (61 cm), Resolution: 1920x1200 pixels, Refresh rate: 60 Hz
- Microphone: Zoom H1 Handy Recorder, f_s: 48 kHz, Bit rate: 16 bit, Channels: 1 (mono recording)
- Eye-tracker: Pupil Core, World cam. Resolution: 1280x720 pixels, fs: 30 Hz, Field of view: 99 degrees x 53 degrees, Eye cam. Resolution: 192x192 pixels, f_s: 120 Hz. Gaze Accuracy: 0.6 degrees, gaze precision: 0.02 degrees.



A lightweight exercise device that can be used while traveling.

Press <SPACE> each time you think of a new idea.

How USEFUL were the words presented with this design problem?

Consider the following design problem:

A lightweight exercise device that can be used while traveling.

A lightweight exercise device that can be used while traveling.

Press <SPACE> each time you think of a new idea.

Figure 1. Experiment procedure. "+" indicates fixation cross. Instructions were shown once.

3.3. Software implementation

The experiment was programmed in the open-source software PsychoPy v2021.1.4 (Peirce et al., 2019) in contrary proprietary software E-Prime used originally. PsychoPy offers a graphical user interface and allows for running custom Python code. Most input data and visual design was retrieved from the published article. Some additional figures and information were obtained from original authors. All word stimuli were presented as black text on a white background in OpenSans font. Letter height were set to 5 percent of the screen height in PsychoPy which translates to 60 pixels on the monitor.

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

3.3.1. Eye-tracking data collection, annotations and time synchronization

Eye-tracking data was collected with Pupil Capture. Pupil Network API was used to synchronize eye-tracking data, audio data, questionnaire responses, timestamped ideas, and stimuli annotations. The Network API control time synchronization of PsychoPy and Pupil Capture by sending a message over the API setting Pupil's clock to the global experiment clock in PsychoPy. Automatic data recording was implemented using the API, ensuring that Pupil Capture began recording once the PsychoPy experiment was launched.

3.3.2. Audio recording

Automatic sound recording of participants thinking aloud was implemented by using high-level functionality from module python-sounddevice to record, and saving functionality in WAV format from SciPy. Each recording was automatically saved with a filename with participant ID, problem ID and stimuli conditions.

3.3.3. Surface tracking

We used Pupil's Surface Tracker plugin to record the gaze of the participants relative to the monitor, not only the video frame. By fixing AprilTags (small binary markers) on the bezel of the monitor the plugin can map out the planar monitor surface, thus marking the exact size of the monitor in recording software. We designed and 3D printed custom monitor mounts to ensure no changes in marker setup.

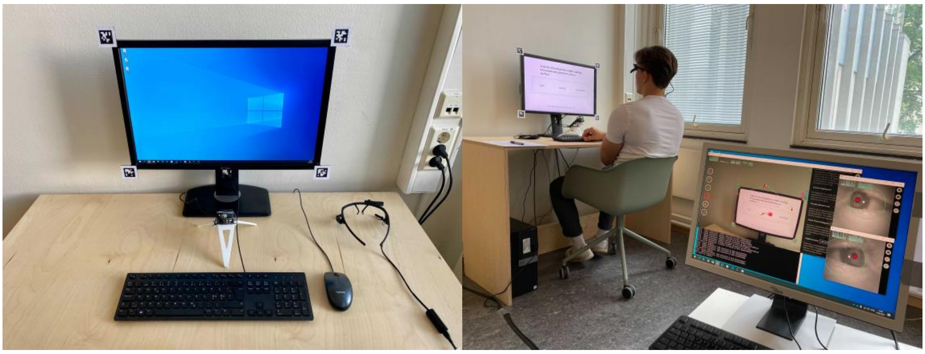


Figure 2. Left) Monitor with apriltags, Pupil Core to the right, and microphone in the middle. Right) Experimenter monitors eye-tracking real time during experimental run.

3.4. Participants

24 healthy adults (18 male/6 female, ages 23-35, mean = 25.8 yrs., SD = 2.9 yrs.,) participated in the study. 22 were right-handed and 2 left-handed. None of the participants were native English speakers, and none wore glasses to not interfere with the head-mounted eye-tracker. 8 participants used lenses. Participants were recruited through internal channels and contacts at relevant departments at the Norwegian University of Science and Technology (NTNU)—the Department of Mechanical and Industrial Engineering (MTP) and the Department of Design (ID)—to ensure a similar educational background as original participants. All participants were graduate level students or higher (minimum 4th year MSc, PhDs), with more than half of the participants being final year Master students. No monetary compensation was given.

3.5. Experiment procedure and calibration

After providing informed consent participants received general information in Norwegian about the experiment, its procedure, and the task. The eye-tracker was then correctly positioned on the participant before calibration of the eye-tracker. A 3D calibration was performed following manufacturers' "Best Practices" (Pupil Labs, 2021). Afterwards, the experiment started by showing additional information, before proceeding to explaining the design ideation task again and the 1-back task. Participants were

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

sequentially assigned to groups in order A, B, C. After completing the approximately 1 hour long experiment participants answered a demographic survey.

3.6. Knowledge from pilot participants

The experiment was piloted, following several procedures by van Teijlingen and Hundley (2001) to remove unexpected technical bugs or ensure clear task description. The experiment was conducted as if would be for the actual participants, the session was timed, feedback from participants to identify potential ambiguities was obtained afterwards, and eye-tracking data quality was inspected. This induced two experimental changes: 1) The initial chairs height led participants to angle their head which degraded eye-tracking data quality, and thus the chair was changed to one with an appropriate height relative to the table. 2) Priming the participants. Participants were unsure of what "rules" that applied during the design ideation. For example, did it only have to be new ideas or realistic ideas? The study's primary purpose wasn't a quality evaluation of generated ideas. The manuscript briefing participants initially was thus written informing participants about ideating freely and without any constraints, and a change in the written instruction in PsychoPy was made. Changes were iteratively implemented before commencing with actual participants.

3.7. Data Analysis

To summarize, the following data modalities were collected during the experiment: eye-tracking data, audio recordings, number and timing of generated ideas, and a questionnaire. The scope of this article is to analyze the number of ideas and subjective ratings from the questionnaire. Analysis of eye-tracking data and audio data is not within the scope of this article and will be published later.

Compared to the original study we hypothesize that this study will result in a similar number of ideas in order of magnitude, perhaps fewer du to having to think aloud and cultural factors. The subjective ratings and differences between ratings will be similar.

Questionnaire: Differences in subjective ratings between conditions were assessed with Friedman's test, a non-parametric test, suitable for ordinal data such as ratings on a 1-5 scale. Post hoc pairwise comparisons were assessed with Wilcoxon signed rank test with a Bonferroni correction for multiple comparisons (Field, 2018). Hedges' g was used as effect size.

Idea generation: For idea generation analysis the number of ideas were aggregated per stimuli and wordset. Differences in the number of ideas generated between conditions was assessed with one-way repeated measures ANOVA, suitable for continuous variables (Field, 2018). ANOVAs assumptions of sphericity and normal distribution of the data were assessed. The Control-Wordset1 contrast failed the normality test, but since ANOVA is relatively robust against normality violations, and all other data exhibited both sphericity and normality, we continued with the analysis. Partial-eta squared (η^2) was used as effect size. Statistics were performed in Python with Pingouin (Vallat, 2018), Pandas (McKinney, 2010), Seaborn (Waskom, 2021), and Matplotlib (Hunter, 2007). The significance level was set at $p<0.05$ for all tests.

4. Results and discussion

4.1. Questionnaire results

There was a highly significant difference between conditions for relevancy and usefulness, insignificant difference for novelty, and a $p = 0.051$ for quality. See Table 1 and Figure 3.

Post hoc pairwise comparisons are listed in Table 2. There was a significant difference in relevancy between Control and Near, and between Far and Near with Near being more relevant than both Control and Far. There was a significant difference in usefulness between Control and Far, and between Control and Near, with the inspirational stimuli conditions being more useful than control. Moreover, the difference between Far and Near reached a significance level of $p=0.05$, which we find interesting and interpret as a strengthening indication of that Near was more useful than far. There was not a significant difference between conditions for Novelty, indicating that participants did not consider their ideas to be more novel in either condition. There was one significant difference for Quality between Far and Near,

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

i.e., participants thought they produced ideas of higher quality in Near. These results exhibit similar trends to the original study with one exception: relevancy. Participants in this study rated Control to be less relevant than both Near and Far inspirational stimuli, whereas the original participants thought Control was more relevant than Near and Far. The experimenter noted during the experiment and in post experimental feedback that participants were unsure of what relevancy rating to give control stimuli, which may explain the difference. This ambiguity could potentially have been removed by clarifying whether to rate the relevancy of the words based on their relevancy for solving the problem or being related to the problem. Participants being non-native English speakers may also affect their interpretation of the question and the word "relevance", as it was read in English but processed and evaluated in Norwegian. Near and far inspirational stimuli was more useful than control, which corroborates the results from the original study. Near was close to significantly more useful than far in this study with a multiple-comparisons-corrected of 0.05, for which the original study reported a p<0.01. We don't know if this value is corrected for multiple comparisons or not, but if not, this might explain the discrepancy since this study's uncorrected p-value was 0.017. The novelty ratings corroborate the insignificant differences of the original study. Even though the trends were similar for questionnaire ratings, overall mean values for novelty and quality were lower. This can indicate lower confidence in the solutions generated by this study’s participants.

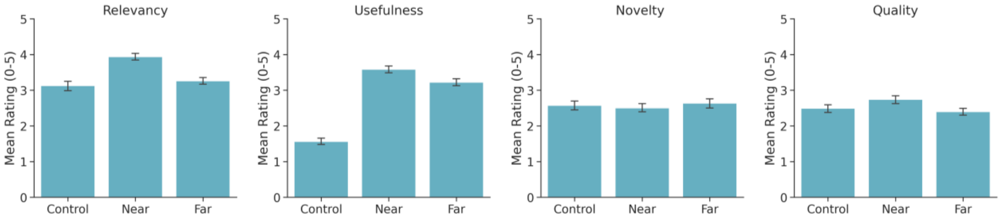


Figure 3. Mean ± 1 SE for participants subjective ratings

Table 1. Results subjective variables

	Control		Near		Far				
Variable	***M***	***SD***	***M***	***SD***	***M***	***SD***	***DOF***	***χ2***	***p***
Relevancy	3.12	1.35	3.94	0.92	3.26	0.94	2	16.587	<0.001**
Usefulness	1.56	0.83	3.58	0.94	3.22	0.99	2	39.758	<0.001**
Novelty	2.57	1.17	2.5	1.14	2.64	1.27	2	2.987	0.225
Quality	2.49	1.14	2.74	1.11	2.4	0.92	2	5.945	0.051

*p < 0.05, **p < 0.01.

Table 2. Pairwise comparisons subjective variables

Variable	**Between**		***W***	***p***	***Corr. p***	***Hedges' g***
Relevancy	Control	Far	126	0.726	1.000	-0.148
	Control	Near	51	0.015	0.044*	-0.886
	Far	Near	3	<0.001	<0.001**	-1.034
Usefulness	Control	Far	0	<0.001	<0.001**	-2.694
	Control	Near	0	<0.001	<0.001 **	-3.588
	Far	Near	35.5	0.017	0.05	-0.606
Novelty	Control	Far	73	0.886	1.000	-0.072
	Control	Near	89	0.362	1.000	0.088
	Far	Near	37.5	0.208	0.625	0.160
Quality	Control	Far	61	0.734	1.000	0.121
	Control	Near	64.5	0.133	0.398	-0.301
	Far	Near	20	0.008	0.023*	-0.428

*p < 0.05, **p < 0.01. Uncorrected p-value included for understanding of the effect of multiple comparisons correction.

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

4.2. Idea generation results

The number of ideas produced in Wordset1 was significantly higher than in Wordset2 for all conditions (Control: t(23) = 7.250 , p < 0.001, Near t(23) = 5.152, p < 0.001, Far: t(23) = 7.052, p < 0.001). See Figure 4 which also illustrates the differences between participants in the numbers of ideas produced. The number of ideas generated in each condition plotted over time in Figure 5 exhibits a similar shape as in the original study, but with an approximate 10 second temporal delay. This may have been caused by participants both being native English speakers, the think aloud protocol, or a combination thereof with or without other influencing factors.

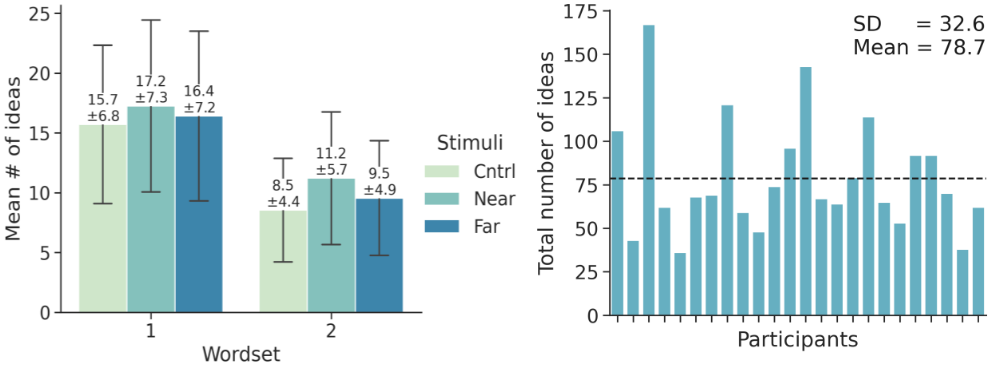


Figure 4. The number of ideas produced across stimuli and participants. Mean ± SD are on aggregated ideas across all wordset-stimuli combinations.

In Figure 4 and Figure 5 Near stimuli appears to generate more ideas than Far stimuli, again generating more ideas than Control stimuli—for both wordsets. However, there was not a statistically significant different number of ideas produced between conditions for Wordset1 (F(2, 46) = 2.241, p = 0.118, η2 = 0.089). This corroborates the original study's results. The number of ideas was significantly different between conditions for Wordset2. (F(2, 46) = 10.316, p < 0.001, η2 = 0.310). Post hoc pairwise comparisons for Wordset2 reveals a significant difference between Control and Near, and between Far and Near. See Table 3. This indicates that inspirational stimuli help participants retaining a higher idea output throughout the ideation session. It also interesting to note that ideas produced with Far inspirational stimuli was not significantly different from Control, contrary to the original study finding this contrast significant. The original study also resulted in a significant Control-Near contrast, but their Far-Near contrast was not significant at p<0.05, although it was close with a p=0.06. Both studies' mean values of idea output were indeed Near > Far > Control. Overall, these results indicate that inspirational stimuli nearer to the problem space facilitates idea generation. Moreover, original results have largely been corroborated.

Table 3. Pairwise comparisons number of ideas

Wordset	**Between**		***T***	***DOF***	***p***	***Corr. p***	***Hedges' g***
1	Control	Far	-1.026	23	0.315	0.946	-0.099
	Control	Near	-2.157	23	0.042	0.125	-0.215
	Far	Near	-1.069	23	0.296	0.888	-0.112
2	Control	Far	-1.906	23	0.069	0.208	-0.211
	Control	Near	-3.968	23	0.001	0.002**	-0.515
	Far	Near	-2.908	23	0.008	0.024*	-0.309

*p < 0.05, **p < 0.01. Uncorrected p-value included for understanding of the effect of multiple comparisons correction.

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

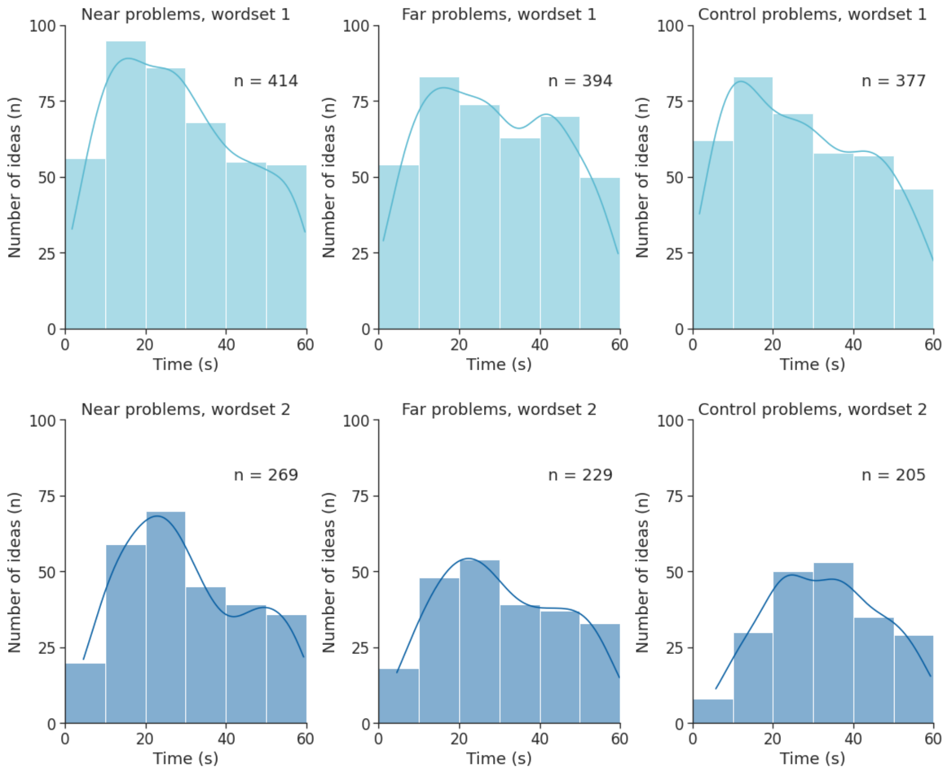


Figure 5. Histograms with number of ideas over time generated across all conditions, binned into bins of width 10 second. All histograms are overlaid with a kernel density estimate (KDE).

4.3. Further work

We are currently analyzing the eye-tracking data and audio recordings to get further insights to what kinds of ideas that were produced for the different problems across the different conditions. The audio recordings aren't published due to privacy considerations and a transcription will therefore be completed, and eventually publicly available. Further replication of the experiment may bring insight into potential differences in results due to culture and/or nationality, and is also of interest to further understand the effects of inspirational stimuli on idea generation.

5. Conclusion

This article described the replication and extension of a design ideation experiment with and without inspirational stimuli. Eye-tracking technology and a think aloud protocol was added to provide new insights into generated ideas. Preliminary results presented here corroborates the original study's results, inspirational stimuli influence idea output and questionnaire ratings. Participants produced more ideas over time when aided by inspirational stimuli, and rated inspirational stimuli as more relevant and useful than control. Future work will analyze eye-tracking data and audio recordings. The experiment and the data, and code are publicly available for reproducibility purposes.

Acknowledgement and Data availability

This work has been published in a master thesis.
Raw data, analysis code, results, and PsychoPy experiment code are all publicly available: code repository (Abelson, 2021), pre-processed data: (Abelson et al., 2021a), and raw eye-tracking data (Abelson et al., 2021b).

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

References

Abelson, F.G., 2021. Code Repository for Design Ideation Experiment (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5130090

Abelson, F.G., Dybvik, H., Steinert, M., 2021a. Dataset for Design Ideation Study. DataverseNO. https://doi.org/10.18710/PZQC4A

Abelson, F.G., Dybvik, H., Steinert, M., 2021b. Raw Data for Design Ideation Study. https://doi.org/10.21400/7KQ02WJL

Borgianni, Y., Maccioni, L., 2020. Review of the use of neurophysiological and biometric measures in experimental design research. AI EDAM 34, 248–285. https://doi.org/10.1017/S0890060420000062

Cao, J., Xiong, Y., Li, Y., Liu, L., Wang, M., 2018. Differences between beginning and advanced design students in analogical reasoning during idea generation: evidence from eye movements. Cogn Tech Work 20, 505–520. https://doi.org/10.1007/s10111-018-0477-z

Carter, B.T., Luke, S.G., 2020. Best practices in eye-tracking research. International Journal of Psychophysiology 155, 49–62. https://doi.org/10.1016/j.ijpsycho.2020.05.010

Colombo, S., Mazza, A., Montagna, F., Ricci, R., Monte, O.D., Cantamessa, M., 2020. NEUROPHYSIOLOGICAL EVIDENCE IN IDEA GENERATION: DIFFERENCES BETWEEN DESIGNERS AND ENGINEERS. Proceedings of the Design Society: DESIGN Conference 1, 1415–1424. https://doi.org/10.1017/dsd.2020.161

Duchowski, A.T., 2017. Eye-tracking Methodology, 3rd ed. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-57883-5

Field, A., 2018. Discovering statistics using IBM SPSS statistics, 5th edition. ed. SAGE Publications, Thousand Oaks, CA.

Gero, J.S., Milovanovic, J., 2020. A framework for studying design thinking through measuring designers' minds, bodies and brains. Design Science 6. https://doi.org/10.1017/dsj.2020.15

Goucher-Lambert, K., Moss, J., Cagan, J., 2019. A neuroimaging investigation of design ideation with and without inspirational stimuli—understanding the meaning of near and far stimuli. Design Studies 60, 1–38. https://doi.org/10.1016/j.destud.2018.07.001

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95. https://doi.org/10.1109/MCSE.2007.55

Kassner, M., Patera, W., Bulling, A., 2014. Pupil: an open source platform for pervasive eye-tracking and mobile gaze-based interaction, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, UbiComp '14 Adjunct. Association for Computing Machinery, New York, NY, USA, pp. 1151–1160. https://doi.org/10.1145/2638728.2641695

Kwon, E., Ryan, J.D., Bazylak, A., Shu, L.H., 2019. Does Visual Fixation Affect Idea Fixation? Journal of Mechanical Design 142. https://doi.org/10.1115/1.4045600

Martin, G.N., Clarke, R.M., 2017. Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. Frontiers in Psychology 8.

McKinney, W., 2010. Data Structures for Statistical Computing in Python. Presented at the Python in Science Conference, Austin, Texas, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349. https://doi.org/10.1126/science.aac4716

Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J.K., 2019. PsychoPy2: Experiments in behavior made easy. Behav Res 51, 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pupil Labs, 2021. Best Practices - Tips for conducting eye-tracking experiments with the Pupil Core eye-tracking platform. [WWW Document]. Pupil Labs. URL https://docs.pupil-labs.com (accessed 4.27.21).

Rayner, K., 2009. Eye movements and attention in reading, scene perception, and visual search. Quarterly Journal of Experimental Psychology 62, 1457–1506. https://doi.org/10.1080/17470210902816461

Shrout, P.E., Rodgers, J.L., 2018. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. Annu. Rev. Psychol. 69, 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Vallat, R., 2018. Pingouin: statistics in Python. Journal of Open Source Software 3, 1026. https://doi.org/10.21105/joss.01026

van Teijlingen, E.R., Hundley, V., 2001. The importance of pilot studies.

Wade, P. of V.P.N., Wade, N., Tatler, B.W., Tatler, L. in P.B., 2005. The Moving Tablet of the Eye: The Origins of Modern Eye Movement Research. Oxford University Press.

Waskom, M., 2021. seaborn: statistical data visualization. JOSS 6, 3021. https://doi.org/10.21105/joss.03021

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

Appendix C6: Academic contribution 6

Dybvik, H., Løland, M., Gerstenberg, A., Slåttsveen, K. B., & Steinert, M. (2021). A low-cost predictive display for teleoperation: Investigating effects on human performance and workload. International Journal of Human-Computer Studies, 145, 102536. https://doi.org/10.1016/j.ijhcs.2020.102536

International Journal of Human-Computer Studies 145 (2021) 102536

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

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

A low-cost predictive display for teleoperation: Investigating effects on human performance and workload

Henrikke Dybvik*, Martin Løland, Achim Gerstenberg, Kristoffer Bjørnerud Slåttsveen, Martin Steinert

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology (NTNU), Richard Birkelands vei 2B, 7491 Trondheim, Norway

ARTICLE INFO

Keywords:
Predictive display
Human operator
Performance
Subjective workload

ABSTRACT

Teleoperation in an environment with latency is difficult and highly stressful for human operators, resulting in high cognitive workload and decreased human performance. This work investigates if a simple predictive display can increase performance and lower subjective workload for the human operator when teleoperating a remotely operated vehicle (ROV). A predictive display based on image transformation was developed by applying positional and scale transformations to the video feed and tested. An experiment was designed, consisting of a simple navigational task (peg-in-hole game) with a ground ROV, in three distinct conditions: C1. Latency, C2. Latency with predictive display (PD) and C3. Baseline (no added latency). Findings from N = 57 participants show a statistically significant increase of 20% in human performance with the aid of the predictive display. Although differences in subjective workload was not statistically significant, both subjective performance and actual game performance did increase significantly by using the predictive display. In fact, the latter almost doubled for participants defining themselves as regular gamers. Lastly, A principle component analysis (PCA) was conducted investigating confounding factors with confirmatory results.

1. Introduction – Predictive technology can combat the detrimental effects of latency in teleoperation

Teleoperation, also called remote operation, is electronic remote control of machines or vehicles and it includes applications of remotely operated vehicles (ROVs) on ground, under water, subsea, aerial and in space (Draper et al., 1998). Teleoperation is a subclass of telepresence; “the perception of presence within a physically remote or simulated site” (Draper et al., 1998). Telepresence is generally viewed as being beneficial to mission performance and is furthermore hypothesized to improve efficiency and/or reduce operator workload (Draper et al., 1998). There are multiple challenges related to teleoperation, one of which is latency. In this work, we are interested in latency, also called time delay, which refers to the delay between operator input action (steering commands) and visible output response of the video feed (Chen et al., 2007). Teleoperation in an environment with latency, especially basic driving, is difficult and highly stressful for the human operator, resulting in high cognitive workload (Matheson et al., 2013) and decreased human performance (Chen et al., 2007), e.g. observed as an increase in task completion time or reduced accuracy (Lane et al., 2002). Approaches to overcome the detrimental effects of latency in teleoperation include increasing the level of automation (which excludes the human from the loop), provide information to increase the situational awareness of the human operator and predictive technology.

Predictive technology spans several approaches, either categorized as dynamic system models or free model approaches. Model free approaches include superimposed information models, 3D graphic models, and video manipulation. Superimposed information and 3D graphics models show promising results by greatly reducing task completion times, but require advanced algorithms, potentially expensive equipment and extensive information regarding the environment and the ROV. Video manipulation can increase performance of human operators' and it is simpler in comparison, as it alters the delayed video feed to mimic movements and environment in real time. Simple video manipulation can provide time efficient and inexpensive means to enhance performance of human operators' in settings where extensive information regarding the ROV and its environment is unavailable, or the opportunity to utilize expensive equipment or highly advanced algorithms is not a possibility.

With basis in existing video manipulation methods based on image transformation, we developed a simpler predictive display by applying image positional and scale transformations to the video feed. This

[*] Corresponding author.
E-mail address: henrikke.dybvik@ntnu.no (H. Dybvik).

https://doi.org/10.1016/j.ijhcs.2020.102536
Received 1 July 2019; Received in revised form 18 August 2020; Accepted 21 August 2020
Available online 25 August 2020

predictive display requires a few lines of code and can be applied to several ROV configurations. In this work, we are interested in human operators' performance and their subjectively experienced workload while using predictive technology. The aim of this article is to investigate if a simple predictive display can increase performance and lower subjective workload of human operators' during teleoperation. To do so an experiment was set up to investigate changes in human operator performance and workload when operating an ROV under three distinct conditions, each condition with a distinct display and latency. The participants were presented with a single, simple navigational task, framed as a peg-in-hole game using a ground ROV with a first-person camera view. The conditions were C1. Latency, 2. Latency with Predictive Display (w/PD) and 3. Baseline. Data collected included objective performance (task score), and subjective workload (RTLX), demographics and other variables. N = 57 participants were recruited and the hypotheses (task performance and subjective workload) tested using ANOVA. A post hoc Exploratory Data Analysis (EDA), specifically a principal component analysis (PCA) explore influencing factors.

Following the introduction, the article is structured as follows; the background second section covers challenges in teleoperation, focusing on latency and its detrimental effects on human performance and workload. Means to compensate for latency are discussed, emphasizing various predictive technologies. The third section describes development and implementation of a predictive display, and the experiment design, including stimuli, data collection, procedure, and data analysis. Section four presents the results of the statistical tests before providing the result from the EDA. A discussion of the presented results follows in section five, before the conclusion.

2. Background – Latency in teleoperation, human performance, and workload

This section describes challenges in telepresence, detailing latency and its detrimental effects, with a focus on human operator performance and subjectively experienced workload. Human operator performance decrease and workload increase as latency is introduced in teleoperation. Means to compensate for latency are discussed, predictive technologies in particular. Lastly, the section provides means to measure human operator performance objectively and workload subjectively.

2.1. Latency in teleoperation and its related challenges

2.1.1. Telepresence and related challenges

Draper et al. (1998) defines telepresence as "the perception of presence within a physically remote or simulated site". Teleoperation is one subclass of telepresence (Sheridan, 1995). Telepresence is beneficial to mission performance and is furthermore hypothesized to improve efficiency and/or reduce operator workload. Chen et al. (2007) reviewed 150 articles investigating factors in telepresence, and how they influence operator performance and challenges related. They found eight main factors; field-of-view (FOV), orientation, camera viewpoint, depth perception, video quality and frame rate, time delay (or latency) and motion.

2.1.2. Latency challenges in telepresence

In this work, we are interested in time delay, or latency, which will be used throughout this article, which refers to the delay between operator input action (steering commands) and visible output response of the video feed (Chen et al., 2007). Latency is usually a result of information having to be conveyed over a communication network (Chen et al., 2007). The total latency of the teleoperation system can further result from a combination of a number of reasons, such as software design, hardware design, physical limitations such as distance and obstacles, signal processing, etc. Thus, total latency can be both fixed and variable (Lane et al., 2002). There are important distinctions between the two, e.g. they influence performance differently (Davis et al., 2010; Neumeier et al., 2019; Oboe and Fiorini, 1998).

The causes of latency are not within the scope of this work, and we consider fixed delay only. We are interested in the total perceived latency; i.e. the time from when the human operator issues a command until they visually perceive a reaction in the vehicle in the video feedback.

2.1.3. Latency in teleoperation and its detrimental effects

Latency produces a mismatch between given input commands and visual feedback of vehicle reactions. This creates a conflict for human perception. To correct for this during operation the human operator must remember the inputs command given until they see the desired output produced by the vehicle in the video feed (Matheson et al., 2013). In addition, as new information is prompted on the video feedback this must be mentally connected with the commands issued previously (i.e. the vehicles previous state), and thereafter combined that with issuing new commands based on this conjunction of information (Ricks et al., 2004). Latencies as low as 10 - 20 ms can be detected by humans' visual perception (Chen et al., 2007). Taken together, this can degrade human performance (Chen et al., 2007) and can increase subjectively experienced workload (Ricks et al., 2004).

2.1.4. Latency in teleoperation degrades human performance

The detrimental effect of latency on human performance can be seen in Table 1, which includes relevant research investigating the effect of video feed latency on human performance in a given task. Human performance includes course completion time, task completion time, task score, accuracy, etc. This table describe the task and the related increase factor, where a 40% increase in task completion time corresponds to an increase factor of 1.40. For example a needle-driving task at 100 ms latency had an increase factor of 1.5 (Xu et al., 2014). The relationship between latency and task completion time is task dependent, notably it is similar for similar tasks. For example; a linear relationship between latency and task completion time was found in a mobile robot operating task (Ando et al., 1999) and a vehicle peg-in-hole task (Lane et al., 2002), whereas an exponential relationship was found in a telerobotic surgical technique task (Xu et al., 2014).

Table 1
Task completion time for a variety of tasks and latencies.

Author	Task	Participants	Latency [ms] and increase factor		
			100 – 300 ms	400 – 700 ms	800 – 1500 ms
(Fabrizio et al., 2000)	Pin transfer	N = 6	1.04 - 1.21*	1.17 - 1.41*	1.11 - 1.58*
(Xu et al., 2014)	Energy dissection	N = 16	1.4 - 1.8	2.7 - 4.3	
(Xu et al., 2014)	Needle-driving	N = 16	1.5 - 2.1	2.5 - 6.2	
(Perez et al., 2016)	Surgical simulator	N = 37	0.75	1.5	
(Lum et al., 2009)	Block transfer	N = 14	1.45	2.04	
(MacKenzie and Ware, 1993)	Target acquisition	N = 8	1.64		

* Estimated from graph.

2.1.5. Latency in teleoperation increase workload

The notion of workload or cognitive load is argued to be predictive of both performance in human-machine interactions as well as the mental state of the operator. Workload is described as a relation between the mental resources a task demands and the resources available from the human operator (Parasuraman et al., 2008). It is a multidimensional construct emerging from the interaction between task, context, operator capabilities, behavior, perceptions and (mental and physical) state (Hart and Staveland, 1988a; Parasuraman et al., 2008). This mental load posed on a human operator by latency in teleoperation negatively affects their ability to control a vehicle in an efficient manner (Ricks et al., 2004). The human operator's *subjectively* experienced workload is important (Hart and Staveland, 1988a), since this might alter behavior. Should an operator experience a situation as high workload, the operator might adopt strategies to mitigate workload. In the specific case of teleoperation human operators tend to perform steering commands correcting for the mismatch in given input and visually perceived output, causing the vehicle to oscillate and limiting top speed (Appelqvist et al., 2007). Teleoperation in an environment with time delay, in particular basic driving, is difficult and highly stressful for the human operator, resulting in high cognitive workload (Matheson et al., 2013). Extended exposure to such an environment can create cognitive overload leading to mental fatigue (Lim et al., 2010; Matheson et al., 2013).

2.1.6. Latency compensation

There are multiple approaches to reduce the detrimental effects of latency. First, increasing the level of automation (LOA) reduces the operator workload and improve safety (Dorais et al., 1999; ENDSLEY, 1999; Goodrich et al., 2001; Luck et al., 2006; Schutte, 2017). A second option is providing the human operator with information and/or previously given input commands, increasing situational awareness and leading to higher performance and/or decreasing subjective workload (Chen et al., 2007; Miller and Machulis, 2005; Nielsen et al., 2007). A third option is predictive technology, which can be displays, control algorithms and graphical models attempting to predict the state of the ROV based on the vehicles current state and commands issued by the operator. Chen et al. (2007) conclude it is the most promising solution if eliminating latency from the system is impossible, and highlight that predictive displays has been shown to reduce task performance time by up to 150%.

2.1.7. Predictive technology

A range of experiments where predictive technology has been used are shown in Table 2, illustrating a wide variety of experimental tasks, robot configurations and predictive method. Exact robot configuration can be known, including examples such as robot-arm manipulators fixed to a user defined reference frame, or not known, such as vehicles subjective to external forces or floating freely. The unknown robot configuration challenges the predictive technology as it must account for unknown and changing external factors. Common for the experiments in Table 2 is that they involve lateral movement in an alignment or aiming task, which are particularly exposed to detrimental effects of latency in video feedback. Correctional behavior commonly occurs, causing operators to overshoot a target or employ a wait-and-move strategy. This behavioral strategy increases task completion time and occurs around one second latency (Lane et al., 2002).

In general, predictive technology calculates a future predicted state of the robot based on different variables and methods. Methods can rely on dynamic system equations, such as Zhang and Li. (2016) who used a spacecraft's state equations and its dynamic properties to calculate the predicted state. An image of the predicted state is provided to the operator which can issue commands accordingly. In contrast, a model free approach, which excludes dynamics, is often used in contexts where accurate modeling of external forces isn't a possibility, such as in space applications. Predictive technology within model free approaches includes superimposed predictive information, 3D graphic models and video manipulation.

The first category superimposes (or overlays) information on a delayed video feed, providing the operator with an estimate of the vehicles future state. Superimposed predictive information is often visualized as vector graphics where lines of dots follow a path. For example, Mathan et al. (1996) superimposed directional velocity information of a lunar rover on a video display. Further, airplane and helicopter displays have a *tunnel in the sky* showing where the aircraft should be going and a cross indicating the predicted trajectory (Grunwald et al., 1981). In cases with large amounts of lateral movement this approach might not be applicable as the predicted heading can come off screen.

3D graphics model (or virtual reality (VR) based predictive display) use sensor technology input such as Monocular Simultaneous Location and Mapping (SLAM), stereo imagery, vision-based structure from motion (SFM), light detection and ranging (LiDAR), or radio detection and ranging (radar), etc., to construct a three-dimensional world, wherein images from ordinary cameras are rendered on the surface of the virtual world.

Then, a virtual camera is placed inside the virtual world in the predicted position of the real camera and operators' are presented with the virtual video feed as virtual reality (VR) or augmented reality (AR). This method is particularly popular in combination with robot arm

Table 2
Predictive technology with task completion time reduction.

Author	Robot system	Predictive technology	Participants	Reduction in task completion time
	Task	Camera	Latency	
(Lu et al., 2018)	Car simulator	Model-free framework	N = 12	8%
	Driving	Simulated human	Not reported	
(Hu et al., 2016)	2-6 DOF manipulator	Simulated 3D	N = 15	33%, 58%, 65%*
	Camera alignment	Virtual	300 ms, 500 ms, 1000 ms	
(Zheng et al., 2016)	Car simulator	Model-free framework	N = 5	35%
	Driving	Simulated human	900 ms	
(Lovi et al., 2010)	Robot arm on Segway	Vision-based monocular modelling	N = 5	33%*
	Object alignment	At end effector	300 ms	
(Matheson et al., 2013)	Rover	Projected field of view	N = 12	48% - 64%*
	Driving	Fixed to car	3000 ms	
(Rachmielowski et al., 2010)	Virtual with Phantom OMNI	Reconstructed 3D environment	N = 12	29% - 30%*
	Alignment	At end effector	300 ms	
(Mathan et al., 1996)	Lunar vehicle	Superimposed directional information	N = 8	24% - 30%
	Manoeuvring	Fixed to car	5000 ms	
(Bejczy et al., 1990)	6DOF PUMA robot	Superimposed phantom robot	N = 2	13% - 34%, 40% - 56%
	Tapping	Fixed	1000 ms, 4000 ms	

* Estimated from graph.

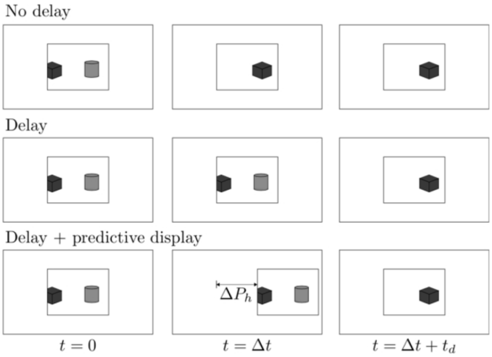


Fig. 1. Monitor for the human operator. The outer box is total screen size, whereas the inner box is the video feed.

manipulators. The 3D environment can be constructed a priori, and exact location of the robot arm is known (Ricks et al., 2004). A limitation arises when tasks are performed in unknown and unstructured areas, and since environment geometry is unknown real time mapping and rendering can be difficult. Additional hardware may be required and calculations can become computationally intensive. Moreover, additional challenges, such as oscillopsia occur when latency is introduced in VR head-mounted displays (Allison et al., 2001).

Video manipulation does not require 3D information about the environment. It alters the delayed video feed to mimic movements and environment in real time. A simple example would be to zoom into the image if the robot is moving forward. Matheson et al. (2013) halved task completion time at a latency of three seconds in an ROV experiment using this method, by cropping and projecting the image. A similar result is obtained by capturing a wide FOV video, possibly 360 degrees, and then only displaying a section of that image to the operator. The section can be moved around in the video as a response to steering commands and thus provide fluid and seemingly real time feedback (Baldwin et al., 1999). Advantageous to video manipulation techniques are low cost, ease of implementation and not requiring a structured environment. Furthermore, prediction error propagation cannot occur since the presented video feed consists only of alterations to the latest image. However, it cannot recreate parallax movement (such as passing an object or corner) which 3D graphics models can achieve.

3. Method - Experiment investigating a predictive display under three conditions

An experiment was set up to investigate changes in human operator performance and workload when operating an ROV under three distinct conditions, each condition with a distinct display and latency. The participants were presented with a single, simple navigational task, framed as a peg-in-hole game which was the same for all three conditions. The conditions were C1. Latency, C2. Latency with Predictive Display (w/PD) and C3. Baseline, and they are described in detail in this section. First, this section describes development and implementation of a predictive display Then, the experiment design follow, which includes research objective, hypotheses, stimuli (description of task and conditions), data collection (objective performance and subjective workload), setup, experimental procedure, and data analysis.

3.1. Predictive display development

Predictive technology that reconstructs a 3D environment based on sensory data requires advanced algorithms, potentially expensive equipment, and extensive information regarding environment and ROV. In cases where this is not a possibility video manipulation provides simple and inexpensive means to increase human operators' performance.

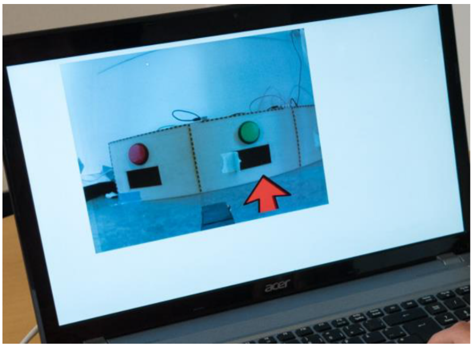


Fig. 2. Predictive display visualization. The operator has recently turned the ROV to the right, and as a result the video has moved to the left. The red arrow has not moved and works as an indication of where the ROV will be heading when the video feed has caught up with the time delay.

The projected display by Matheson et al. (2013) is the simpler video manipulation method of the ones considered in Table 2, while retaining a great increase in human operator performance. However, information on the vehicles' ground trajectory is required to calculate changes in perspective. By disregarding the effects of change in perspective and applying positional and scale transformations to the video feed we obtain an even simpler approach. As such, by applying positional and scale transformations to the video feed we developed a predictive display based on image transformation. The predictive display can be applied to several robot configurations though it was developed for ROVs initially; It is appropriate only for screen-based systems and other alternatives are needed for predictive head-mounted display systems.

3.1.1. Predictive display implementation in detail

The developed predictive display repositions the delayed video feed on the monitor so objects in the video feed appear in correct size and position on the screen as if there was no latency (see Figs. 1 and 2). It uses user input (i.e. steering commands) and predefined ROV speed to predict how the FOV would move in the scene, repositioning and scaling the video feed accordingly.

The positional transformation can be explained by considering an ROV with an onboard camera rotating about its center of mass, turning with an angular velocity of $\omega°/s$. The camera FOV is $\varphi°$, with horizontal resolution R_h pixels. A counterclockwise rotation for Δt s moves the ROV $\Delta\theta°$. Objects in the video feed moves $(R_h{\cdot}\omega/\varphi){\cdot}\Delta t{=}\eta{\cdot}\Delta t{=}\Delta P_h$ pixels to the right. η = *pixel turn rate*, which depends on screen resolution, angular velocity and camera FOV. The pixel turn rate, user input and total system delay t_d is used to create the predictive display. The video feeds' position on the monitor is calculated at a set interval *dt* (preferably at a minimum video frame rate (FPS)). If the ROV moves to the left, time since last update *dt* multiplies with pixel turn rate to find change in horizontal video position ΔP_h. The video feed then moves ΔP_h to the right on the monitor. When a time t_d has passed (system delay has caught up), the video feed is moved back.

For backward and forward translation, similarly as for *pixel turn rate*, a *pixel scale rate* can be found and used to scale the video feed. For backwards and forwards ROV translation, scaling of objects depends on how close they are to the camera. An average distance is used as an approximation. The video feed scale transformation works as the aforementioned positional transformation.

Finally, the predictive display uses a red arrow to visualize the future position as illustrated in Fig. 2.

3.2. Research objective and hypotheses

Research objective: Investigate if such a simple predictive display still increase human operators' performance and reduce workload.

Based on the research objective, we sought to test the following hypotheses:

- A simple predictive display significantly increases human operators' performance (objectively measured by task score performance, - i.e. the number of hits achieved in 90s by the participant).
- A simple predictive display significantly decreases human operators' subjective workload (subjectively measured by RTLX's six dimensions, mental demand, physical demand, temporal demand, performance, effort and frustration, evaluated on eleven-point scales (Hart, 2006)).

3.3. Stimuli – Peg-in-hole-game under three conditions

The experiment encompassed a single, navigational task, in which we measured operator performance by means of an achieved score over a fixed time period.

3.3.1. Rationale behind task selection

Chen et al. (2007) reports benefits of predictive technology to be very task dependent. A peg in hole task was selected due to its applicability in teleoperation (Lane et al., 2002). A task encompassing as much lateral navigation as possible was selected, as this is where the predictive display can provide the most help, in contrast to for example navigational tasks with longer stretches of forward motion (and the maximal velocity of the ROV would create a ceiling effect). A short timeframe of 90 seconds was chosen to reduce any learning effect that might accompany a longer maneuvering course. A fixed time period made total experiment length predictable, participants used 10 min and 56 s on average (SD 1 min and 12 s). This aided in recruiting new participants. Furthermore, time pressure in combination with score achievement made participants fully devoted to the task at hand, and we argue this led to participants performing close to the best of their ability. We further argue that a single, simple task will minimize the effect of other factors on performance, e.g. trouble understanding the task, or being highly experienced in related tasks such as gaming, driving, or other navigational tasks.

3.3.2. Task

Participants were given a modified 'peg-in-hole' task. The peg was mounted on a remotely controlled ground vehicle, and there were three rectangular holes in three rectangular boxes with accompanying LEDs. One LED would light up at a time, in random order, to which the participant was instructed to perform as many 'hits' as possible by inserting the peg in the hole within the given timeframe. Task and time given (90s) was the same for three distinct conditions. During the task, a red timer indicating remaining time was constantly visible in the screens' upper right corner.

3.3.3. Three conditions

All participants repeated the task three times, under three distinct conditions. The display provided to the participants would differ in each condition. The conditions, latency and displays were as follows:

Condition 1. Latency: 700 ms delay (250 ms inherent system delay + 450 added delay). No predictive display.
Condition 2. Latency with Predictive Display (PD): 700 ms delay (250 ms inherent system delay + 450 added delay). With predictive display.
Condition 3. Baseline: 250 ms inherent system delay.[1] No predictive display.

Throughout the paper we refer to the conditions as:

C1. Latency
C2. Latency w/PD
C3. Baseline

3.3.4. 3 × 3 Latin Square Design

The sequence of the conditions was randomized according to a 3 × 3 Latin Square Design to avoid potential order and/or learning effects. All six combinations were used. Each participant was automatically assigned to one of the combinations, ensuring equal group sizes across conditions as far as possible. Due to the number of participants recruited; three of the combinations had 10 participants, and three combinations had 9 participants.

3.4. Data collection – Performance measured objectively and workload measured subjectively

N = 58 participants were recruited to test the predictive display. We collected objective measures of human performance and subjective measures of workload. Demographic data were also collected.

3.4.1. Participants

Participants were voluntary selected from NTNU, Department of Mechanical and Industrial Engineering. Our aim was to recruit as many participants as possible within the time constraint we were working with. A total of 58 participants performed the experiment, one participant was excluded in the analysis due to incomplete information. The remaining N = 57 participants received the same information and were included in the analysis. Age ranged from 23 to 30 years (24.7 ± 1.5). There were 19 female and 38 men. We gathered level of education, how often they played video games, how often they use a computer and eye health information, which can be found in Table 3.

3.4.2. Objective performance measurements

Two performance measurements are common among experiments on predictive technology: course completion time and task score (Lu et al., 2018; Mathan et al., 1996; Matheson et al., 2013; Zhang and Li, 2016; Zheng et al., 2016). In the former, the task is to navigate through a predefined pathway with the vehicle and measuring the time necessary to complete the course. In the latter, the task typically involves aligning or aiming at a given target, assigning a score to the number of times the target was met. Using a task score as a performance measure enables a fixed time for experiments, which was desirable for us to be able to recruit more participants. The number of hits made by participants in each of the 90 s test period was used as a performance measure.

Additional objective data collected included total number of hits made in all three test periods, and number of key presses in each of the test periods.

3.4.3. Subjective workload measurements

NASA Task Load Index (TLX) is common and highly accepted for remote operation and ROV applications (Hart, 2006; Hill et al., 1992; Hu et al., 2016; Ma and Kaber, 2006; Zhang and Li, 2016), and was initially developed for experimental tasks that include cognitive and

[1] The variability of inherent system delay was repeatedly quantified (10 times) to 1 – 5 ms difference each time. The average of those 10 measurements was used as inherent system delay.

Table 3
Participant data.

Variable	Options	Frequency	Percent
Gaming	Daily	2	3.5
	Weekly	15	26.3
	Monthly	8	14.0
	Yearly	17	29.8
	Never	15	26.3
Education	Nursery school	1	1.8
	Some college credit, no degree	38	66.7
	Bachelor's degree	10	17.5
	Master's degree	8	14.0
Eye health	No visual aid	32	56.1
	Spectacles	4	7.0
	Contact lenses	10	17.5
	Both spectacle and contact lenses	11	19.3

manual control tasks, and supervisory control tasks (Hart and Staveland, 1988b). TLX is multidimensional, provide good diagnostic properties for assessing underlying mechanisms of subjective workload, and has been shown to have high sensitivity (Hart, 2006; Hendy et al., 1993; Hill et al., 1992; Vidulich and Tsang, 1987). A modified version of TLX, Raw TLX (RTLX) was chosen to assess workload. The six dimensions (mental demand, physical demand, temporal demand, performance, effort and frustration) were rated on eleven-point scales. The weighting process in TLX consists of pairwise comparison of all six dimensions. It was not conducted, since we are not interested in the subjective importance of each dimension in a specific task, rather we're interested in comparing the subjective workload of different tasks (the three conditions). Furthermore, this weighing process consumes time, and, in this context, it was deemed more important to have a short survey, leaving more time for recruiting participants and conducting experimental runs. This modification is what is referred to as RTLX. One additional modification was made to the survey, as a pilot study of the experiment showed that a participant found it more intuitive to rate good performance with a high number. In the original survey a low value corresponds to good performance. Therefore, this metric and the corresponding description was reversed, such that a high value corresponded to good performance. After data collection, this value was reversed back for conventional analysis and reporting.

Furthermore, a question of perceived delay time was added to the survey, to investigate participants' subjective experienced latency in each individual condition and to compare the individual conditions, the latter in hopes of providing a measure of effectiveness for the predictive display in reducing the subjectively perceived latency of the system.

3.4.4. Data collection procedure

Both survey data and experiment data were recorded with the ROV computer using an SQLite database.

3.5. Setup

A 17” laptop running a 2.3GHz Intel Core i7-3610QM CPU and Windows 10 was used. The laptop screen served as monitor and the keyboard's arrow keys were used to steer the ROV. The keyboard and a remote mouse were used to answer the surveys. The ROV was running a Raspberry Pi 3 Model B+, and equipped with a forward facing Raspberry Pi Camera V2 and a wide angle lens with horizontal FOV of 76.5°. The robot was constructed using three wheels, two of them connected to a DC motor and the third a caster wheel for support (see Fig. 4). A wooden box with three holes and LEDs were used to register task performance. The distance between the holes (center to center) was $D = 30$ cm while the holes itself has a width of $W = 10$cm. This translates to a Fitts's index of difficulty of $Id = \log2\ (2D/W) = 2.58$ bits (Fitts, 1954). The robot ran eduROV[2] software, which provided an interface to control the robot, handling control commands, adding desired latency to the communication, and logging data.

3.6. Experimental procedure

After entering the experiment room, participants were shown the setup to ensure that they understood the situation and what they were tasked to do. The participant was placed in a chair at a desk with a laptop, with their back to the game (see Fig. 3). The participant would have no visual perception of the physical setup during the experiment. To ensure there was no auditory perception of the ROV, participants wore an ear protection headset. Information was given in writing on the computer screen. After giving consent to participate in the experiment, participants filled out a demographic survey. Information describing the experiment was provided; How to steer the vehicle, the task and performance measure, and the following procedure of the experiment. Each participant was automatically assigned to one of the groups corresponding to the 3 × 3 Latin Square Design. The participant would then conduct a 30s practice period followed by a 90s test period. After each block of practice and test period the participant filled out a survey of mental workload and perceived delay time. The starting position (indicated by the black mark in Figs. 3 and 4) was identical for all periods. The third block concluded the experiment and the participants were escorted out.

The participant was not informed of the fact that one of the conditions would have a predictive display, nor how it worked. To be able to take advantage of the predictive display is therefore dependent on the individual participants ability to intuitively understand the display. It was assumed that the practice period before each test would suffice in giving the participant the needed training in the display for the game. However, the questionnaire included a question of time delay, which could have influenced participant's attention to delay in the next conditions.

3.7. Analysis – Classical statistics and exploratory multivariate analysis

3.7.1. Classical statistics – Analysis of variance (ANOVA)

Subjective measurements used for analysis were collected after each condition and performance measurements were collected continuously during each condition. An analysis of variance (ANOVA) was conducted to investigate the effects of the predictive display on both subjective and objective measurements, i.e. this statistical test investigated the predetermined hypotheses. The characteristics of the data was inspected and in the case of violations of assumptions, the non-parametric alternative to one-way repeated measures ANOVA, the Friedman test was conducted. Data distribution was visually inspected using Normal Q-Q Plots for all variables and conditions. ANOVA F-test is found to be insensitive or robust (Krishnaiah, 1980; Schmider et al., 2010) to general nonnormality, and can for equal group sizes be used with confidence in most practical situations. We consider the sample size of 57 to be high, and we have continued with the analysis and when possible conducted a Friedman test for comparison purposes. Mauchly's test evaluates sphericity, an assumption which is considered difficult not to violate in practice (Weinfurt, 2000), over-detecting deviations from sphericity in large samples (Kesselman et al., 1980). Maxwell and Delaney (2003), recommend using an adjusted test, interpreting the result of using a Greenhouse-Geisser correction and thus ignoring the result of Mauchly's test. This was done here, calculating epsilon according to Greenhouse & Geisser (1959), and using it to correct the one-way repeated measures ANOVA. The Bonferroni post hoc test (Maxwell, 1980, Maxwell &Delaney 2004) was used to test all possible pairwise combinations of conditions. Statistical tests were performed using SPSS Statistics (*IBM SPSS Statistics 25*, 2017).

[2] https://github.com/trolllabs/eduROV/.

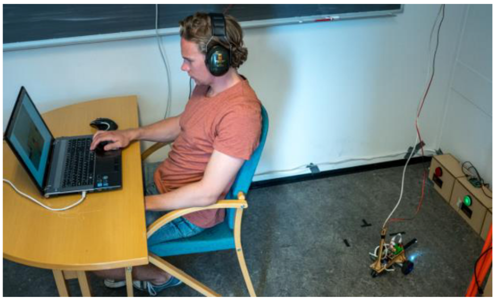


Fig. 3. Experiment setup. The participant can only see the robot through the display provided on the laptop screen, which is a first-person camera view.

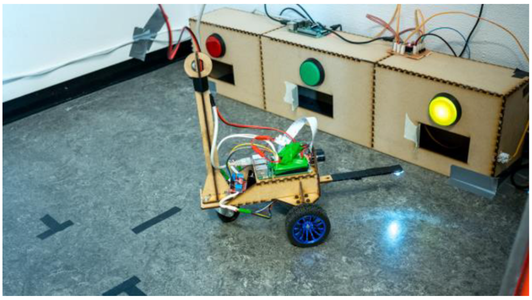


Fig. 4. Experiment setup. The three wheeled ROV with the peg mounted and the wooden box.

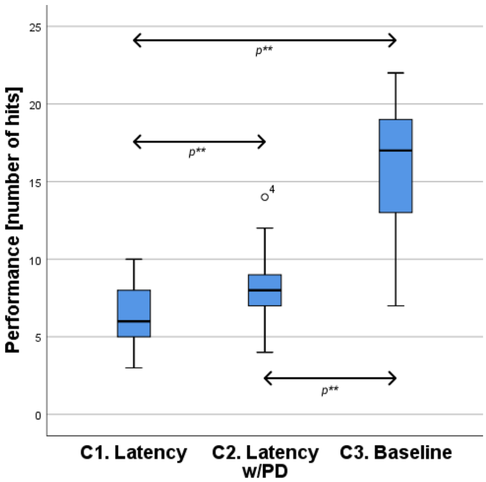


Fig. 5. Descriptive statistics of performance (objective). Original data reported. Statistically significant differences at p < 0.01 are indicated by *p***.

3.7.2. Exploratory data analysis

Furthermore, we wanted to explore and hypothesize regarding other potential relationships between the variables. Therefore, an exploratory data analysis (EDA) (Tukey, 1977), specifically a principal component analysis (PCA) was conducted to explore whether there were any interesting patters or observations in the data collected. Here, no hypothesis was determined, and all effects described emerged post-hoc. The PCA was conducted using scikit-learn (Pedregosa et al., 2011) and Jupyter Lab Notebook (Kluyver et al., 2016).

Table 4
One-way repeated measured ANOVA F-test for performance (objective).

Variable Performance	N	Outliers	Normality	Sphericity	Epsilon (ε)	F-statistic	Sig.	Sample effect size	Population effect size [c]
Performance [number of hits]	57	Yes (1)[a]	Approx. normal[b]	, $\chi 2(2) = 25.7$, $p < 0.0001$	0.728	$F(1.46, 81.54) = 316.34$	p <.0001**	$\eta^2 = 0.850$	$\omega^2 = 0.787$
	56[d]	No	Approx. normal[b]	$\chi 2(2) = 26.4$, $p < 0.0001$	0.721	$F(1.44, 79.32) = 308.69$	p <.0001**	$\eta^2 = 0.849$	$\omega^2 = 0.786$

*: $p < 0.05$, **: $p < 0.01$.

a) There was one outlier in C2. Latency w/ PD, as assessed by visual inspection of a boxplot. SPSS Statistics defines outliers as values greater than 1.5 box-plots from the edge of the plot. This value (14 hits) is genuinely unusual, we know from the experiment and the data that this participant performed above average in all three conditions. We ran the analysis both with and without outliers, reporting both results.

b) Visual inspection of Normal Q-Q Plots and histograms for all three conditions.

c) Calculated according to Wickens and Keppel (2004).

d) outlier removed.

Table 5
Pairwise comparisons of performance (objective).

Variable	C1. Latency – C2. Latency w/PD			C1. Latency – C3. Baseline			C2. Latency w/PD – C3. Baseline		
	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]
Performance [number of hits]	-1.298	0.264	$p < 0.0001$**	-9.754	0.491	$p < 0.0001$**	-8.456	.471	$p < 0.0001$**

b: Adjustment for multiple comparisons: Bonferroni.
*: $p < 0.05$, **: $p < 0.01$.

Table 6
Pairwise comparisons subjective variables.

Variable	C1. Latency – C2. Latency w/PD			C1. Latency – C3. Baseline			C2. Latency w/PD –C3. Baseline		
	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]
Subjective Overall Workload	0.336	0.163	$p = 0.133$	1.775	0.141	$p = 0.000$**	1.439*	0.177	$p = 0.000$**
Mental Demand 0-10	0.158	0.235	$p = 1.000$	2.105	0.226	$p = 0.000$	1.947*	0.306	$p = 0.000$**
Physical Demand 0-10[f]	0.035	0.221	$p = 1.000$	0.702	0.227	$p = 0.009$	.667*	0.211	$p = 0.008$**
Temporal Demand 0-10	0.175	0.221	$p = 1.000$	0.456	0.236	$p = 0.176$	.281	0.288	$p = 1.00$
Subjective Performance 0-10	0.789	0.244	$p = 0.006$*	2.825	0.240	$p = 0.000$	2.035	0.212	$p = 0.000$**
Effort 0-10[f]	0.246	0.234	$p = 0.894$	1.351	0.213	$p = 0.000$	1.105	0.241	$p = 0.000$**
Frustration 0-10	0.679	0.283	$p = 0.059$	3.179	0.304	$p = 0.000$	2.500	0.306	$p = 0.000$**

b) Host hoc Pairwise comparisons were adjusted for Bonferroni.
f) A Friedman test with pairwise comparisons using a Bonferroni correction for multiple comparisons was carried out for comparison purposes. Results were corroborated.
*: $p < 0.05$, **: $p < 0.01$.

4. Results

The following section presents the results of the statistical tests before providing the result from the EDA.

4.1. Performance (objective)

A one-way repeated measures ANOVA was conducted to determine whether differences in human performance (the number of hits, an objective measure) between the three conditions were statistically significant. Descriptive statistics of performance data are illustrated in Fig. 5, and Table 5 shows all pairwise comparisons of the conditions. Table 4 contains the ANOVA F-test statistic, data characteristics and pretests.

Performance was statistically significant different in the three conditions, with a performance increase of 20% from C1. Latency to C2. Latency w/PD. Performance increased from $M = 6.2$ *hits* in C1. Latency, to $M = 7.5$ *hits* in C2. Latency w/PD, to $M = 16$ *hits* in C3. Baseline. There was a statistically significant increase in performance of $M = 1.3$ hits ($SD = 0.26$) from C1. Latency to C2. Latency w/PD. In summary, there was a statistically significant difference between means and, therefore, we accept the alternative hypothesis; The predictive display significantly increases performance of the human operator.

4.2. Subjective workload

This section presents the results from statistical analysis of subjective workload measures. Overall Subjective Workload is presented first, before also presenting the individual workload dimensions.

Since we conducted RTLX, the values of the individual workload dimensions (mental, physical, temporal, performance, effort and frustration) were averaged to obtain an estimate of the overall workload (Hart, 2006). This averaged score is addressed as Subjective Overall Workload in the following. Separate one-way repeated measures ANOVA was conducted for overall workload and the six individual workload dimensions to determine the effects of the predictive display on lowering subjective workload in the three conditions. The results from the ANOVA F-test, including pretests for all variables can be found in Table A2,

Appendix A, whereas descriptive statistics and pairwise comparisons can be found in Table A1,

Appendix A, and Table 6, respectively. The following paragraphs describe individual results before providing an overall explanation.

4.2.1. Subjective overall workload

Subjective Overall Workload was statistically significant different under the three conditions. There was a decreased subjective workload from $M = 5.3$ $(SD = 0.2)$ in C1. Latency, to $M = 4.9$ $(SD = 0.2)$ in C2. Latency w/PD, to $M = 3.5$ $(SD = 0.2)$ in C3. Baseline. Pairwise comparisons of the three conditions was carried out using the Bonferroni post hoc test, which revealed that the mean decrease in subjective workload from C1. Latency to C2. Latency w/PD was not statistically significant ($M = 0.35, SD = 0.16, p = 0.133$). There was a statistically significant mean decrease in subjective workload from C2. Latency w/PD to C3. Baseline ($M = 1.44$, $SD = 0.18$, $p < 0.001$), and from C1. Latency to C3. Baseline ($M = 1.775$, $SD = 0.14$ $p < 0.001$). A Friedman test produced corroborating results. Therefore, we cannot reject the null hypothesis and cannot accept the alternative hypothesis. The predictive display does not decrease human operators' subjective overall workload.

4.2.2. Mental demand (individual workload dimension)

Mental demand was statistically significantly different in the three conditions, however, post hoc analysis with a Bonferroni adjustment revealed that mental demand did not significantly decrease from C1. Latency to C2. Latency w/PD. There was a statistically significant decrease in mental demand from C1. Latency to C3. Baseline and from C2. Latency w/PD to C3. Baseline. The predictive display did not reduce participants' mental demand.

4.2.3. Physical demand (individual workload dimension)

Physical demand was statistically significantly different in the three conditions, however, post hoc analysis with a Bonferroni adjustment revealed that physical demand did not significantly decrease from C1. Latency to C2. Latency w/PD. A Friedman test with pairwise comparisons using a Bonferroni correction for multiple comparisons was carried out for comparison purposes, which gave the same result. There was a statistically significant decrease in physical demand from C1.

Table 7
One-way repeated measured ANOVA F-test subjective latency.

Variable	N	Outliers	Normality	Sphericity	Epsilon (ε)	F-statistic	Sig.	Sample effect size	Population effect size[c]
Subjective latency	57	Yes (13)[a] (10 unique)	Approx.[c]	Yes χ2(2) = 5.575, p = 0.062	-	F(2,112) = 45.734	$p < 0.001$**	$\eta^2 = 0.450$	$\omega^2 = 0.343$
	52[b]	4 unique	Approx.[c]	Yes χ2(2) = 3.617, p = 0.164	-	F(2,102) = 44.684	$p < 0.001$**	$\eta^2 = 0.467$	$\omega^2 = 0.359$

*: $p < 0.05$, **: $p < 0.01$.

a) Number of outliers in parentheses. There was no reason to exclude any outliers and so they were kept in first analysis. In the second, 5 extreme outliers were excluded, which yielded a dataset with 4 unique outliers. Further reduction did not yield a dataset without outliers. Both results are reported here.

b) 5 extreme outliers removed.

c) Visual inspection of Normal Q-Q Plots and histograms for all three conditions.

Latency to C3. Baseline, and from C2. Latency w/PD to C3. Baseline. The predictive display did not reduce participants' physical demand.

4.2.4. *Temporal demand (individual workload dimension)*

Temporal demand was not statistically significantly different in the three conditions, according to both ANOVA and Friedman test. The predictive display did not reduce participants' temporal demand.

4.2.5. *Subjective performance (individual workload dimension)*

Subjective Performance was statistically significantly different in the three conditions. Subjective Performance was evaluated at $M = 5.53$ in C1. Latency, $M = 4.74$ in C2. Latency w/PD, and $M = 2.70$ in C3. Baseline, with a low value corresponding to a performance closer to perfect. There was a statistically significant decrease of $M = 0.79$ *($SD = 0.24, p = 0.006$)* between C1. Latency and C2. Latency w/PD, a statistically significant decrease of $M = 2.83$ *($SD = 0.24, p < 0.001$)* between C1. Latency to C3. Baseline, and a statistically significant decrease of $M = 2.04$ *($SD = 0.21, p < 0.001$)* between C2. Latency w/PD to C3. Baseline. A Friedman test with a Bonferroni correction for multiple comparisons was carried out for comparison purposes, corroborating result at $p < 0.001$. The median of Subjective Performance was statistically significant different between C1. Latency ($Mdn = 5$) and C3. Baseline ($Mdn = 2$) ($p < 0.001$), statistically significant between C2. Latency w/PD ($Mdn = 5$) and C3. Baseline ($p < 0.001$), but not statistically significant different between C1. Latency condition and C2. Latency w/PD ($p < 0.132$). In addition to a statistically significant decrease from both latency conditions (C1. Latency and C2. Latency w/PD) to C3. Baseline, it is also noteworthy that the mean decrease towards C3. Baseline is greater from C1. Latency than the decrease from C2. Latency w/PD; Which means participants thought they performed better with the predictive display than without it, given different latencies, and given equal latency. In summary, the predictive display increased participants subjective performance, i.e. participants thought their performance was better with the predictive display.

4.2.6. *Effort (individual workload dimension)*

Effort was statistically significant different in the three conditions, however, post hoc tests with a Bonferroni adjustment revealed that there was not a statistically significant difference between C1. Latency and C2. Latency w/PD. There was a statistically significant decrease from C3. Baseline to the two latency conditions (C1. Latency and C2. Latency w/PD). A Freidman test with a Bonferroni correction for multiple comparisons corroborated these results. The predictive display did not reduce participants' effort.

4.2.7. *Frustration (individual workload dimension)*

Frustration was statistically significantly different in the three conditions and post hoc tests with a Bonferroni adjustment revealed that there was a statistically significant decrease in frustration from C1. Latency to C3. Baseline, as well as from C2. Latency w/PD to C3. Baseline, though not from C1. Latency to C2. Latency w/PD. The predictive display did not reduce participants' frustration.

4.2.8. *Overall result for workload*

The analysis of Subjective Overall Workload and the individual workload dimensions did not show a statistically significant difference between C1. Latency and C2. Latency w/PD, with the exception of the individual variable Subjective Performance, in which participants reported a statistically significant mean increase of 0.789.

The predictive display does not reduce participants' mental demand, physical demand, temporal demand, effort, nor frustration; However, the predictive display increased participants' subjective performance, i.e. participants' thought they performed better with the predictive display. In summary for subjective workload, we cannot reject the null hypothesis, i.e. the predictive display does not reduce participants subjective workload.

Table 8
Pairwise comparisons subjective latency.

Variable	C1. Latency – C2. Latency w/PD			C1. Latency – C3. Baseline			C2. Latency w/PD – C3. Baseline		
	Mean Diff.	SD	Sig.^{b}	Mean Diff.	SD	Sig.^{b}	Mean Diff.	SD	Sig.^{b}
Subjective Latency [ms]	78.33	68.42	$p = 0.771$	582.21	73.14	$p = 0.000$**	503.88	55.37	$p = 0.000$**

b) Host hoc Pairwise comparisons were adjusted for Bonferroni.
*$p < 0.05$, **: $p < 0.01$.

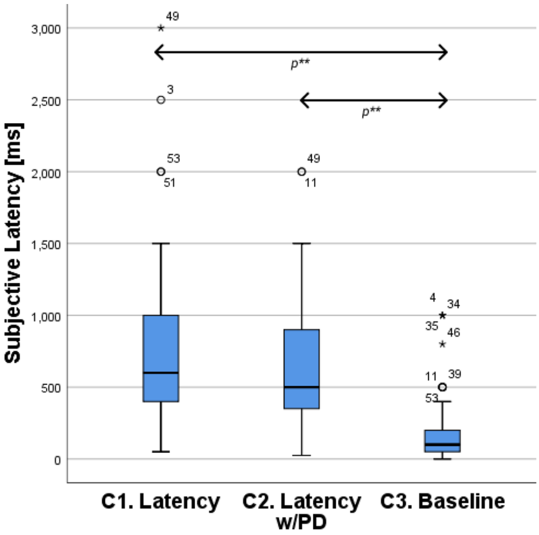


Fig. 6. Descriptive Statistics Subjective Latency. Original data reported. Statistically significant differences at p < 0.01 are indicated by *p***.

4.3. Subjective latency

A one-way repeated measures ANOVA was conducted on Subjective Latency to determine if there was a statistically significant reduction from C1. Latency – to C2. Latency w/PD condition. Subjective latency (evaluated in ms by the participants) was statistically significantly different in the three conditions. Post hoc tests with a Bonferroni adjustment revealed that there was a statistically significant decrease in Subjective Latency from C1. Latency to C3. Baseline condition as well as from C2. Latency w/PD to C3. Baseline condition, but not from C1. Latency to C2. Latency w/PD. A Friedman test corroborated these results, (Tables 7 and 8).

There were multiple outliers for this measure, and we reran the analysis with the extreme outliers removed, which only slightly increased effect size, however it did not change the overall result. The predictive display did not reduce participants' estimation of latency when comparing C1. Latency and C2. Latency w/PD, see Fig. 6.

4.4. Gamers

Those who play games weekly or more often were defined as gamers. The potential increased performance gain, measured objectively, by gamers is investigated here. A two-way mixed ANOVA was conducted, comparing the mean differences of performance (objective) between two independent groups, Gamers and Non-Gamers, under the three conditions. Descriptive statistics can be found in Table 9 and

Table 9
Descriptive statistics gamer vs non-gamer. Original data reported.

Variable	Group	N	C1. Latency		C2. Latency w/PD		C3. Baseline	
			Mean	SD	Mean	SD	Mean	SD
Performance [number of hits]	Gamer	17	6.47	0.42	8.41	0.46	17.71	0.92
	Non-Gamer	40	6.08	0.27	7.10	0.30	15.20	0.60

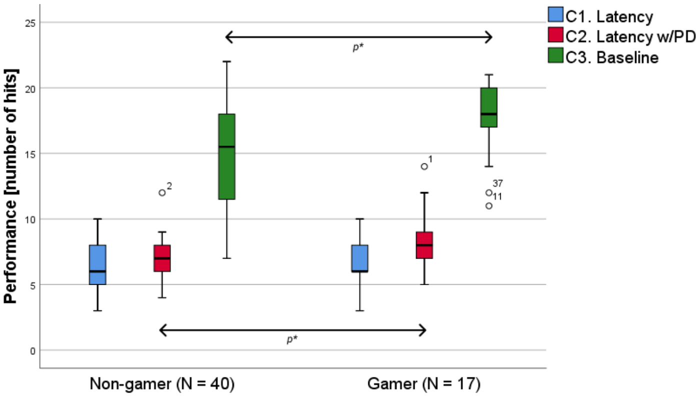


Fig. 7. Performance of Gamers vs. Non-gamer. Original data reported. Statistically significant differences at p < 0.01 are indicated by *p***.

Table 10
Two-way mixed ANOVA F-test on performance (objective) for gamers vs. non-gamer.

Variable	N	Outliers	Normality	Homogeneity	Sphericity	Epsilon	F-statistic	Sig.	Sample effect size	Population effect size [g]
Performance[a] [number of hits]	57	Yes (4)[bc]	Yes[d]	Yes[e]	No $\chi^2(2)$ = 24.895, p = .000	ε = 0.728	F(1.46, 80.32) = 2.72[f]	p = 0.088	η^2 = 0.047	ω^2 = 0.02
	53[h]	No	Yes[d]	Yes[e]	No $\chi^2(2)$ = 21.406, p = .000	ε = 0.742	F(1.48, 75.65) = 5.769[f]	p = 0.009**	η^2 = 0.102	ω^2 = 0.057

*: $p < 0.05$, **: $p < 0.01$.

a) Performance was separated for the two independent groups.

b) Assessed by visual inspection of a boxplot. SPSS Statistics defines outliers as values greater than 1.5 box-plots from the edge of the plot. The 4 outliers were kept in the first analysis as there was no reason to exclude them. In the second, they were excluded. Both results are reported here.

c) By examination of studentized residuals for values greater than ± 3, one outlier was found with a studentized residual value of 3.04. The outlier was kept in the subsequent analysis since its value is close to the threshold and as there was no reason to exclude it.

d) Visual inspection of Normal Q-Q Plots of the distribution and the distribution of studentized residuals for all three conditions.

e) Levene's test assessed homogeneity of variance, and Box's test evaluated homogeneity of covariances.

f) A Greenhause Geisser correction was applied.

g) Calculated according to Wickens and Keppel (2004).

h) outlers removed.

Fig. 7. The ANOVA F-test can be found in Table 10.

The interaction between gaming experience and conditions on performance had a level of significance of $p = 0.088$. Univariate post hoc tests indicated that there was not a statistically significant difference between gamers (*M = 6.5 hits*) and non-gamers (*M = 6.1 hits*) in C1. Latency (*F(1,55) = 0.622, p = 0.43*, sample effect size $\eta^2 = 0.01$). However, there was a significant increase in performance for gamers in C2. Latency w/PD (*F(1,55) = 5.71, p = 0.02*, sample effect size $\eta^2 = 0.094$), in which gamers had *M = 8.4 hits*, whereas non-gamers had *M = 7.1 hits*. Furthermore, there was a significant increase in performance for gamers in C3. Baseline (*F(1,55) = 5.203, p = 0.026*, sample effect size $\eta^2 = 0.086$), in which gamers had *M = 17.7 hits*, whereas non-gamers had *M = 15.2 hits, (Tables 10 and 11)*.

When considering the two independent groups (Gamer, Non-Gamer), there was a significant main effect of gaming (*F (1,55) = 6.311, p = 0.015,* sample size effect $\eta^2 = 0.103$), with gamers performing better than non-gamers. Gamers performed on average *M = 10.9 hits*, which is *M =1.4 hits (SD = 0.6)* above the performance of non-games with *M = 9.5 hits*.

The analysis was also conducted without outliers (see Table 11), which yielded more than a doubling of effect size (sample effect size η^2 = 0.102 and population effect size ω^2 = 0.057), and a lower p value (p = 0.009), which means that the interaction between gaming experience and conditions on objective performance reached statistical significance. Univariate post hoc tests (on the pruned dataset) indicated a statistically insignificant difference between gamers (*M = 6.3 hits*) and non-gamers (*M = 6.0 hits*) in C1. Latency (*F(1,51) = 0.240, p = 0.63*, sample effect size $\eta^2 = 0.005$), and a statistically insignificant difference increase in performance for gamers in C2. Latency w/PD (*F(1,51) = 3.52, p = 0.066*, sample effect size $\eta^2 = 0.065$), in which gamers had a *M = 7.9 hits*, whereas non-gamers had *M = 6.0 hits.* There was a significant increase in performance for gamers in C3. Baseline (*F(1,51) = 8.249, p = 0.006*, sample effect size $\eta^2 = 0.1.39$), in which gamers achieved *M = 18.4 hits*, whereas non-gamers had *M = 15.1 hits.* When considering the two independent groups (Gamer, Non-Gamer), there was a significant main effect of gaming (*F (1,51) = 6.929, p = 0.011,* sample size effect $\eta^2 = 0.12$), with gamers performing better than non-gamers. Gamers performed on average *M = 10.9* hits, which is *M =1.5 hits (SD = 0.6)* above the performance of non-games with *M = 9.4 hits.*

Gamers performed better than non-gamers on average.

4.5. *Exploratory data analysis - PCA*

A principal component analysis (PCA) was conducted to explore whether there were any interesting patters or observations in the data collected. We had no predetermined hypothesis, and all effects described in this section emerged post-hoc.

A total of 35 variables collected during the experiment were standardized (removing the mean and scaling to unit variance) and used in the PCA. Fig. 8 shows a Scree plot of the Principal Components (PCs) eigenvalues. The first 10 eigenvalues are larger than 1, the first 5 have an eigenvalue above 2, the first two are greater than 4 and the first eigenvalue is greater than 7. Fig. 9 shows the cumulative sum of explained variance, which did not have a clear ‘elbow-shape’, however the first 7-10 PCs retains 72.1% – 82.8%[3] of the variance of the original data.

A Score plot, and a Loading plot of the two first Principal Components (PCs) can be found in Figs. 10 and 11, and 12 respectively. The first two PCs explains 22.6% and 13.6% of the total variance. The following result emerged post-hoc, and thus interpretations made accordingly.

4.5.1. *Interpreting the Score plot*

Fig. 11. Score Plot of the first two principal components with gaming

[3] Accurate percentage of explained variance retrieved from data, and not estimated from graph.

Table 11
Pairwise comparisons of differences in performance (objective) for gamer vs. non-gamer.

Variable	C1. Latency			C2. Latency w/PD			C3. Baseline			
Performance [number of hits]	Mean Diff.	SD	Sig.^{b}	Mean Diff.	SD	Sig.^{b}	Mean Diff.	SD	Sig.^{b}	
Gamer – Non-Gamer (outliers removed)	57	.396	.501	*p = 0.434*	1.312	.549	*p = 0.020**	2.506	1.099	*p = 0.026**
	53	0.260	0.53	*p = 0.626*	0.954	0.509	*p = 0.066*	3.300	1.149	*p = 0.006***

* $p < 0.05$, **: $p < 0.01$.
a) outliers removed.
b) Host hoc Pairwise comparisons were adjusted for Bonferroni.

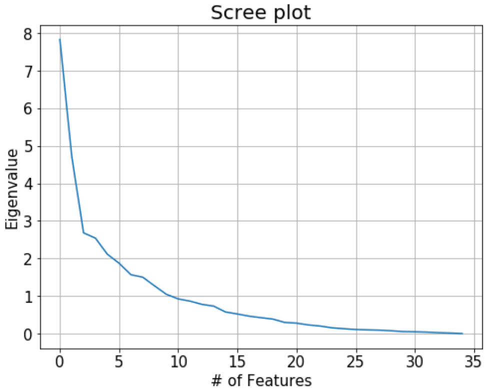


Fig. 8. Scree plot.

experience.Fig. 10. shows the Score Plot of PC1 and PC2 with legends indicating gender, from which we can see that PC2 tend to separate woman and men quite accurately. The woman cluster in the positive range of PC2 and the men in the negative range, with only a few datapoints crossing zero. Fig. 11 show the Score plot with gaming experience, in which we observe subtle trends in the scatterplot based on gaming experience; Those who never gamed predominantly resides in the positive range of PC2; Furthermore those who games more often tended to cluster in the negative range of PC2. When viewing Figs. 10 and 11 simultaneously we observe that the women in this experiment typically gamed yearly or never, with two exceptions of woman gaming on a monthly basis. Men gamed most often, typically yearly, monthly, weekly and two participants daily.

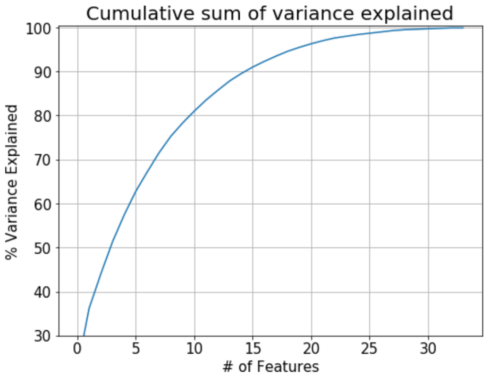


Fig. 9. Cumulative sum of variance explained.

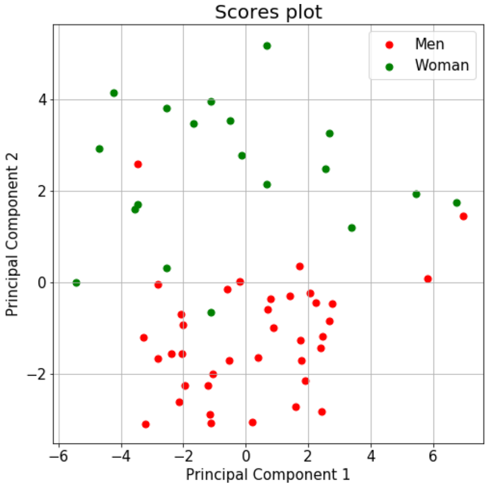


Fig. 10. Score Plot of the first two principal components gender.

4.5.2. Interpreting the Loading plot

From the loading plot in Fig. 11, we see that the total hits in each of the conditions seem to be correlated as they cluster together. Total hits in each condition (Total hits C1 – C3) and total hits for all conditions combined (Total hits C1 + C2 + C3) cluster together, as does subjective performance (Performance C1 – C3). We observe that eye health and computer usage have a loading close to zero, thus not contributing to the definition of the principal components, and unimportant for defining the direction of some underlying latent variable. All participants used a computer daily, thus this variable had the same value across the participant population. The eye health, level of education, key strokes in C1. Latency (Key strokes C1) and C2. Latency w/PD (Key strokes C2), and age also have a loading

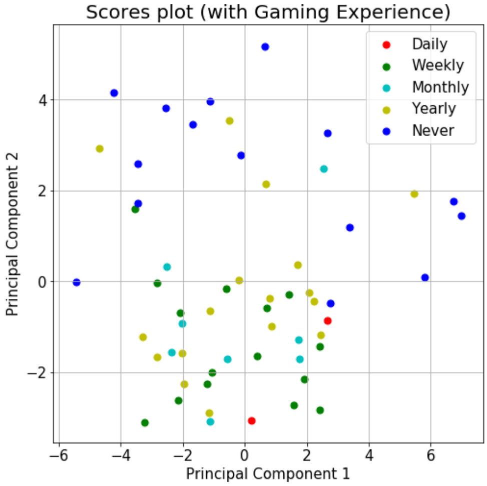


Fig. 11. Score Plot of the first two principal components with gaming experience.

close to zero and are less important for the model.

The subjective performance is negatively correlated with the other TLX dimensions, especially noteworthy is the opposite positions of subjective performance (Performance C1 – C3) and frustration (Frustration C1 – C3). The hits in each condition, and the total hits for all three conditions are clustered together and are therefore correlated. Furthermore, we see that gender and gaming experience have a high loading on PC2, thus contributing greatly to defining PC2 (that gender contributed to PC2 we also knew from the scores plot), and that they are positioned quite close together in comparison to the other variables and therefore are correlated. Among these participants men tended to game more than woman, which is also reflected when looking at the raw data.

5. Discussion

5.1. Performance

The results show that there is a statistical difference in performance when controlling the ROV without and with the help of the predictive display. Subjects performed on average 20% better with a sample effect size $\eta^2 = 0.850$, and population effect size of $\omega^2 = 0.787$. This can be categorized as a medium to large effect (Kirk, 2013), especially when considering the simplicity and low cost of implementing the predictive display. Previous research describes a wide range (8% to 65%) of task time reduction from predictive technology. A direct comparison to any specific

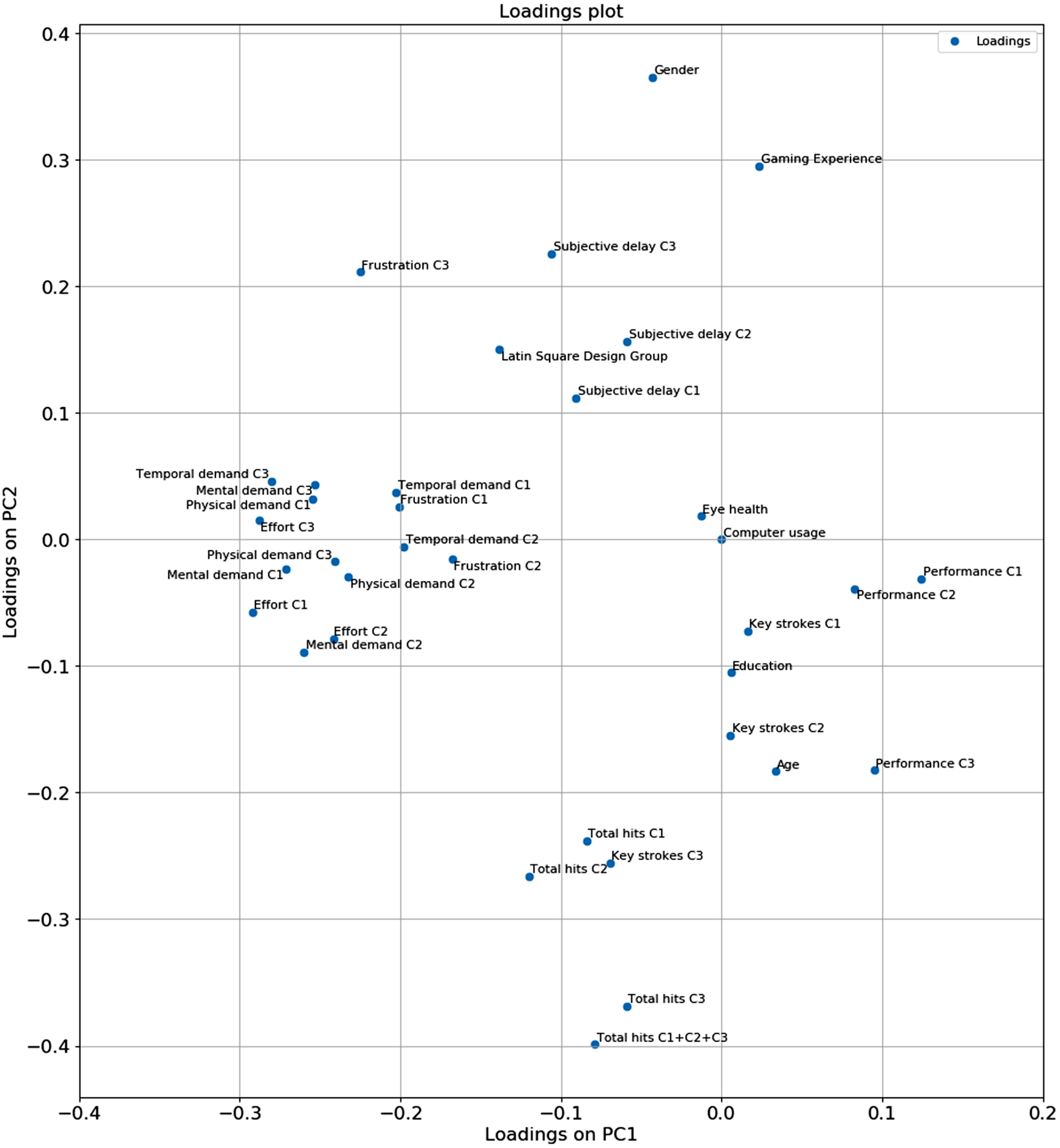


Fig. 12. Loadings plot.

experiment is challenging, however a performance increase of 20% in this experiment is probably in the lower range of what was found in the other experiments in Table 1. The task time reduction measure is considered to be comparable to the performance gain measured in this experiment. However, the predictive method used here is the simplest solution to implement at the lowest cost. Moreover, the participants only had 30 seconds to intuitively learn and train in using the predictive display, since none of the participants were told that there would be a predictive display nor how it worked. Some immediately understood what the predictive display was trying to tell them, others did not understand that there had been a predictive display until the experiment was over. The ones who tried to use the predictive display the way it was intended typically performed better than those who did not use it. It may be that performance could have been improved more if participants were informed about the predictive display's functionality.

As expected, participants performed significantly better in the baseline condition, in which there was only 250 ms latency. This latency is well above what human perception is able to pick up on, which most participants did. As discussed under 5.4, most participant underestimated the latency in the third condition reporting barely above 0 ms.

5.2. Subjective workload

Subjects reported minimal differences between C3. Baseline, C1. Latency and C2. Latency w/PD. There was a statistically significant difference in subjective overall workload between the three conditions, however Bonferroni post hoc tests revealed that differences between C1. Latency and C2. Latency w/PD was not statistically significant. Therefore, we cannot say that the predictive display reduces subjective workload. The only significant difference was found in subjective performance, in which participants felt that they on average performed 14% better when using the predictive display. The actual performance increase was 20%. They also reported that they felt 11% less frustrated using the PD, though this is not statistically significant. Participants also stated that C3. Baseline was better in all metrics, with an exception of temporal demand where the difference was not significant.

Participants reported no significant difference in mental, physical and temporal demand between C1. Latency and C2. Latency w/PD. We consider these three metrics to be a good description of the total subjective workload in this experiment setup. Some participants, especially those who did not understand what the predictive display was trying to tell them, even reported it as distracting. Due to the predictive display's functionality, the video feed is constantly moving around and scaling up and down. This can understandably be distracting. Some participants immediately understood how the predictive display worked, and they typically reported the predictive display as helpful. To the experimenter they also seemed to be more relaxed, however there are no recorded data illustrating this. During the task, a red timer indicating the remaining time was constantly visible for participants to see in the upper right corner. In addition, the ROV had rapid acceleration and was able move fast if the operator managed to do so. Overall, this made for a hectic and exiting experience for the subjects. This may explain why there is no significant change in the temporal demand, even compared to C3. Baseline. The fact that the participants reported a better value (smaller) in the other five metrics for the no delay condition, is as expected. The experimenter also observed a tendency of participants performing correcting steering commands, causing the ROV to oscillate greatly before hitting or missing the target, which corroborates prior research (Appelqvist et al., 2007). This was particularly prominent in the C1. Latency condition, again illustrating the detrimental effect of latency on both human performance and behavior. These findings support earlier research describing how video latency negatively affects the user experience in teleoperation.

5.3. Gaming

The gamers performed 30% better with the predictive display, while non-gamers performed 17% better. Interestingly the gamers increased their score almost twice as much as non-gamers when shifting from C1. Latency to C2. Latency w/PD, though the exact reason for this is unclear. The arrow in the predictive display acts as an aiming device, which could be a more familiar concept for gamers. This finding could also indicate that gamers are more used to having to adapt to unfamiliar setting and interfaces in a computer competing context. Furthermore, when comparing the scores of gamers and non-gamers, it is interesting to note that gamers only performed better than non-gamers in C2. Latency w/PD and C3. Baseline, but not in C1. Latency. This could indicate that the amount of experience may not be crucial for obtaining a high score (equal to high performance) in a situation with considerable latency. Thought post hoc tests on the pruned dataset were not statistically significant at $p<0.05$ in C2. Latency w/PD, the level of significance $p = 0.066$ was close to that threshold. A level of statistical significance may have been achieved with additional participants conducting the experiment, and equal group sizes, as both may have a large effect on p-values (Krishnaiah, 1980). In both analysis, there was a significant main effect of gaming, meaning gamers performed better on average. More interesting is the population effect size, which increased from $\omega^2 = 0.02,$ a small association to $\omega^2 = 0.057,$ a medium association (Kirk, 2013), which means that the effect of gaming, and the ability gamers had to take advantage of the PD, reaches some practical significance. Taken together, we interpret this to mean gamers were better able to take advantage of the predictive display to increase objective performance.

We observe that the combination of predictive display and related training (in the form of playing similar games at least once a week) results in twice a performance gain compared to only predicative display. In this experiment participants were not informed of the predictive display's functionality, which leads us to consider what the performance gain might have been if participants' were aware of the functionality a priori and if they received training in using the predictive display. Simultaneously considering an increased effect size when removing outliers, i.e. a stronger result, leads us to believe that a greater performance gain might have been the result of specialized training prior to the experiment. Therefore, we hypothesize that the combination of predictive display and extensive training produces a greater increase in performance. Research corroborates this; A priori gaming experience have been found to relate to performance in desktop and immersive virtual environments (Richardson et al., 2011), and video gaming suggested a s a training regimen to increase processing speed, which contributes to increased cognitive performance (Dye et al., 2009). Moreover, studies investigating causality supports action video gaming as a training method (Dye et al., 2009; Green & Bavelier, 2003; Richardson et al., 2011). Generally, we hypothesize that assistive technology in combination with (potentially minimal) training produces high performance gain (output). When compared to the necessary implementation of technology and training (input), we consider this a good trade-off between input and output.

5.4. Subjective latency

About 75% of the participants underestimated the latency in the third condition. Many of them barely reported over 0 ms, but the actual latency was 250 ms. These findings support previous research, which states that smaller latencies closer to zero is difficult to differentiate from no latency. Questioning participants about latency could have influenced their attention to latency in the forthcoming conditions. However, the randomized Latin Square Design of conditions should account for any order effects caused by this question. Furthermore, this question was primarily included to investigate whether participants experienced lower latency with the aid of the predictive display when comparing conditions with equal latencies, which was not the case. The predictive display did not decrease the subjectively experienced latency for participants in this experiment.

5.5. Exploratory data analysis discussion

Effects discussed here emerged post hoc; Thus, is interesting to see effects of gender and gaming experience show up in the PCA, since there are known effects of both. From the scores plots (Figs. 10 and 11) we see that PC2 separates women and men quite accurately with a few exceptions. Furthermore, PC2 tends to separate participants by their gaming experience, and by combining the loadings plot (Fig. 11) and scores plot (Figs. 10 and 11) we observe that the male participants, the exceptions in the upper regions of PC2, never gamed. When further investigating the loadings plot (Fig. 11) and scores plots (Figs. 10 and 11) simultaneously we see that gender and gaming both had high loadings on PC2, thus contributing to PC2. In the loadings plot (Fig. 11) we see participants objective performance (Total hits C1 – C3) having a high negative loading, which means it also contributes to the definition to PC2. Males are generally more experienced in gaming (Richardson et al., 2011), and in both studies investigated by Richardson et al. (2011) high gaming experience was related to higher task performance. Video gaming involves several spatial and cognitive abilities, and studies investigating causation show that gaming experience can improve mental rotation and visual attention (Moffat et al., 1998; Richardson et al., 2011). For instance, performance in visual search tasks, visual attention, visual memory, contrast sensitivity, and judging relative velocity have all been shown to improve with gaming experience (Dye et al., 2009; Moffat et al., 1998; Richardson et al., 2011). Performance in dynamic spatial tasks that required reasoning about moving stimuli (e.g. tracking objects) also improved (Richardson et al., 2011); And all those abilities are important for a high objective performance (Total hits C1 – C3) in this experiment. When specifically considering spatial abilities, there are known gender differences, including visuospatial abilities such as spatial orientation and spatial visualization (Moffat et al., 1998). Males outperform females in spatial performance tasks; In particular when it involves mental rotations, whether that task is paper-and-pencil (manipulations and transformations of geometric figures and forms) or in a virtual environment (Moffat et al., 1998; Richardson et al., 2011). Since males generally have more gaming experience than females and video game experience influence visuospatial processes, this might further contribute to gender differences in spatial tasks (Richardson et al., 2011), and moreover the objective performance (Total hits C1 – C3) in this experiment. In fact, females and males with similar levels of gaming experience did not differ in dynamic spatial ability, and gender differences were eliminated when gaming experience was included as a covariate (Richardson et al., 2011). Since the females in our experiment generally had less gaming experience, and those who did tended to cluster towards the male gamers, and since non-gaming males tended to cluster towards the females, we therefore identify an effect of gaming experience. We do recognize the high collinearity between gender and gaming experience, both had a high loading on PC2 (Fig. 11); However, further analysis is needed to examine what exactly separates the data here. Still, PC2 consists mainly of objective measures, e.g. gender, gaming experience, and objective performance (Total hits C1 – C3). For PC1, we have high loadings on individual workload dimensions (which are subjective), in which all are correlated except for subjective performance, and so they contribute to the definition of PC1. In summary, PC1 consists mainly of subjective variables from surveys, whereas PC2 consists mainly of objective variables collected in the experiment.

6. Conclusion – An increase in human performance

This work investigated human operators' performance and their subjectively experienced workload in a teleoperation context when using a predictive display. Human operator performance decrease and workload increase as latency is introduced in teleoperation, but there exist several approaches to combat these detrimental effects; One of which is predictive technology. A predictive display based on image transformation was developed by applying positional and scale transformations to the video feed and tested experimentally. An experiment was set up to test the predictive display and investigate changes in human operator performance and workload when operating an ROV. N = 57 participants conducted a simple navigational task (peg-in-hole game), under three conditions: C1. Latency, C2. Latency with predictive display and C3. Baseline. ANOVAs showed a statistically significant increase of 20% in human performance with the aid of the predictive display. Differences in overall subjective workload was not statistically significant, except for with subjective performance where participants felt they performed better with the predictive display. Gaming experience was advantageous, in fact gamers increased their score with almost twice as much as non-gamers. An exploratory data analysis (EDA) investigated confounding factors with confirmatory results.

CRediT authorship contribution statement

Henrikke Dybvik: Writing - original draft, Writing - review & editing, Visualization, Formal analysis, Validation. **Martin Løland:** Software, Investigation, Data curation, Formal analysis, Conceptualization, Methodology. **Achim Gerstenberg:** Supervision. **Kristoffer Bjørnerud Slåttsveen:** Supervision. **Martin Steinert:** Supervision.

Declaration of Competing Interest

None.

Acknowledgements

This research is supported by strategic funds from the department for Mechanical and Industrial Engineering (MTP) at NTNU.

Appendix A

Tables A1, A2.

Table A1
Descriptive statistics subjective variables.

Variable	C1. Latency		C2. Latency w/PD		C3. Baseline	
	Mean	SD	Mean	SD	Mean	SD
Subjective Overall Workload	5.263	0.197	4.927	0.192	3.488	0.193
Mental Demand 0-10	5.667	0.273	5.509	0.301	3.561	0.271
Physical Demand 0-10	2.877	0.285	2.842	0.293	2.175	0.245
Temporal Demand 0-10	5.842	0.277	5.667	0.280	5.386	0.307
Subjective Performance 0-10[1)]	5.526	0.307	4.737	0.274	2.702	0.214
Effort 0-10	6.018	0.260	5.772	0.266	4.667	0.278
Frustration 0-10	5.625	0.319	4.946	0.275	2.446	0.243

Table A2
One-way repeated measured ANOVA F-test.

Variable	N	Outliers[a, e]	Normality	Sphericity	Epsilon (ε)	F-statistic	Sig.	Sample effect size	Population effect size[c]
Subjective Overall Workload	57	Yes (1)	Yes	Yes $\chi^2(2) = 3.787$, $p = 0.151$	-	$F(2, 112) = 68.322$	$p < 0.001$**	$\eta^2 = 0.55$	$\omega^2 = 0.441$
	56[i]	No	Yes	Yes $\chi^2(2) = 3.999$, $p = 0.135$	-	$F(2, 110) = 68.311$	$p < 0.001$**	$\eta^2 = 0.55$	$\omega^2 = 0.445$
Mental demand	57	No	Yes[d]	No $\chi^2(2) = 9.962198$, $p = 0.007$	$\varepsilon = 0.858$[b]	F(1.716, 96.082) = 41.286	$p < 0.001$**	0.424	$\omega^2 = 0.32$
Physical demand	57	Yes (1) [f]	Yes[d]	Yes $\chi2(2) = 0.357$, $p = 0.837$.	-	F(2, 112) = 6.474	$p = 0.002$**	$\eta^2 = 0.104$	$\omega^2 = 0.060$
	56[i]	No	Yes[d]	Yes $\chi2(2) = 1.198$, $p = 0.549$.	-	F(2, 110) = 5.601	$p = 0.005$**	$\eta^2 = 0.092$	$\omega^2 = 0.052$
Temporal demand	57	Yes (2) [f]	Yes[d]	No $\chi2(2) = 6.498$, $p = 0.039$.	$\varepsilon = 0.900$[b]	F(1.799, 100.771) = 1.690	$p = 0.192$	$\eta^2 = 0.029$	$\omega^2 = 0.008$
	55[i]	No	Yes[d]	No $\chi2(2) = 6.504$, $p = 0.039$.	$\varepsilon = 0.896$	F(1.793, 96.819) = 1.686	$p = 0.193$	$\eta^2 = 0.030$	$\omega^2 = 0.008$
Subjective performance	57	Yes (2) [f]	Yes[d]	Yes $\chi2(2) = 1.552$, $p = 0.460$.	-	F(2, 112) = 78.578	$p < 0.001$**	$\eta^2 = 0.584$	$\omega^2 = 0.476$
	55[i]	Yes (3)[g]	Yes[d]	Yes $\chi2(2) = 1.972$, $p = 0.373$.	-	F(2, 108) = 81.030	$p < 0.001$**	$\eta^2 = 0.600$	$\omega^2 = 0.492$
Effort	57	Yes (1) [f]	Yes[d]	Yes $\chi2(2) = 1.143$, $p = 0.565$	-	$F(2, 112) = 19.641$	$p < 0.001$**	$\eta^2 = 0.260$	$\omega^2 = 0.179$
	56[i]	No	Yes[d]	Yes $\chi2(2) = 1.277$, $p = 0.528$	-	$F(2, 110) = 18.627$	$p < 0.001$**	$\eta^2 = 0.253$	$\omega^2 = 0.173$
Frustration	56	No[h]	Yes[d]	Yes $\chi2(2) = 0.519$, $p = 0.771$	-	F(2,112) = 63.275	$p < 0.001$**	$\eta^2 = 0.535$	$\omega^2 = 0.426$

*:$p < 0.05$, **: $p < 0.01$.

a) Number of outliers in parentheses.

b) A Greenhause Geisser correction was applied.

c) Calculated according to Wickens and Keppel (2004).

d) Visual inspection of Normal Q-Q Plots and histograms for all three conditions.

e) Visual inspection of a boxplot.

f) Outliers was kept in the first ANOVA as there was no reason for excluding them and a Friedman test with pairwise comparisons using a Bonferroni correction was carried out for comparison purposes, as this test is less affected by outliers. Results were corroborated. We also reran the analysis with outliers excluded, which resulted in somewhat higher effect size. The overall result was the same.

g) Excluding initial outliers did not yield a dataset without outliers. Further outlier removal was not conducted to avoid constructing a highly reduced, and thus unrepresentative dataset.

h) There were three outliers in the sample with N = 57. One outlier in C2. Latency w/PD, reported the highest frustration while feeling like their performed the worst with PD. We assume this was due to not understanding what PD was trying to do and thus we removed this participant from the analysis of Frustration. The two outliers in C3. Baseline reported high frustration in all three conditions and were therefore kept. When excluding the abovementioned participant and rerunning the analysis, there were not outliers in the data. N = 56 data points were used for this specific analysis.

i) outlier removed.

References

Allison, R.S., Harris, L.R., Jenkin, M., Jasiobedzka, U., Zacher, J.E., 2001. Tolerance of temporal delay in virtual environments. Proc. IEEE Virtual Reality 2001, 247–254. https://doi.org/10.1109/VR.2001.913793.

Ando, N., Lee, J.-H., Hashimoto, H., 1999. A study on influence of time delay in teleoperation—quantitative evaluation on time perception and operability of human operator. In: IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics. 5. pp. 1111–1116. https://doi.org/10.1109/ICSMC.1999.815712. Cat. No.99CH37028), 5.

Appelqvist, P., Knuuttila, J., Ahtiainen, J., 2007. Development of an Unmanned Ground Vehicle for task-oriented operation—considerations on teleoperation and delay. In: 2007 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pp. 1–6. https://doi.org/10.1109/AIM.2007.4412567.

Baldwin, J., Basu, A., Zhang, H., 1999. Panoramic video with predictive windows for telepresence applications. In: Proceedings 1999 IEEE International Conference on Robotics and Automation. 3. pp. 1922–1927. https://doi.org/10.1109/ROBOT.1999.770389. Cat. No. 99CH36288C), 3.

Bejczy, A.K., Kim, W.S., Venema, S.C., 1990. The phantom robot: predictive displays for teleoperation with time delay. In: IEEE International Conference on Robotics and Automation Proceedings. 1. pp. 546–551. https://doi.org/10.1109/ROBOT.1990.126037.

Chen, J.Y.C., Haas, E.C., Barnes, M.J., 2007. Human performance issues and user interface design for teleoperated robots. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) 37 (6), 1231–1245. https://doi.org/10.1109/TSMCC.2007.905819.

Davis, J., Smyth, C., McDowell, K., 2010. The effects of time lag on driving performance and a possible mitigation. IEEE Trans. Rob. 26 (3), 590–593. https://doi.org/10.1109/TRO.2010.2046695.

Dorais, G., Bonasso, R. P., Kortenkamp, D., Pell, B., & Schreckenghost, D. (1999). Adjustable autonomy for human-centered autonomous systems. 16–35.

Draper, J.V., Kaber, D.B., Usher, J.M., 1998. Telepresence. Hum. Factors 40 (3), 354–375. https://doi.org/10.1518/001872098779591386.

Dye, M.W.G., Green, C.S., Bavelier, D., 2009. Increasing speed of processing with action video games. Curr. Dir. Psychol. Sci. 18 (6), 321–326. https://doi.org/10.1111/j.1467-8721.2009.01660.x.

ENDSLEY, M.R., 1999. Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics 42 (3), 462–492. https://doi.org/10.1080/001401399185595.

Fabrizio, M.D., Lee, B.R., Chan, D.Y., Stoianovici, D., Jarrett, T.W., Yang, C., Kavoussi, L.R., 2000. Effect of time delay on surgical performance during telesurgical manipulation. J. Endourol. 14 (2), 133–138. https://doi.org/10.1089/end.2000.14.133.

Fitts, P.M., 1954. The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. 47 (6), 381. https://doi.org/10.1037/h0055392.

Goodrich, M.A., Olsen, D.R., Crandall, J.W., Palmer, T.J., 2001. Experiments in adjustable autonomy. In: Proceedings of IJCAI Workshop on Autonomy, Delegation and Control: Interacting with Intelligent Agents, pp. 1624–1629.

Green, C.S., Bavelier, D., 2003. Action video game modifies visual selective attention. Nature 423 (6939), 534–537. https://doi.org/10.1038/nature01647.

Hart, S.G., 2006. Nasa-task load index (NASA-TLX); 20 years later. Proc. Hum. Factors Ergonomics Soc. Ann. Meet. 50 (9), 904–908. https://doi.org/10.1177/154193120605000909.

Hart, S.G., Staveland, L.E., 1988a. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183.

Hart, S.G., Staveland, L.E., 1988b. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183.

Hendy, K.C., Hamilton, K.M., Landry, L.N., 1993. Measuring subjective workload: when is one scale better than many? Hum. Factors 35 (4), 579–601.

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklade, A.L., Christ, R.E., 1992. Comparison of four subjective workload rating scales. Hum. Factors 34 (4), 429–439. https://doi.org/10.1177/001872089203400405.

Hu, H., Perez, C., Sun, H., Jagersand, M., 2016. Performance of predictive display teleoperation under different delays with different degree of freedoms. In: 2016 International Conference on Information System and Artificial Intelligence (ISAI), pp. 380–384. https://doi.org/10.1109/ISAI.2016.0087.

IBM SPSS Statistics 25. (2017). IBM Corp.

Kirk, R., 2013. Experimental Design: Procedures for the Behavioral Sciences. SAGE Publications, Inc. https://doi.org/10.4135/9781483384733.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Loizides, F., Schmidt, B. (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, pp. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87.

Krishnaiah, P.R., 1980. Analysis of *Variance* 1 North-Holland.

Lane, J.C., Carignan, C.R., Sullivan, B.R., Akin, D.L., Hunt, T., Cohen, R., 2002. Effects of time delay on telerobotic control of neutral buoyancy vehicles. In: Proceedings 2002 IEEE International Conference on Robotics and Automation. 3. pp. 2874–2879. (Cat. No. 02CH37292), 3. https://doi.org/10.1109/ROBOT.2002.1013668.

Lim, J., Wu, W., Wang, J., Detre, J.A., Dinges, D.F., Rao, H., 2010. Imaging brain fatigue from sustained mental workload: an ASL perfusion study of the time-on-task effect. Neuroimage 49 (4), 3426–3435. https://doi.org/10.1016/j.neuroimage.2009.11.020.

Lovi, D., Birkbeck, N., Herdocia, A.H., Rachmielowski, A., Jägersand, M., Cobzaş, D., 2010. Predictive display for mobile manipulators in unknown environments using online vision-based monocular modeling and localization. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5792–5798. https://doi.org/10.1109/IROS.2010.5649522.

Lu, S., Zhang, M.Y., Ersal, T., Yang, X.J., 2018. Effects of a delay compensation aid on teleoperation of unmanned ground vehicles. In: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 179–180. https://doi.org/10.1145/3173386.3177064.

Luck, J.P., McDermott, P.L., Allender, L., Russell, D.C., 2006. An investigation of real world control of robotic assets under communication latency. In: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction, pp. 202–209. https://doi.org/10.1145/1121241.1121277.

Lum, M.J.H., Rosen, J., King, H., Friedman, D.C.W., Lendvay, T.S., Wright, A.S., Sinanan, M.N., Hannaford, B., 2009. Teleoperation in surgical robotics – network latency effects on surgical performance. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6860–6863. https://doi.org/10.1109/IEMBS.2009.5333120.

Ma, R., Kaber, D.B., 2006. Presence, workload and performance effects of synthetic environment design factors. Int. J. Hum. Comput. Stud. 64 (6), 541–552. https://doi.org/10.1016/j.ijhcs.2005.12.003.

MacKenzie, I.S., Ware, C., 1993. Lag As a determinant of human performance in interactive systems. In: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, pp. 488–493. https://doi.org/10.1145/169059.169431.

Mathan, S., Hyndman, A., Fischer, K., Blatz, J., Brams, D., 1996. Efficacy of a predictive display, steering device, and vehicle body representation in the operation of a lunar vehicle. In: Conference Companion on Human Factors in Computing Systems, pp. 71–72. https://doi.org/10.1145/257089.257147.

Matheson, A., Donmez, B., Rehmatullah, F., Jasiobedzki, P., Ng, H.-K., Panwar, V., Li, M., 2013. The effects of predictive displays on performance in driving tasks with multi-second latency: aiding tele-operation of lunar rovers. Proc. Hum. Factors Ergonomics Soc. Ann. Meet. 57 (1), 21–25. https://doi.org/10.1177/1541931213571007.

Maxwell, S.E., Delaney, H.D., 2003. Designing Experiments and Analyzing Data: a Model Comparison Perspective. Routledge.

Miller, D.P., Machulis, K., 2005. Visual aids for lunar rover tele-operation. In: Battrick, R. (Ed.), Proceedings of 8th International Symposium on Artificial Intelligence, Robotics and Automation in Space. ESA Publishing, Noordwijk, Netherlands.

Moffat, S.D., Hampson, E., Hatzipentelis, M., 1998. Navigation in a "virtual" maze: sex differences and correlation with psychometric measures of spatial ability in humans. Evol. Hum. Behav. 19 (2), 73–87. https://doi.org/10.1016/S1090-5138(97)00104-9.

Neumeier, S., Wintersberger, P., Frison, A.-K., Becher, A., Facchi, C., Riener, A., 2019. Teleoperation: the holy grail to solve problems of automated driving? sure, but latency matters. In: Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 186–197. https://doi.org/10.1145/3342197.3344534.

Nielsen, C.W., Goodrich, M.A., Ricks, R.W., 2007. Ecological interfaces for improving mobile robot teleoperation. IEEE Trans. Rob. 23 (5), 927–941. https://doi.org/10.1109/TRO.2007.907479.

Oboe, R., Fiorini, P., 1998. A design and control environment for internet-based telerobotics. Int. J. Robot. Res. 17 (4), 433–449. https://doi.org/10.1177/027836499801700408.

Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2008. Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Mak. 2 (2), 140–160. https://doi.org/10.1518/155534308X284417.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Perez, M., Xu, S., Chauhan, S., Tanaka, A., Simpson, K., Abdul-Muhsin, H., Smith, R., 2016. Impact of delay on telesurgical performance: study on the robotic simulator dV-Trainer. Int. J. Comput. Assist. Radiol. Surg. 11 (4), 581–587. https://doi.org/10.1007/s11548-015-1306-y.

Rachmielowski, A., Birkbeck, N., Jägersand, M., 2010. Performance evaluation of monocular predictive display. In: 2010 IEEE International Conference on Robotics and Automation, pp. 5309–5314. https://doi.org/10.1109/ROBOT.2010.5509652.

Richardson, A.E., Powers, M.E., Bousquet, L.G., 2011. Video game experience predicts virtual, but not real navigation performance. Comput. Hum. Behav. 27 (1), 552–560. https://doi.org/10.1016/j.chb.2010.10.003.

Ricks, B., Nielsen, C.W., Goodrich, M.A., 2004. Ecological displays for robot interaction: a new perspective. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems. 3. pp. 2855–2860. IROS) (IEEE Cat. No.04CH37566), 3. https://doi.org/10.1109/IROS.2004.1389842.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., Bühner, M., 2010. Is it really robust. Methodology 6 (4), 147–151. https://doi.org/10.1027/1614-2241/a000016.

Schutte, P.C., 2017. How to make the most of your human: design considerations for human–machine interactions. Cogn. Technol. Work 19 (2), 233–249. https://doi.org/10.1007/s10111-017-0418-2.

Sheridan, T.B., 1995. Teleoperation, telerobotics and telepresence: a progress report. Control Eng. Pract. 3 (2), 205–214. https://doi.org/10.1016/0967-0661(94)00078-U.

Tukey, J.W., 1977. Exploratory Data Analysis 2 Reading, Mass.

Vidulich, M.A., Tsang, P.S., 1987. Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proc. Hum. Factors Soc. Ann. Meet. 31 (9), 1057–1061. https://doi.org/10.1177/154193128703100930.

Wickens, T.D., Keppel, G., 2004. Design and Analysis: a Researcher's Handbook. Pearson Prentice-Hall.

Xu, S., Perez, M., Yang, K., Perrenot, C., Felblinger, J., Hubert, J., 2014. Determination of the latency effects on surgical performance and the acceptable latency levels in telesurgery using the dV-Trainer® simulator. Surg. Endosc. 28 (9), 2569–2576. https://doi.org/10.1007/s00464-014-3504-z.

Zhang, Y., Li, H., 2016. Handling qualities evaluation of predictive display model for rendezvous and docking in lunar orbit with large time delay. In: 2016 IEEE Chinese Guidance, Navigation and Control Conference (CGNCC), pp. 742–747. https://doi.org/10.1109/CGNCC.2016.7828878.

Zheng, Y., Brudnak, M.J., Jayakumar, P., Stein, J.L., Ersal, T., 2016. An Experimental Evaluation of a Model-Free Predictor Framework in Teleoperated Vehicles**This work was supported by the Automotive Research Center (ARC) in accordance with Cooperative Agreement W56HZV-14-2-0001 U.S. Army Tank Automotive Research, Development and Engineering Center (TARDEC) Warren, MI. UNCLASSIFIED: Distribution Statement A. Approved for public release. #27479. IFAC-Papers OnLine 49 (10), 157–164. https://doi.org/10.1016/j.ifacol.2016.07.513.

Appendix C7: Academic contribution 7

Veitch, E., Dybvik, H., Steinert, M., & Alsos, O. A. (2022). Collaborative Work with Highly Automated Marine Navigation Systems. Computer Supported Cooperative Work (CSCW). https://doi.org/10.1007/s10606-022-09450-7

C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16

Computer Supported Cooperative Work (CSCW)
https://doi.org/10.1007/s10606-022-09450-7

RESEARCH ARTICLE

Collaborative Work with Highly Automated Marine Navigation Systems

Erik Veitch*[1], Henrikke Dybvik[2], Martin Steinert[2] & Ole Andreas Alsos[1]

*[1]*Department of Design, Faculty of Architecture and Design, Norwegian University of Science and Technology (NTNU), Produktdesign, 341, Gløshaugen, Kolbjørn Hejes Vei 2 B, 7491 Trondheim, Norway (E-mail: erik.a.veitch@ntnu.no);* [2]*Department of Mechanical and Industrial Engineering, Faculty of Engineering, Norwegian University of Science and Technology (NTNU), Verkstedteknisk, P317, Gløshaugen, Richard Birkelandsvei 2B, 7034 Trondheim, Norway*

Accepted: 8 September 2022

Abstract. In navigation applications, Artificial Intelligence (AI) can improve efficiency and decision making. It is not clear, however, how designers should account for human cooperation when integrating AI systems in navigation work. In a novel empirical study, we examine the transition in the maritime domain towards higher levels of machine autonomy. Our method involved interviewing technology designers (n = 9) and navigators aboard two partially automated ferries (n = 5), as well as collecting field observations aboard one of the ferries. The results indicated a discrepancy between how designers construed human-AI collaboration compared to navigators' own accounts in the field. Navigators reflected upon their role as one of 'backup,' defined by ad-hoc control takeovers from the automation. Designers positioned navigators 'in the loop' of a larger control system but discounted the role of in-situ skills and heuristic decision making in all but the most controlled takeover actions. The discrepancy shed light on how integration of AI systems may be better aligned to human cooperation in navigation. This included designing AI systems that render computational activities more visible and that incorporate social cues that articulate human work in its natural setting. Positioned within the field of AI alignment research, the main contribution is a formulation of human-AI interaction design insights for future navigation and control room work.

Keywords: Collaborative work, Interaction design, Navigation, Human–computer interaction, Autonomous ships, Artificial intelligence, Control rooms

1 Introduction

High levels of machine autonomy and Artificial Intelligence (AI) have the potential to improve work efficiency and improve human decision making. McCarthy (2007) defined AI as 'the science and engineering of making intelligent machines,' and intelligence as 'the computational part of the ability to achieve

Published online: 08 October 2022

goals in the world.' Since the field's inception in the 1950s, one of the frontiers of AI research has been navigation. Navigation – the process of moving a vehicle from one place to another – exemplifies the primary goal of computational intelligence: the capacity to execute planned action, as if by its own agency. In this study, we examine a transition currently underway in maritime navigation – a transition characterized by increasingly high levels of machine autonomy and incorporation of AI tools designed to collaborate with skilled navigators. Given the breakthroughs in AI technology in the past decade, we explore the extent to which a new human–machine interface is at hand and the extent to which systems design must realign to demands underlying a new order of work.

Driven by advances in computational power and the availability of hardware, examples of high levels of autonomy and AI in maritime applications are becoming more commonplace. Autonomous Surface Vehicles (ASVs) are plying the oceans for scientific data (e.g., Dallolio et al., 2019; Dunbabin et al., 2009; Kimball et al., 2014), autonomous passenger ferries are offering new alternatives to urban mobility (e.g., Reddy et al., 2019; Wang et al., 2019; MiT, 2020; Reddy et al., 2019; Wang et al., 2019), and Maritime Autonomous Surface Ships (MASSs) are introducing new ways to transport payload more efficiently across integrated ports (e.g., Burmeister et al., 2014; Peeters et al., 2020) In this study, we look at the case of partially automated Roll-On/Roll-On (Ro-Ro) ferries operating in Norway, where navigators complete crossings and dockings at the press of a few buttons (e.g., Kongsberg, 2020; Rolls-Royce, 2018). Looking ahead, we can expect implementation of machine learning tools designed to aid navigators make decisions (e.g., Martinsen & Lekkas, 2018; Gjærum et al., 2021; Wu et al., 2021), and computer vision to identify targets and automatically avoid collisions (e.g., Brekke et al., 2019; Helgesen et al., 2022).

High levels of autonomy and AI in sociotechnical applications like navigation rely upon collaboration with skilled human operators. ASVs need remote supervision (Utne et al., 2020), urban autonomous passenger ferries need human safety hosts, (Goerlandt and Pulsifer, 2022), and MASSs need supervision and remote control (Veitch and Alsos, 2022). Aboard the partially automated ferries we study in this article, operations depend upon the presence of a navigator who remains responsible for the vessels and its passengers and stands ready to take over control from the automated system. Despite more advanced systems that automate human manual control tasks and support decision making, the transition underway is not one of less human involvement, as one might expect, but of more collaboration between machines and humans. For designers, such systems present significant challenges. Recent accidents in aviation and car automation serve as dramatic examples of how the transition to human–machine collaboration can lead to accidents. In the years 2018 and 2019, two Boeing 737 MAX crashes revealed that the flight crew fatally lost control when counteracting a non-existent stall. A faulty airflow sensor feeding inputs to the Maneuvering Characteristics

Augmentation System (MCAS) was to blame: an automated pitch controller that the flight crew did not know how to override due to its hasty implementation (Nicas et al., 2019). In another instance, a fatal Tesla 'Autopilot' crash was found to be caused by 'system limitations' combined with 'ineffective monitoring of driver engagement, which facilitated the driver's complacency and inattentiveness' (National Transportation Safety Board, 2020, p. 58). As expressed by a leading autonomous car company in their safety report: 'While the benefits of automation are obvious, it can actually become a problem if people get tired or bored from having too little to do' (Waymo, 2020, p. 37). Whether it is an airplane, car, or even a ship, those individuals in control are increasingly finding themselves in a supervisory role, a role that Brian Christian has provocatively called the 'sorcerer's apprentice.' 'We conjure a force, autonomous but totally compliant, give it a set of instructions, and scramble like mad when we realize our instructions are imprecise or incomplete' (Christian, 2020, p. 31). In sociotechnical systems like that exemplified by a ship, where control is not executed by a single person but a whole team acting as one (Hutchins, 1995), this role, defined by the crossover between human and machine control, presents new challenges when considering work as fundamentally social action.

The premise for our study is that increased collaboration with computationally intelligent machines places new demands on its human counterparts, and that these demands can be discovered through observation and data collections efforts. Framing the current period of transition in maritime navigation as an opportunity to study these new demands, our aim is to incorporate perspectives of navigators experiencing this transition into further design iterations. Motivated by the potential of machine autonomy to enhance work efficiency and improve decision making, we seek to contribute to system design featuring a more seamless interface for coordinating action.

2 Related literature

Drawing on computer science, engineering, design, human–computer interaction, and sociology, we explore how current knowledge gaps and issues are compelling a new research direction positioned at these disciplines' crossroads. The background literature we present here sets the stage for our study, deepening and expanding the discussion about how technology designers are shaping human-AI collaborative work.

2.1 Levels of autonomy and artificial intelligence

AI has no formal definition. Far from presenting a problem for the field's practitioners, though, this lack of definition has, in the eyes of its leading experts, been precisely what has driven the field forward (Stone et al., 2016). The sociologist Levi Strauss used the term 'floating signifier' to describe phenomena like AI which, in evading definition, strengthen its suggestive power (Lechte, 1994, p. 26). The consequence of such a suggestive power, however, is captured in the so-called 'AI effect,' which describes the tendency for any new technology produced by the field, once accepted, to cast off its claims to AI. AI, in this sense, is precisely what is under development. In the development of autonomous vehicles, which represents the field's idyllic mission of imbuing agency in a computational object, traces of the AI effect can be detected in the taxonomies commonly adopted to establish 'how autonomous' a vehicle is. These 'Levels of Autonomy' (LoA) taxonomies are not binary (autonomous or not) as one might expect. Rather, LoAs are more like standardized yardsticks for the extent to which a vehicle's agency is independent of the human driver's. These taxonomies have their origin in road transportation (SAE International, 2017) and have more recently been developed for maritime transportation (IMO, 2018; Rødseth, 2017). While LoA taxonomies vary, their basic structures remain the same, laying out an integer scale starting at zero or one, which represents full human control, and extending incrementally to some number that represents full machine control. For the vast majority of technology developers, this top number, like the field that proposed it, is a floating signifier. Only the intermediate numbers, which presume a collaborative approach to the myriad actions involved in driving a vehicle, are considered feasible.

Despite the apparently intractable goals underlying machine autonomy, the field of AI has been remarkably productive in producing technologies and techniques enabling intermediate LoAs. The theoretical underpinnings of modern computational machine learning techniques like Deep Neural Network (DNNs) have been around for decades, but only in the past decade has computational power enabled their widespread use. Advancements in machine learning techniques, too, have rapidly advanced the field, including in areas like natural language processing, image and video classification and generation, planning, decision making, and integration of vision and robotics. In the face of such advancements, however, a major new challenge has arisen. As expressed in Stanford University's '*AI100 Report*,' the field's most influential experts recognized that, 'Perhaps the most inspiring challenge is to build machines that can cooperate and collaborate seamlessly with humans' (Littman et al., 2021, p. 19). In response to this challenge, an active research community has sprung up. These researchers are dedicated to 'AI alignment,' and include not just computer engineers and designers, but also anthropologists and sociologists, safety specialists and organizational scientists. In the context of sociotechnical systems, like that exemplified by our focus on the transition in maritime navigation, there is a growing need for such multidisciplinary efforts to understand the implications of high levels of autonomy and AI in safety–critical work. We position our work within the efforts of AI alignment research, interpreting the transition underway as one necessitating a realignment of design practices with the social actions coordinating human work.

2.2 Centres of coordination

The supervisory role taking shape in the wake of higher levels of autonomy has generated interest in centres of coordination for autonomous vehicles. For maritime navigation, this is exemplified by the concept of land-based supervisory control of highly automated ships, variously referred to in the literature as ‘shore control centre,’ ‘remote control centre,’ or ‘remote operating centre.’ These terms, which have surfaced in the past decade, capture a renewed interest in control rooms. Control rooms were a topic of academic interest in human factors and cognitive engineering in the 1970s and 80s especially in the context of complex, sociotechnical systems like nuclear power plants (Rasmussen, 1986; Vicente, 1999). In the 1990s, control rooms were of academic interest in the field of Computer Supported Collaborative Work. Researchers in CSCW studied the sociality of computer use in natural settings like line control rooms (Heath & Luff, 1992), airline scheduling (Goodwin & Goodwin, 1996), and emergency dispatch (Whalen, 1995). Today, in the wake of technological developments enabling higher levels of autonomy, the spotlight is once again directed towards the control room, the stage upon which supervisory control and time-critical action is orchestrated, enabling the coordination of highly automated vehicles across distributed locations. In this context, we revisit Lucy Suchman’s definition of ‘centre of coordination’:

> ‘Centres of coordination are characterized in terms of participants’ ongoing orientation to problems of space and time, involving the deployment of people and equipment across distances, according to a canonical timetable or the emergent requirements of rapid response to a time-critical situation.’ (Suchman, 1997, p. 42)

For autonomous ships, the ‘shore control centre’ as a centre of coordination presents significant challenges to designers. The International Maritime Organization (IMO), the inter-governmental agency for standardisation of safety at sea, outlined such outstanding challenges in their ‘Regulatory scoping exercise for the use of maritime autonomous surface ships (MASS)’ (IMO, 2021). In their report, the highest priority issues concerned the role of the navigator working in a location separate from the ship environment. While navigators’ responsibility for the safety of the ship remained unchanged, the environment in which they work was substituted by an information-rich landscape necessitating new skills and competencies (IMO, 2021, p. 8). In revisiting centres of coordination, we explore what concepts and theories that emerged from seminal control room studies remain relevant today, and what gaps emerge in the light of new technological and organizational developments.

2.3 Design for AI collaborative systems

The need for improved human–machine collaboration predicated by recent technological development has led to new frameworks adopting human-centred design principles to AI systems. Shneiderman (2020), for example, proposed a design framework for 'human-centred AI' based on the principles of safe, reliable, and trustworthy system interactions. The field of Human–Computer Interaction (HCI) has put forward practical guidelines for designers adapting to such frameworks (e.g., Amershi et al., 2019; Mahadevan et al., 2018). The rapidly growing field of explainable AI (XAI), too, focuses on the interaction between humans and machines, aiming to establish human-based values of interpretability and understandability at the core of 'black box' machine learning techniques (Voosen, 2017). Expanding the audience of XAI towards users, organisations, and even non-governmental agencies, Arrieta Barredo et al. (2020) envision a 'Responsible AI' initiative, which embraces values of fairness and accountability along with the mandate of model explainability at the core of XAI. The multidisciplinary field of 'machine behaviour' has also emerged recently, which sets out as its mission the empirical treatment of the ways in which human social interactions are modified by the introduction of intelligent machines (Rahwan et al., 2019). The field of CSCW, with its interest in computationally infused environments and enacted elements of work, also stands to offer distinct contributions to this discussion. Ethnomethodological works on social interactions during navigation and control of ships (Hutchins, 1995) and airplanes (Nevile, 2001) lay the theoretical groundwork for such contributions, while more recent discussions exploring 'ethical AI issues' (Fleischmann et al., 2019) and 'challenges in human-AI collaboration' (Park et al., 2019) pave the way for current research directions. The aim of our study continues in this vein, motivated by lack of knowledge about how the transition to higher levels of autonomy affects the social underpinnings enacting work in its natural setting.

2.4 Ironies of automation

Bainbridge (1983), writing in her seminal paper 'Ironies of Automation,' described the paradoxical decrease in human abilities resulting from machines designed to improve that very ability. Among human factors specialists the effect is well-known, but despite its articulation three decades ago, its consequences persist in modern system design. For example, skill degradation associated with automation emerged as a key factor in the high-profile crash of flight Air France flight 447 in 2009, which fatally stalled over the Atlantic Ocean after the automatic flight system handed control to the flight crew shortly after detecting faulty readings from an airspeed sensor. As the accident report stated, one of the contributing causes of the stall and resulting crash was 'The absence of any training, at high altitude, in manual aeroplane handling' (Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, 2012, p. 201). Consequently, guidelines now recommend pilots practice manual flying regularly, highlighting that 'continuous use of automation does not strengthen pilots' knowledge and skills in manual flight operation and in fact could lead to degradation of the pilots' ability to quickly recover the aircraft from an undesired state' (IATA, 2020, p. 5). The consequences of skill degradation are exacerbated in systems with high LoA that require timely and decisive preventative action from a skilled operator. Taking irony to be a poor premise for design, we consider instead how design activities better aligned to the needs of collaboration can avoid the pitfalls associated with automation-induced skill degradation.

3 Methodology

Our methodology consisted of field study observations and semi-structured interviews. The research design was motivated by the practical need to inform design efforts implementing high levels of machine autonomy and AI techniques in maritime navigation applications. The aim was to describe the extent to which design practices currently shaping a transition in the maritime domain are aligned with the realities of skilled, safety–critical work in the field.

3.1 Data collection

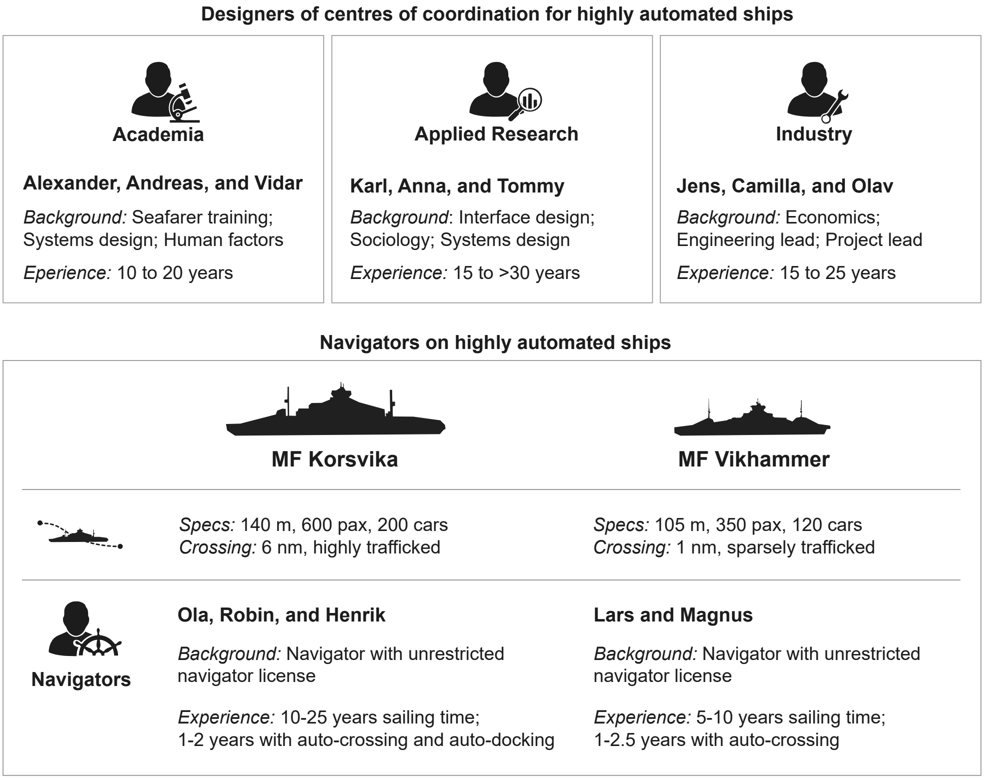
The empirical data consisted of semi-structured interviews with individuals in the design and research communities (n = 9) and navigators working aboard two partially automated passenger ferries (n = 5). To provide context about the natural setting in which the navigation work takes place, we also report on field observations conducted at the site of one of the ferries featured in the interviews. All data were collected in Norway.

Selection of interview participants was guided by theoretical sampling commonly employed in Grounded Theory Methods (Corbin and Strauss, 2015; Glaser and Strauss, 1999). This allowed us to follow up on themes of interest and target subsequent participants as new open-ended questions presented themselves. After completing nine interviews with technology designers and researchers in autumn 2019, it became clear that the perspective of navigators in the field would be of interest. Turning to this gap, a field study was conducted aboard the navigation deck of a ferry outfitted with state-of-the-art automated navigation technology. Field notes and images were collected by the first author, and once again new questions were posed. Interviews were subsequently held in summer 2021 with three of the navigators aboard this ferry, followed by two more aboard a similar ferry.

Two researchers were present for all interviews, with one leading the conversation while the other transcribed, verbatim. The first author was present for all interviews, which were conducted either in-person or via video call and ranging in length from 45 to 60 minutes. Participants consented to data collection before and after the interviews, and all interviews that were held in Norwegian were translated into English.

3.2 Interview respondents

A total of fourteen informants were interviewed, each of whom we refer to in this study with a pseudonym (Figure 1). The group whom we refer to as ‘Designers’ originated from academia, applied research, and industry. This group consisted of individuals with expert domain knowledge about the development of centres of coordination for highly automated ships. Their unique contribution was insights about activities shaping the transition towards higher levels of autonomy in marine navigation work. This group captured a wide breadth of perspectives



Designers of centres of coordination for highly automated ships

Academia

Alexander, Andreas, and Vidar

Background: Seafarer training; Systems design; Human factors

Eperience: 10 to 20 years

Applied Research

Karl, Anna, and Tommy

Background: Interface design; Sociology; Systems design

Experience: 15 to >30 years

Industry

Jens, Camilla, and Olav

Background: Economics; Engineering lead; Project lead

Experience: 15 to 25 years

Navigators on highly automated ships

MF Korsvika

Specs: 140 m, 600 pax, 200 cars
Crossing: 6 nm, highly trafficked

Ola, Robin, and Henrik

Background: Navigator with unrestricted navigator license

Experience: 10-25 years sailing time;
1-2 years with auto-crossing and auto-docking

MF Vikhammer

Specs: 105 m, 350 pax, 120 cars
Crossing: 1 nm, sparsely trafficked

Lars and Magnus

Background: Navigator with unrestricted navigator license

Experience: 5-10 years sailing time;
1-2.5 years with auto-crossing

Navigators

Figure 1 Description of interview informants

on technology development, holding different titles and originating from distinct professional networks separate from the networks held by the authors.

The group whom we refer to as 'Navigators' represented captains and chief mates working aboard two ferries outfitted with state-of-the-art automation technology. All navigators had a 'D1' deck officer license, the highest maritime navigation license in Norway. At the time of writing, the number of navigators working aboard partially automated ferries represented a small population. As such, we were careful to characterise them broadly to avoid de-anonymising them. The *Korsvika* (a pseudonym) was, at the time of this writing, the world's only ferry operating regularly with both auto-crossing and auto-docking, making it relatively easy to identify. It was on this ferry that field observations took place. There is a total of eight deck officers on the *Korsvika* and we interviewed three of them. The second ferry in our case study, called the *Vikhammer* (also a pseudonym), had just auto-crossing installed. The *Korsvika* and *Vikhammer* were owned and operated by different companies.

3.3 Data analysis

Our analytical approach was inspired by Grounded Theory Methods (Corbin and Strauss, 2015; Glaser and Strauss, 1999; Morse et al., 2009). Observations made on the navigation deck aboard the *Korsvika* also served an important role in the analysis, describing the context in which navigation work took place. During interviews and field observations, insights were recorded as 'memos:' dated text excerpts ranging from short notes to long, descriptive passages. No less than 101 memos were recorded in total, which served as precedents to a more structured analysis aimed at synthesizing these early insights.

In structuring the analysis, we used the software tool NVivo (NVivo, 2020). At its most fundamental level, the analysis comprised of 'codes' – units of highlighted text representing potentially relevant findings. Our analysis consisted of several hundred codes, which we assigned to categories called 'axial codes.' Special attention was afforded to retaining terms and phrases used by informants and to resisting re-interpretation in our own wording. For example, the term 'backup,' emerged as an important axial code. While only two navigators used the term expressly, the saliency of the theme was made apparent through other related codes (e.g., Ola: 'you become an operator who monitors the systems and is ready to press a button if there's a bug;' Henrik: 'When what you see on the screen no longer shows the correct thing, that's when things get interesting'). The axial coding process was iterative and was conducted by the first author and two graduate students, involving many rounds of discussion with the authors over the study period.

Eventually, we distilled our analysis into an overarching narrative structure, focused on the discrepancy between designers' construal of navigators' work and navigators' own reflective accounts. These findings are presented in Sect. 5. Before presenting this, however, it is necessary to provide some context to the findings. In Sect. 4 that follows, we outline the work activities making up a regular crossing aboard the *Korsvika*, constructed from first-hand field observations.

3.4 Methodological limitations

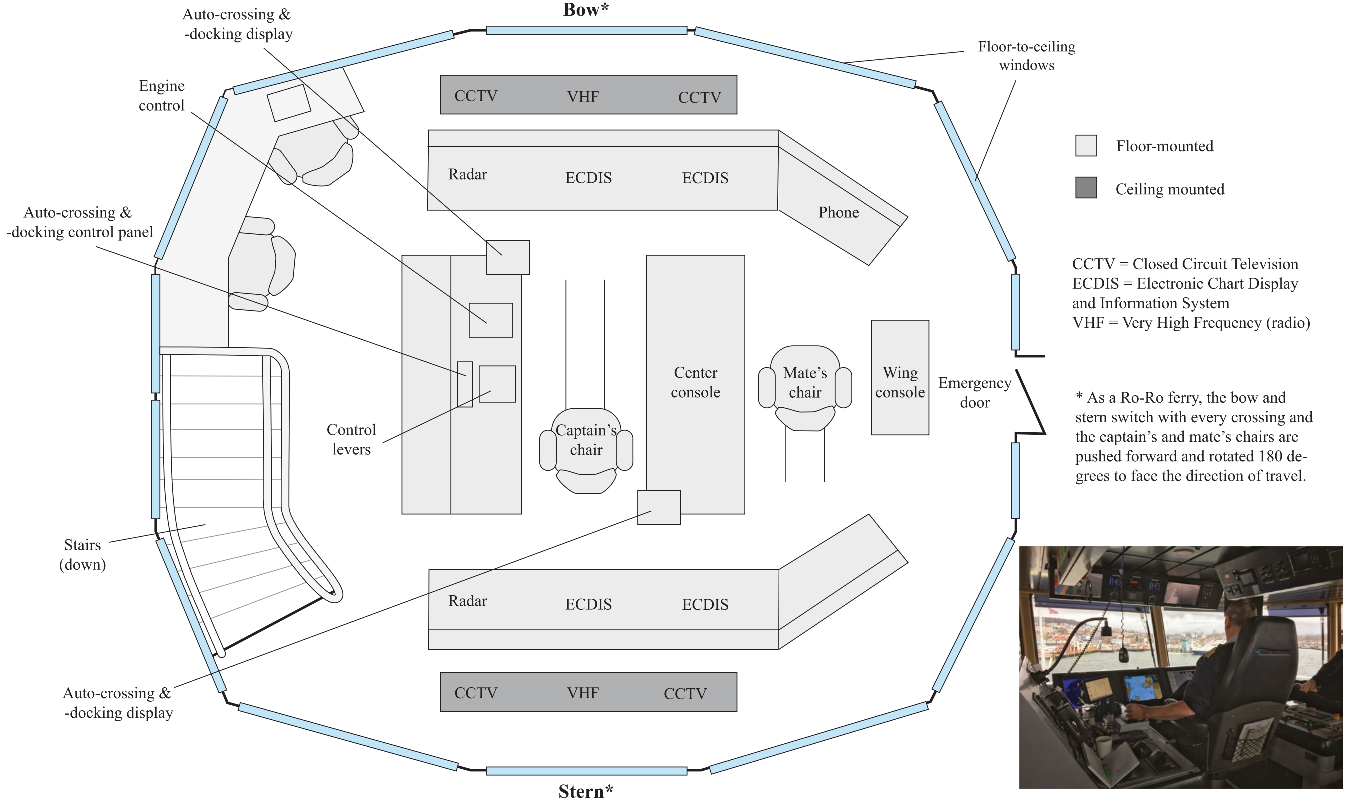
The empirical study consisted of both field observations and semi-structured interviews, lending our research design some distinct advantages as well as limitations. One advantage, for instance, presented itself from conducting independent, one-on-one interviews, as it led to the discovery of discrepancies between designers' and navigators' accounts of the same core activities. Similarly, this approach lent itself to making comparisons within groups. For example, when we compared accounts of navigators on different ferries, insights emerged linking their use of automation with skill degradation (Sect. 5.1). The conditions of confidentiality and anonymity, too, proved to be helpful in a way that field observation alone could not be. Informants were free to express their opinions without the potentially self-censoring effect of their colleagues' or managers' presence and reflected on their work activities as if observing them from the outside. Having interviewed the operators during a global COVID-19 pandemic, video conferencing provided a useful platform for data collection during social distancing.

4 Field observations aboard the *Korsvika*

In this section, we present field observations from the *Korsvika*. The *Korsvika* (a pseudonym) is the world's first ferry in regular service equipped with auto-crossing and auto-docking: two technologies representing a step change in the transition towards higher levels of machine autonomy in the maritime domain. For simplicity, we refer to the two technologies together as 'auto-systems.' The account that follows is a description of work during a regular crossing, as well as the environment of the navigation deck and the functionality of the auto-systems. The aim is to provide context about the roles, responsibilities, and tasks of the navigators, how these navigators interact with each other and the auto-systems, and how the adoption of higher levels of autonomy impacts their work activities. The diagram in Figure 2 can be used to orient the reader on the *Korsvika*'s navigation deck.

4.1 The ***Korsvika***'s ferry service

The *Korsvika* connects vehicle traffic and foot passengers between two busy ports in Norway. The crossing takes less than 45 minutes. Operations are going smoothly when this 140-m-long roll-on/roll-off ferry, with capacity for almost 600 passengers and 200 cars, is on time with an even gap behind the other ferries that sail the same route. Because several ferries traverse the same crossing, issues can arise when one ferry is delayed, forcing the ferry behind to wait for it outside the dock. There are many factors that can affect the ferry's service, including the weather conditions and even the



▢ Floor-mounted

■ Ceiling mounted

CCTV = Closed Circuit Television
ECDIS = Electronic Chart Display and Information System
VHF = Very High Frequency (radio)

* As a Ro-Ro ferry, the bow and stern switch with every crossing and the captain's and mate's chairs are pushed forward and rotated 180 degrees to face the direction of travel.

Figure 2 Layout of the bridge aboard the *Korsvika* (image taken by first author)

sailing styles of different navigators on duty. The new auto-systems installed aboard the *Korsvika* were intended to improve the efficiency of ferry service, saving fuel while providing customers with a more consistent service.

4.2 The ***Korsvika***'s crew

The captain has overall responsibility for the safety of passengers and crew. The chief mate (often shortened to 'mate') shares much of this responsibility. The captain and mate relieve one another's shifts throughout the ten-and-a-half hour working day, exchanging regular handovers in what the navigators call 'sharing a voyage.' Two bosuns handle the physical work on the main deck: loading, unloading, fitting cars, maintenance, and checks of safety equipment. One of the two navigators (captain or mate) communicates to the bosuns over a local radio and observes their actions from the bridge windows or on Closed Circuit Television (CCTV). The shipowner requires that two crew personnel must always be on the bridge, so after handovers between the captain and mate, a bosun comes up to the bridge and joins as a lookout. Other than the navigators and bosuns, the crew consists of a chief engineer, a mechanic, and cafeteria crew. Of all the crew aboard the *Korsvika*, the new auto-systems directly affect only the navigators' day-to-day work.

4.3 Loading and leaving dock

At the dock, the ferry loads vehicles and foot passengers. When loading is completed, the command 'Lift up!' is radioed to the bosuns, cueing them to close the ramp door and secure it for crossing.

Leaving the dock can be accomplished by the navigators either manually using thruster controls ('at the handles,' to use their terminology) or by pressing a button on the new auto-docking system. Currently, the auto-docking is used for 50–70% of all voyages.

Leaving the dock, the captain or mate reports their departure to the local Vessel Traffic Services (VTS) centre that they have left the dock, and VTS replies with any relevant information about traffic in the area. The navigator also keeps an eye out for small recreational boats, which are typically not detected by VTS. The new auto-systems are not yet equipped with cameras to detect possible collision targets, so the navigator must be attentive even when in auto-mode.

4.4 A regular crossing in 'auto-mode'

Shortly after the ferry is clear of the dock, auto-crossing is engaged by pressing the 'AUTO CROSS' button on the console. Nearby, on a small screen the size of a tablet computer, a touchscreen indicates that auto-crossing has been engaged and displays system information like thrust and heading. The handles on the thruster controllers move by themselves as the ship settles into its route and adjusts its speed for the crossing. The captain sits back in the chair and looks out the window, occasionally glancing at the Electronic Chart Display and Information System (ECDIS) and radar. The lookout sits in the mate's chair beside the captain, looking out the window and glancing occasionally at the captain.

Sometimes, small boats are encountered enroute. On weekends and summer holidays, there may be many such recreational boats in addition to regular commercial traffic. These small boats warrant special attention, because unlike commercial ships with trained crew, their occupants may be unfamiliar with the rules of navigation and may occasionally end up on a collision course. The auto-crossing is not yet capable of avoiding collisions. Avoiding collisions remains one of the core duties of the navigators. For larger ship traffic, whose navigators manoeuvre their vessels in accordance with Collision Regulation conventions (COLREGs), there are generally no issues avoiding collisions. Should a ship cross from either port or starboard, an agreement is usually made over the radio regarding who will adjust course or speed to pass behind the other, even if it is the give-way vessel that does so. In a give-way situation, the navigator takes over manual control by pressing the 'MANUAL' button on the auto-system console. Pulling back on the thruster, the other ship can cross ahead, whereafter the navigator can press 'AUTO CROSS,' resuming the crossing and losing little time to the timetable.

4.5 Arriving at dock and unloading

Approaching the dock, the auto-system alerts the crew with a loud beep followed by a pre-recorded voice announcing that docking is about to start. The alarm is acknowledged by the captain by pressing the 'AUTO DOCK' button that starts the auto-docking stage. Were the captain to ignore the alert, a safety measure is built in to stop the ferry in station-keeping mode, holding position some distance away from the dock.

As the ferry heads to the dock slowly under auto-docking control, the mate joins the captain (or vice-versa) in time for the docking sequence. At this point, the bosun who was on lookout duty during the crossing heads down to prepare for unloading. 'Betty's taking care of it,' announces the captain, using a nickname referring to the auto-docking system. The mate acknowledges, confirming they understood that the ferry is docking automatically.

At the dock, the captain communicates with the bosuns over radio and the ramp is lowered and unloading commences. Shortly after unloading, loading begins again. The captain's and mate's chairs are slid forward and rotated 180 degrees and the *Korsvika* sets out for its other port in the direction from where it came.

4.6 Higher levels of autonomy and centres of coordination

Currently, there are cameras installed in the *Korsvika* bow that record all marine traffic it encounters. Technology developers behind the auto-crossing and auto-docking initiatives are working towards enhancements; for example, they can use the recordings to train machine learning algorithms that can classify objects and be used in collision avoidance algorithms. As development of more advanced automation continues, there have been discussions about reducing crew aboard the ferries and controlling fleets of highly automated ferries from a land-based centre of coordination. Higher levels of autonomy have already proven successful on the *Korsvika*, improving the efficiency of fuel consumption and consistency of service in the face of highly variable external factors. Unlocking the potential benefits of higher levels of machine autonomy, though, depends on seamless integration of the AI systems with what is, at its heart, human work.

5 Interviews with designers and navigators

In this section, we present the findings of the interviews with navigators both aboard the *Korsvika* and the *Vikhammer*, as well as with technology designers and researchers shaping the transition towards higher levels of autonomy in the maritime domain. We start with the navigators, who recounted a shift to a 'backup' role subsequent to the introduction of auto-systems aboard their ships. Then, we compare this to accounts of the designers, whose construal of working with automated systems seemed misaligned with navigators' own accounts of working with automation in the field.

5.1 Navigators' perspectives: shifting to a backup role

The navigators attributed agency – a capacity for action – to the auto-systems. The influence of this agency was most evident in their descriptions of transitioning from 'hands-on' to 'backup' navigation.

The nickname assigned to the auto-systems by some of the navigators ('Betty')exemplified how machine agency could be manifested. Betty could 'take care of it,' as Robin reported, referring to the complex process of docking the 1400-ton *Korsvika* to the dock. In fog, Betty was 'ingenious' given her ability to dock in zero visibility. Betty could be a 'nag,' however, and 'do weird things,' according to Ola, who, as if by way of assuring themselves, told us that 'she has no thoughts of her own.'

> Robin: 'My captain and I, if we're auto-docking, we say that "Betty's taking care of it." Then he knows that auto mode is on. If we have normal autopilot on then I say that "Betty's not taking care of it."'

The nickname ‘Betty’ was used by two of the five navigators we interviewed, both aboard the *Korsvika*. Traditionally a woman’s name, Betty was chosen owing to the system’s female voice announcements, played at intervals to announce stages of operations or to alert navigators’ attention to some procedure or sequence. Personified in this way, the navigators described interactions with the auto-system in human terms.

The agency attributed to the auto-systems underpinned the emerging ‘backup’ role described by the navigators. We adopted the term ‘backup’ from Robin, who, describing a transition in their work in recent years, said, ‘We are the backup if something happens.’ Other navigators described a similar role. ‘You go from being the one who performs something to just monitoring something,’ said Henrik of the transition, ‘but when what you see on the screen no longer shows the correct thing, that’s when things get interesting.’

One limitation of the auto-crossing was that it did not yet have automated collision avoidance capabilities, meaning such manoeuvres were left to navigators. Collision avoidance manoeuvres are regulated in the 1972 Convention on the International Regulations Preventing Collisions at Sea (COLREGs). The convention lays out traffic rules, like Rule 8 stating that collision avoidance actions must be ‘made in ample time and with due regard to the observance of good seamanship’ (IMO, 1972). Rules work best if everyone knows them, which is not always the case. ‘The biggest problem is with small boats and sailboats,’ Ola reported. ‘They don’t have the same knowledge about rules, speed, and direction,’ explained Henrik. ‘They think we move slower,’ said Robin, ‘so we have to press “MANUAL” … you don’t want to run someone over.’ In collision avoidance, the navigators’ backup role to the auto-systems was clearly defined: take over control to adhere to the COLREGs, with special attention to small boats. Another backup role emerged, however, with less clearly defined parameters. This was illustrated by Robin who recounted an instance when they took over control to make a crossing more comfortable for passengers:

> Robin: ‘… these days we [the navigators] say: if it’s blowing, we steer manually. Auto-crossing can be used at any time, but manual mode is more comfortable for passengers.’
> Interviewer (Erik): ‘You steer the ferry [manually] so it’s more comfortable for passengers?’
> Robin: ‘If you have rolling, people can fall and hurt themselves. Instead of rolling all the way over, I sail a little North and then a little South to go across the waves.’

Robin’s interaction with the automated system in this case is not determined by safety–critical and timely intervention, but rather on the system’s inability to account for comfort of passengers. Whether in taking over control to avoid hazardous traffic situations or simply to attend to passenger comfort, the shift from a hands-on role to a backup role underscored the most salient change in navigator work after the auto-system's introduction.

One effect of shifting to a backup role was skill degradation associated with more time spent in a passive, monitoring role relative to hands-on, manual control. Skill degradation was especially apparent when comparing navigators' accounts from the *Korsvika*, who reported that 50–70% of crossings were in auto-mode, to the *Vikhammer*, who reported close to 100% automated crossings. As reported by Henrik, the crew of the *Korsvika* had taken to driving the ship manually 'at least twice per shift so as not to forget how that works.' This suggested that skill degradation set in quickly, possibly over the course of days, and that regular practice was an effective countermeasure. '…when I have driven a lot of auto,' said Henrik, 'I have to steer a couple of times myself to get the feel of it again.' Robin expanded on the subject, noting that operators' propensity for regular manual sailing practice resulting in it being incorporated into the shipowner's operating procedures:

> Robin: 'We've set it up so you'll sail it [the ferry] yourself during the day to maintain your driving. That's written in our procedures now. If you've had a holiday, you're allowed to steer the whole shift, there's no one that says you have to use auto-crossing.'

On the *Vikhammer*, in contrast, the crew had seldom sailed manually since the auto-crossing was implemented. This implied a more significant skill degradation, which might compromise safety in the eventuality of a manual takeover.

> Magnus: 'We only use auto-crossing now – every day, every trip.'
> Interviewer (Erik): 'Do you ever turn it off to take manual control?'
> Magnus: 'No.'
> Interviewer (Erik): 'When was the last time you drove manually?'
> Magnus: 'We might occasionally drive if we have an ambulance dispatched. Auto-crossing must have the lowest energy consumption, but with an ambulance it's life and health. Apart from that … it's been one-and-a-half years since I stopped doing it [driving manually] myself.'

Given how fast de-skilling was a factor among the crew of *Korsvika*, one cannot help but wonder if the crew of *Vikhammer* are prepared for an ambulance dispatch. Manual skill practice procedures, even in situations well-suited to the automation, appeared to be a useful countermeasure to skill degradation for the navigators aboard the *Korsvika*.

5.2 Designers' perspectives: prescribing action for distributed work

The interviews we held with technology designers and researchers yielded insights into how development activities are shaping the transition to increased human-AI collaboration in maritime navigation. Here we outline what this group identified as the most important design goals and what methods they are adopting to address interaction challenges between humans and machines. Then, we compare designers' construals of working with higher levels of machine autonomy with navigators' own corresponding accounts.

To begin, we outline some of the major design goals, the approaches being adopted in the industry and research communities, and what specific challenges represented outstanding gaps and issues. The main goals driving the transition towards higher levels of autonomy in the maritime domain included achieving improved 'logistics,' 'system design,' and 'centralized control.' These goals, it was envisioned, will be accomplished primarily through crew reduction relative to ship payload, as well as through centralized management of employees and ship assets from a centre of coordination. 'The whole problem statement,' said Vidar, 'can be defined as moving work farther from the pointy end to more distributed locations.' By 'pointy end,' Vidar referred to operational work in the field, a term coined by organizational scientist Rhonda Flin (Flin et al., 2008) and used often in the context of exposure to hazardous working environments. Asked to describe the vision of autonomous ships, interviewees described fleets of ships with reduced crew (or in some cases, no crew at all), whose whereabouts were tracked by trained operators in a centralized control centre. Prompted further to describe the control room, images of data-rich information displays were invoked in all interviews ('there will be large-screen displays displaying the "big picture,"' reported Karl; 'through the screen [the operators] will have access to the data they need,' said Andreas). Many of the technological artefacts located in a conventional ship were mentioned, including ECDIS, marine radio, and software for ship scheduling and voyage plans. What distinguished the control room from ship's bridge was the amount of additional data (e.g., video streams, sensor displays, and the like) and, crucially, the ability to take direct control over the ship. Here the analogy was made by five of the interviewees to VTS operators, who, tasked with monitoring traffic in busy port areas, can indirectly direct traffic by contacting navigation crew over radio. In a control room for highly automated ships, such actions could be taken directly instead of indirectly, making the control room operator effectively a remote captain in addition to traffic director.

Two interviewees described interactions at the screen interfaces in terms of 'top down' and 'bottom up' processing. As explained by Karl, this was intended to support decision making at the cognitive level, combining top-down processing ('information search') with bottom-up driven processes ('information that catches the attention of operators'). Two opposing viewpoints emerged, however: some interviewees argued that the control room should be designed to accommodate work as it takes place aboard a ship's bridge; others argued that the control room will require a ground-up approach, requiring specifications drawn up according to distinct requirements. Among the latter group, 'human-centred design,' 'prototyping,' and 'systems engineering' featured as methodologies to uncover these distinct requirements. Discussions about design strategies met a significant challenge: for highly automated ships, there were no standardised guidelines aimed for accommodating approval like those akin to conventional ships. Conventional passenger ship design, for example, is standardized according to design guidelines laid out by classification societies like DNV GL in their 'Rules for Classification' (DNV GL, 2017). For highly automated ships, adopting 'goal-based approaches' were, in place of prescriptive approaches, the most viable option towards approval of designs by regulating authorities. Characterizing this goal-based design process, five interviewees called it a 'transition,' involving testing, verification, and approval – lengthy processes typical in the highly regulated industry of shipping.

The technology for enabling high levels of machine autonomy, it appeared, was more or less available; orchestrating this technology in a real-world context, though, remained the challenge. In the boundary between human and machine, several gaps and issues were identified. The number of vessels, for instance, that each operator should control was unknown. This number was linked to the LoA of the vessel, but the LoA, too, was ambiguous, referencing various taxonomies each with its own configuration of how automated tasks and human tasks combine to navigate a ship. Specific LoA taxonomies mentioned by the interviewees included DNV GL (2018), IMO (2018), and NFAS (Rødseth, 2017). A central problem was the amount of time it takes to take over control. On the premise that such control takeovers are preventative and time-critical, the maximum allowable takeover time emerged as perhaps the single most important factor in goal-based design directed towards collaboration with the automated system.

> Tommy: 'You must quantify the person's response time. This will help a lot with the approval of a shore control centre, because then you can document, for example, that the system gives ten seconds warning and that we have done the research showing that the operators are trained for this. Today, nobody knows.'

5.3 Discrepancies between navigator and designer accounts

Comparing interviews of designers and researchers with those of navigators, certain discrepancies came to light. Two such discrepancies pointed to ways in which designers' construal of human–machine interaction diverged from those who inhabited this interface in their work. The first related to how the two groups treated decision making for control takeovers; the second related to how they reconciled their safety responsibilities while relinquishing tasks and decision making to machines.

Designers, in their efforts to build interfaces that supported navigators' work, adopted practical models for decision making based on sensory input and cognitive processing. The model of 'situation awareness,' attributed to Endsley (1995), was especially prominent, appearing independently in four of the nine interviews we held with designers. Navigators, by contrast, did not refer to situation awareness, neither directly nor by its characteristic features, which decompose decision making into distinct information processing stages. In the following excerpt, for instance, Karl, a designer, described design needs for a control room to support work for navigating highly automated ships, framed in terms of 'situation awareness' needs:

> 'What data is needed to control and monitor the [highly automated] ships: that is situation awareness need number one. Then situation awareness need number two is to display that into something understandable. Situation awareness need number three is to project that into the future. That could be a way to approach the concept [of operating highly automated ships] in a more… systematic way, perhaps.'

By contrast, navigators invoked heuristic approaches to decision making, drawing from in-situ skills informed by experience. One example of such a heuristic was illustrated by Robin who recounted taking over control to attend to passenger comfort (see Sect. 5.1). In that example, rather than following a sequence of information processing stages, the decision to take over manual control stemmed simply from imagining how passengers would experience the crossing in the given sea state.

Four of the nine designers we interviewed expressed the concept of being 'in the loop,' referring to the state of mind one must be in to take over control from automation. The navigators, by contrast, referred to this same state as 'backup.' Being backup reserved the sense of responsibility that comes with being a navigator, while losing the agency involved in manoeuvring a ship under one's own hand ('The job hasn't changed,' reported Ola, 'but in auto you can sit back and let the system do it'). Being 'in the loop,' by contrast, construed the navigators as components in a larger, cognitive system. In this 'loop,' whose terminology is rooted in control theory, the navigator was expected to passively monitor the closed loop of automated control and immediately close this loop – through timely and decisive takeover action – the moment the loop's integrity was compromised. As explained by Alexander, a designer, 'The key challenge will be to get the operator, in the shortest possible time, to get in the loop of what is going on.'

6 Discussion

In this section, we explore the implications of the field observations and interview study results, framed in terms of the knowledge gaps and issues introduced in Sect. 1 and outlined in more detail in Sect. 2. Towards this aim, we focus discussions around three relevant themes: (i) the agencies of humans and machines in collaborative navigation, (ii) the transition to centres of coordination, (iii) the social implications of AI collaboration, and (iv) control rooms of the past, present, and future.

6.1 Agencies of humans and machines in collaborative navigation

One of the most salient themes uncovered in the analysis was a transition to a 'backup' role, defined by peremptory control interventions, or 'takeovers.' For technology designers and researchers, the transition toward higher levels of autonomy in shipping culminated in centres of coordination, where operators were 'in the loop' of the system. Navigators' accounts of inhabiting this transition in their own work reflected a preoccupation with their own agency and expressed a desire to recover this agency. Backup implied two mutually exclusive activities: passive monitoring in situations for which the automation was well-suited, and active control in situations for which it was not. Backup invoked the 'sorcerer's apprentice' role (Sect. 1), necessitating timely intervention to stop the conjured force of a machine imbued with agency.

Lucy Suchman, in her 'Plans and Situated Actions' (Suchman, 2007), framed the human–machine interface in terms of co-existing intentions entrenched both in control algorithms (plans) and in-situ skills (situated actions). In the context of the backup role, navigators co-existed as passive operators when plans represented by the automation proceeded as expected, and as skilled operators when those plans were inevitably jettisoned to deal with some situation at hand.

Navigators' accounts also underscored the extent to which the canonical 'ironies of automation' applied to the present transition (Sect. 2.4). One such example emerged from the observation of skill degradation in navigators' ship-handling (Sect. 5): the auto-systems were, in effect, compromising the very thing it was designed to improve. Given the central importance of in-situ takeovers in the backup role, the manual ship handling skills seemed, paradoxically, of heightened importance in the face of increasing levels of machine autonomy.

As part of the backup role aboard the *Korsvika* and *Vikhammer*, there was a sense that in order for operations to go smoothly, navigators depended on the automation system as much as the automation system depended on the navigators. While the auto-crossing and auto-docking systems onboard represented relatively low levels of autonomy, the stakes introduced by this inter-dependence appeared to be getting higher for higher levels of machine autonomy. Demski and

Garrabrant (2019), envisioning the system requirements for an ideal cooperative AI, called this inter-dependence ‘embedded agency.’ By this design, a cooperative AI must be self-referential, capable of modelling is own impact on its environment, including how its users adapt to its presence. Dautenhahn (2007) framed this same capacity in terms of social interactivity, pointing out that activities requiring increasing degrees of interactivity require the computational system to be able to reflexively adapt to constantly changing conditions – a form of artificial ‘social’ intelligence. Navigation is exemplary of such an activity, requiring attention not just to what tasks can be automated, but how they should be automated in the context of a socially organized activity.

Whether it was framed as ‘in the loop’ by designers or ‘backup’ by navigators, being continually prepared for takeovers emerged as the defining feature in a new landscape of joint human-AI agency. The takeover, which symbolized the boundary between machine and human control, helped bring to light two specific design issues: firstly, operators’ sense of agency was upended, manifesting in skill degradation over longer periods of passive monitoring; secondly, effective collaboration between operators and highly automated navigation systems was left hanging in the balance of situated actions and computational plans in a flux of changing situations.

6.2 Transition to centres of coordination

The need for supervisory control of highly automated ships has generated renewed interest in centres of coordination for marine navigation. Referred to as ‘shore control centres,’ ‘remote operating centres,’ or ‘remote control centres’ by the informants in our study, centres of coordination of this type have emerged only in the past decade and have since grown significantly in the scientific literature (Veitch and Alsos, 2022). This renewed interest warrants a closer look at the guiding principles presented in Suchman’s original articulation of the centres of coordination concept (Suchman, 1997), which was aimed especially for designers (see Sect. 2.2). In revisiting the theoretical considerations associated with centres of coordination, we also ask whether they are still relevant given the recent technological developments in the decades following the concept’s introduction.

To begin, it is worth reiterating how centres of coordination relate to maritime navigation and to the transition to higher levels of machine autonomy. After all, the original case used to characterise them encompassed airline ground operations, a domain distinct from shipping both in sociotechnical and cultural aspects. Regardless of the differences, however, many of the core elements of centres of coordination were reified in the ‘shore control centre’ case. Specifically, the need to orient workers to the emerging requirements of safety- and time-critical situations was front and centre. The emphasis of locating technology use within socially organized activities, too, was of central interest, as were the requirements for workers to maintain competencies in reacting appropriately to emerging situations. Additionally, like in airline operations, the marine operations had to be orchestrated across different locations (e.g., port authorities, Vessel Traffic Services, other ferries and ships) and in line with a timetable.

The treatment of technology interactions as a strictly material practice, however, should be re-evaluated in the context of highly automated ships. Centres of coordination originally laid out technologies as an 'assembly of heterogeneous devices' (ibid., p. 44), placing the locus of particular actions at particular technological artifacts. Observing technology trends towards openness and interconnectedness, Monteiro et al. (2013) shifted this locus from mere 'artefacts' to 'information infrastructures,' showing the latter were distinct by virtue of networks that obscure any fixed notions of user, and even time or place of use. Recently, scholars have shifted the locus of interactivity even further from the material boundary, attributing not just agency to computational systems, but also the capacity to enact this agency in their natural environment – conditions allowing for the emergence of behavioural characteristics (Rahwan et al., 2019). Researchers in the field of 'machine behaviour' correspondingly describe as their mission 'the study of ways in which introduction of intelligent machines into social systems can alter human beliefs and behaviours' (ibid., p. 483). Experiments using games have already indicated that interacting with algorithms can increase human collaboration and may even improve group performance (Crandall et al., 2018; Shirado and Christakis, 2017). Whether the same holds true for work collaboration and navigation activities, though, remains uncertain.

Despite this shift away from the materiality of technology interaction, the core issues associated with centres of coordination raised several decades ago by-and-large still apply for today's transition in maritime navigation. Whether framed as artefacts, information infrastructures, or enacted AI agents in the CSCW sense, the interactivity of technologies in socially organized activities is still met with an inherent 'otherness' from their human collaborators. Moreover, the degree of interactivity is accentuated, rather than attenuated, for higher levels of machine autonomy.

6.3 The social implications of AI collaboration

The discrepancies we observed between designers and navigators at the human–machine interface (Sect. 5.3) reinforced the need to reorient design activities towards improved incorporation of user feedback and in-situ observation. Here, we briefly examine the role of social dynamics in this design reorientation, discussing the extent to which discrepancies can be addressed by a better understanding of social implications of AI collaboration.

Discrepancies that arose in designers' and navigators' interview accounts betrayed the ostensible straightforwardness of how decisions are reached in day-to-day work. Work, for navigators, did not unfold as a neatly distilled, stagewise process, as inferred by designers. Rather, the navigators invoked a more intuitive, heuristic decision making based on common sense and tacit knowledge gained from experience. Reflecting on their role, the navigators were more than just 'in the loop' and ready to take preventative action. They were custodians of the automated system, presiding over its operation and arbitrating in its decision making capacity in the context of real-world events. The question of how to address this gap, though, remained largely open.

Methods employed in CSCW may provide useful tools for addressing design challenges presented by developing more aligned collaborative systems. These methods, in contrast to the more prevalent cognitivist and computer science perspectives in AI systems design, consider the sociality of technology use in its natural setting. As Bødker (1991) observed with engineers immersed in computer-aided drawing, the interface between human and computer can become a site in its own right, with its own physical form and possibilities. Revisiting the seminal CSCW control room studies of the 1990s and early 2000s sheds light on how this type of site can form within a socially organised, collaborative setting (see Sect. 2.2). Extending these studies to the case of AI collaboration, what Heath and Luff (1992) described as 'mutual monitoring' in line control rooms, for instance, may be recast in the present context. Mutual monitoring originally involved instances where operators divided their attention between their own tasks and the perception of colleagues' actions through myriad cues, signals, and gestures – subtle yet essential coordinating actions in their work. A parallel can be drawn to modern 'explainable AI' (XAI) techniques, where one strategy involves generating heatmaps tracing where image recognition is 'looking' when classifying an image. In one such example, a machine learning algorithm trained to assist physicians diagnose skin cancer was designed to output details about what pixels it was analysing to reach its predictions (Esteva et al., 2017). Output in so-called 'saliency maps,' the algorithm in effect showed its collaborators 'where it was looking.' A recent review suggests that such collaborative approaches in diagnostics leads to better performance than either physician or AI working alone (Tschandl et al., 2020). Efforts like this are in line with the 'cooperative eye hypothesis,' a theory positing that humans evolved to have large sclera (whites of the eyes) to enable them to follow the gaze of others in cooperative activities, favouring selection of those able to coordinate communicative interactions (Kobayashi & Kohshima, 1997, 2001; Tomasello et al., 2007). Following this logic, enhancing explainability by 'showing where the AI is looking' may be considered among XAI efforts shifting to a more social view of computer interactivity, efforts whose merits are also recognizable in the collaborative control room setting from previous generations of ethnographic CSCW studies.

6.4 Control rooms of the past, present and future

Examining control rooms of the past and present (through literature review, expert interviews, and comparisons to other domains) has compelled us to make inferences about control rooms of the future. Here, we briefly discuss the extent to which such future explorations are rigorous and valid in the sense typically invoked by scientific research. We make the case that despite the speculative nature of our results, they constitute relevant contributions to the CSCW and design communities through their ability to articulate an under-constrained problem and generate design insights.

Our study results were speculative because, although grounded in expert interviews and field observations, they were exploratory in nature and aimed to generate rather than converge new design ideas. The starting point for the research was not a clearly defined problem calling for a clearly defined procedure; on the contrary, it was an under-specified problem calling for a correspondingly open-ended approach. In the design community, such problems often call for a 'research through design' approach (Frayling, 1993), where the goal is generating ideas through a range of pragmatic and conceptual insights. As intimated by Frayling, design is concerned with 'the new,' and as such has a close relationship with research despite the futility of its meeting the rigorous standards of a scientific research method. Inspired by Frayling's thinking, Zimmerman et al. (2007) defended 'research through design' approaches based on their propensity to produce the 'right thing' in the face of under-constrained problems. The approach's underlying contribution, they argued, was based on the strength of its potential to 'transform the world from its current state to a preferred state.' It is partly this preferred state that is so important for the researcher to articulate. In this article, we described the preferred state of future control rooms for highly automated ships through expert interviews as well as through literature review and comparisons to other domains. Set into a multi-disciplinary conceptional framework, this articulation is among the main contributions of the work, asserting that in order to effectively address a problem, it is necessary first to formulate the situation at hand. In this case, it was especially elements of social interactivity in future control rooms that was articulated (who will work in the control rooms, and what will it be like to work with increasingly automated systems?).

Similar scholarly approaches have been applied to design of centres of coordination for highly automated ships, where 'future workshops' stand out as a popular approach (e.g., Hoem et al., 2022; Lützhöft et al., 2019). In this approach, experts are invited to discuss open-ended issues under the pretext of informing design activities. What future workshops have in common with the expert interviews we utilized is that they both imagined future sociotechnical systems that fit a defined situation, shifting the focus from generating tangible solutions to eliciting insights and generating a better understanding (Lindley and Coulton, 2015). Such methods necessarily yield ambiguities, which, like Frayling's 'research through design,' reflect results that, as expressed by Gaver (2012), are 'provisional, contingent, and aspirational.' Yet, the strength of such results lies in its exploration of real issues, gaps, and opportunities, as well as in its ability to articulate the situation at hand. In this sense, the rigour and validity of such future explorations lie in its relevance, however speculative, to designers shaping that future.

6.5 Conceptual limitations and future work

Our practice-based research consisted of interviews in addition to field observations more in line with the CSCW tradition. While ethnographic methods remain indispensable in CSCW research, the addition of interview-based methodologies in our case helped to open the self-contained nature of what Monteiro et al. (2013) call the 'here and now' of field studies. Acknowledging that human–machine interaction is always in transition, such an approach contributed towards a more open-ended treatment of themes that remain constant – the 'otherness' of machine agency, the sociality of technology use – and that offer stable reference points in an inherently transitory study domain.

There were, however, some key limitations of the research design. One such limitation involved the extrapolation of work activities in a maritime navigation setting to that of a control room, and the extent to which this extrapolation provided a representative case. At the time of this study, no shore control centres for coordination of highly automated ships existed in full operational scale. The choice of studying navigation aboard highly automated ferries was used as an approximation for the control centre case. A future shore control centre will, it was argued, be organised around the same core activities – just with higher levels of autonomy and at a remote location. The choice of studying professional design activities towards building shore control centres helped to ground this extrapolation. Although this extrapolation identified several relevant themes, future work must be tuned in to the ways in which human collaboration is affected in a real control centre environment.

7 Conclusion

Maritime navigation work is in transition, marked by collaboration with increasingly high levels of machine autonomy. In this study, we framed this transition as an opportunity to study how designers are shaping work and how navigators are adjusting to the changes. Maritime navigation in this sense served as a representative case study for broader applications of safety–critical, distributed work in sociotechnical, computationally infused environments. Interviews with technology designers and navigators indicated a discrepancy between designers' construals of working with higher levels of autonomy and navigators' own reflective accounts of this work. This discrepancy was centred around the task of taking over control from the automation, a role designers called 'in the loop' and navigators called 'backup.' The discrepancy suggested a need to realign design strategies to real-world operational demands. The risks of not doing so are heightened in the face of increasing levels of autonomy and ongoing development of centres of coordination – efforts that paradoxically place more expectations on human operators, rather than less.

Considering the importance of mutual monitoring – the reflexive social articulations that coordinate work in control room environments – it was clear that collaboration with AI systems depended to a large extent on rendering computational activities more visible. Aligning with the needs of human collaborators involved displaying the AI system's actions more transparently, akin to following the gaze of a collaborator's eyes. Better alignment also pointed to designing AI systems that incorporate cues, gestures, and exclamations of their human collaborators. At least in theory, it may even require machine learning techniques incorporating embedded agency, reflexively adapting to adjustments of users influenced by the AI's presence.

The main contributions of this work are positioned within the emerging field of AI alignment research. Located at the crossroads of computer science, engineering, design, HCI, and sociology, alignment research strives to understand how people can seamlessly interact with machine autonomy. CSCW, with its preoccupation with the sociality of computer use especially in work environments, is uniquely positioned to lend perspective on the transition towards centralized control centres for highly automated maritime navigation. The contribution of this work involved the articulation of the situation at hand to help align design to the preferred real-world interplays of computational plans and human actions. Methodologically, we demonstrated the combined use of literature review, expert interviews, and field observations to ground speculative design insights for future control rooms. Conceptually, the contributions raise the relevance of multi-disciplinary theoretical frameworks and reify theory from past control rooms studies and HCI considerations.

In maritime navigation as in other applications of collaborative work, improvements to efficiency and decision making are among the potential benefits of implementing higher levels of machine autonomy and AI. The extent to which these benefits rely upon seamless coordination with human supervisors, though, remains the domain of research oriented towards the implications of collaborating with intelligent machines in work's natural settings.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the Norwegian Research Council under the MAROFF-2 project ‘Land-based Operation of Autonomous Ships’ (LOAS) directed by project leader Kongsberg Maritime AS and partnered with the Institute for Energy Technology (IFE). The authors also acknowledge the support of the Department of Design at NTNU and strategic funding from the Department of Mechanical and Industrial Engineering at NTNU. The authors gratefully acknowledge the support of doctoral candidates Kim Alexander Christensen and Dag Rutledal who helped with the interviews, as well as to research assistant Elise Korsmo Fjeld for help with visualisations. The authors are immensely grateful for the feedback of several people who reviewed the manuscript, including Eric Monteiro, Elena Parmiggiani, Mariann Villa, Yngve Dahl, and Hedvig Aminoff.

Funding Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital)

Declarations

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

References

Amershi, Saleema; Weld, Dan; Vorvoreanu, Mihaela; Fourney, Adam; Nushi, Besmira; Collisson, Penny; Suh, Jina; Iqbal, Shamsi; Bennett, Paul N.; Inkpen, Kori; Teevan, Jaime; Kikin-Gil, Ruth; and Eric Horvitz (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, 4–9 May 2019*. New York: Association for Computing Machinery, pp. 1–13.

Arrieta Barredo, A.; Díaz-Rodríguez, N.; Del Ser, Javier; Bennetot, Adrien, Tabik; Siham, Barbado, Alberto; Garcia, Salvador; Gil-Lopez, Sergio; Molina, Daniel; Benjamins, Richard;

Chatila, Raja; and Francisco Herrera (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, vol. 58, pp. 82–115.

Bainbridge, Lisanne (1983). Ironies of Automation. *Automatica*, vol. 19, no. 6, pp. 775–779.

Bødker, Susanne (1991). *Through the Interface: A Human Activity Approach to User Interface Design*. Hillsdale, NJ: Erlbaum.

Brekke, Edmund Førland; Wilthil, Erik F.; Eriksen, Bjørn-Olav; Kufoalor, D. K. M.; Helgesen, Øystein K.; Hagen, Inger B.; Breivik, Morten; and Tor Arne Johansen (2019). The Autosea project: Developing closed-loop target tracking and collision avoidance systems. *Journal of Physics: Conference Series*, vol. 1357.

Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile (2012). *Final Report: On the accident on 1st June 2009 to the Airbus A330–203 registered F-GZCP operated by Air France flight AF 447 Rio de Janerio—Paris*. Le Bourget Cedex, France: BEA.

Burmeister, Hans-Christoph; Bruhn, Wilko; Rødseth, Ørnulf Jan; and Thomas Porathe (2014). Autonomous unmanned merchant vessel and its contribution towards the e-Navigation implementation: The MUNIN perspective. *International Journal of E-Navigation and Maritime Economy*, vol. 1, pp. 1–13.

Christian, Brian. (2020). *The Alignment Problem: Machine Learning and Human Values*. New York, NY: W. W. Norton & Company.

Corbin, Juliet; and Anselm Strauss (2015). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (4th Edition). Thousand Oaks, CA: Sage Publications.

Crandall, Jacob W.; Oudah, Mayada; Tennom, Ishowo-Oloko; Fatimah, Abdallah Sherief; Bonnefon Jean-François; Cebrian Manuel; Shariff Azim; Goodrich, Michael A.; and Iyad Rahwan (2018). Cooperating with machines. *Nature Communications*, vol. 9, no. 1, p. 233.

Dallolio, Alberto; Agdal, Bendik; Zolich, Artur; Alfredsen, Jo Arve; and Tor Arne Johansen (2019). Long-Endurance Green Energy Autonomous Surface Vehicle Control Architecture. *OCEANS 2019 MTS/IEEE SEATTLE, Seattle, WA, USA, 27–31 Oct. 2019*. New York: IEEE, pp. 1–10.

Dautenhahn, Kerstin (2007). Socially intelligent robots: Dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1480, pp. 679–704.

Demski, Abram; and Scott Garrabrant (2019). Embedded agency. *ArXiv Preprint* ArXiv:1902.09469.

DNV GL (2017). *Rules for Classification: Ships (Part 5 Ship types, Chapter 4 Passenger ships, Edition January 2017, Amended July 2017)*. Oslo, Norway: DNV GL.

DNV GL (2018). *Remote-controlled and autonomous ships in the maritime industry*. Position paper. Oslo, Norway: DNV GL.

Dunbabin, Matthew; Grinham, Alistair; and James Udy (2009). An Autonomous Surface Vehicle for Water Quality Monitoring. In S. Scheding (ed): *Australasian Conference on Robotics and Automation (ACRA), Sydney, Australia, 2–4 December 2009*. Australia: Australian Robotics and Automation Association, pp. 1–6.

Endsley, Mica R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, vol. 37, no. 1, pp. 32–64.

Esteva, Andre; Kuprel, Brett; Novoa, Roberto A.; Ko, Justin; Swetter, Susan M.; Blau, Helen M.; and Sebastian Thrun (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, vol. 542, no. 7639, pp. 115–118.

Fleischmann, Kenneth R.; Greenberg, Sherri R.; Gurari, Danna; Stangl, Abigale; Verma, Nitin; Day, Jaxsen R.; Simons, Rachel N.; and Tom Yeh (2019). Good Systems: Ethical AI for CSCW. *CSCW '19: Conference Companion Publication of the 2019 on Computer Supported*

Cooperative Work and Social Computing, Austin, TX, USA, 9–13 November 2019. New York, NY, USA; Association for Computing Machinery, pp. 461–467.

Flin, Rhona H.; O'Connor, Paul; and Margaret Crichton (2008). *Safety at the sharp end: A guide to non-technical skills*. Farnham, UK: Ashgate Publishing, Ltd.

Frayling, Christopher (1993). Research in Art and Design. *Royal College of Art Research Papers*, vol. 1, no. 1. London, UK: Royal College of Art.

Gaver, William (2012). What Should We Expect from Research through Design? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, TX, USA, 5–10 May 2012*. New York, NY, USA; Association for Computing Machinery, pp. 937–946

Gjærum, Vilde B.; Strümke, Inga; Alsos, Ole A.; and Anastasios M. Lekkas (2021). Explaining a Deep Reinforcement Learning Docking Agent Using Linear Model Trees with User Adapted Visualization. *Journal of Marine Science and Engineering*, vol. 9, no. 11, p. 1178.

Glaser, Barney G.; and Anselm L. Strauss (1999). *Discovery of grounded theory: Strategies for qualitative research*. Hawthorne, NY, USA: Aldine de Gruyter.

Goerlandt, Floris; and Kenzie Pulsifer (2022). An exploratory investigation of public perceptions towards autonomous urban ferries. *Safety Science*, vol. 145, January 2022, p. 105496.

Goodwin, Charles; and Marjorie Harness Goodwin (1996). Seeing as a situated activity: Formulating planes. In Y. Engeström and D. Middleton (eds.): *Cognition and Communication at Work*. Cambridge, UK: Cambridge University Press.

Heath, Christian; and Paul Luff (1992). Collaboration and Control: Crisis Management and Multimedia Technology in London Underground Line Control Rooms. *Computer Supported Cooperative Work (CSCW)*, vol. 1, nos 1–2, March 1992, pp. 69–94.

Helgesen, Øystein Kaarstad; Vasstein, Kjetil; Brekke, Edmund Førland; and Annette Stahl (2022). Heterogeneous multi-sensor tracking for an autonomous surface vehicle in a littoral environment. *Ocean Engineering*, vol. 252, May 2022, p. 111168.

Hoem, Åsa S.; Veitch, Erik; and Kjetil Vasstein (2022). Human-centred risk assessment for a land-based control interface for an autonomous vessel. *WMU Journal of Maritime Affairs*, vol. 21, no. 2, pp. 179–211.

Hutchins, Edwin (1995). *Cognition in the Wild*. Cambridge, MA, USA: MIT Press.

IATA (2020). *Aircraft Handling and Manual Flying Skills*. Montreal, Quebec: International Air Transport Association (IATA).

IMO (1972). *International Regulations Preventing Collisions at Sea (COLREGs) (adopted 20 October 1972, entered into force 15 July 1977) 1050 UNTS 16 (COLREGs)*. London, UK: International Maritime Organization.

IMO (2018). *IMO takes first steps to address autonomous ships*. Press release, 25 May 2018. http://www.imo.org/en/MediaCentre/PressBriefings/Pages/08-MSC-99-MASS-scoping.aspx

IMO (2021). *Outcome of the Regulatory Scoping Exercise for the Use of Maritime Autonomous Surface Ships (MASS) (MSC.1/Circ.1638)*. London, UK: International Maritime Organization, 3 June 2021.

Kimball, Peter; Bailey, John; Das, Sarah; Geyer, Rocky; Harrison, Trevor; Kunz, Clay; Manganini, Kevin; Mankoff, Ken; Samuelson, Katie; Sayre-McCord, Thomas; Straneo, Fiamma; Traykovski, Peter; and Hanumant Singh (2014). The WHOI Jetyak: An autonomous surface vehicle for oceanographic research in shallow or dangerous waters. *2014 IEEE/OES Autonomous Underwater Vehicles (AUV), Oxford, MA, USA, 6–9 October 2014*. New York: IEEE, pp. 1–7.

Kobayashi, Hiromi; and Shiro Kohshima (1997). Unique morphology of the human eye. *Nature*, vol. 387, no. 6635, pp. 767–768.

Kobayashi, Hiromi; and Shiro Kohshima (2001). Unique morphology of the human eye and its adaptive meaning: Comparative studies on external morphology of the primate eye. *Journal of Human Evolution*, vol. 40, no. 5, pp. 419–435.

Kongsberg (2020). *First adaptive transit on Bastøfosen VI*. Accessed 28 May 2021.

Lechte, John (1994). *Fifty key contemporary thinkers: From structuralism to postmodernity*. London, UK: Routledge.

Lindley, Joseph; and Paul Coulton (2015). Back to the Future: 10 Years of Design Fiction. *Proceedings of the 2015 British HCI Conference*, *Lincoln, Lincolnshire, UK, 13–17 July 2015*. New York, NY, USA; Association for Computing Machinery, pp. 210–211.

Littman, Michael L.; Ajunwa, Ifeoma; Berger, Guy; Boutilier, Craig; Currie, Morgan; Doshi-Velez, Finale; Hadfield, Gillian; Horowitz, Michael C.; Isbell, Charles; Kitano, Hiroaki; Levy, Karen; Lyons, Terah; Mitchell, Melanie; Shah, Julie; Sloman, Steven; Vallor, Shannon; and Toby Walsh (2021). *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report.* Stanford, CA, USA: Stanford University, 16 September 2016.

Lützhöft, Margareta; Hynnekleiv, Agnieszka; Earthy, Jonathan V.; and Erik S. Petersen (2019). Human-centred maritime autonomy—An ethnography of the future. *Journal of Physics: Conference Series*, vol. 1357, p. 012032.

Mahadevan, Karthik; Somanath, Sowmya; and Ehud Sharlin (2018). Communicating Awareness and Intent in Autonomous Vehicle-Pedestrian Interaction. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montréal, QC, Canada, April 21–26, 2018*. New York: Association for Computing Machinery, pp. 1–12.

Martinsen, Andreas B.; and Anastasios M. Lekkas (2018). Curved Path Following with Deep Reinforcement Learning: Results from Three Vessel Models. *OCEANS 2018 MTS/IEEE, Charleston, SC, USA, October 22–25, 2018*. Piscataway, NJ: IEEE, pp. 1–8.

McCarthy, John (2007). What is Artificial Intelligence? Stanford, CA, USA: Computer Science Department, Stanford University, 12 November 2007. http://jmc.stanford.edu/articles/whatisai/whatisai.pdf. Accessed 21 Sept 2021.

MiT (2020). Roboat. http://www.roboat.org. Accessed 20 September 2021.

Monteiro, Eric; Pollock, Neil; Hanseth, Ole; and Robin Williams (2013). From Artefacts to Infrastructures. *Computer Supported Cooperative Work (CSCW)*, vol. 22, nos 4–6, June 2021, pp. 575–607.

Morse, Janice M.; Bowers, Barbara J.; Charmaz, Kathy; Corbin, Juliet; Clarke, Adele E.; and Phyllis Noerager Stern (2009). *Developing grounded theory: The second generation*. Walnut Creek, CA, USA: Left Coast Press Inc.

National Transportation Safety Board (2020). *Collision Between a Sport Utility Vehicle Operating with Partial Driving Automation and a Crash Attenuator, Mountain View, California, March 23, 2018 (Accident Report NTSB/HAR-20/01).* Washington, DC, USA: National Transportation Safety Board, 25 February 2020.

Nevile, Maurice (2001). *Beyond the black box: Talk-in-interaction in the airline cockpit*. Ph.D. dissertation. The Australian National University, Canberra.

Nicas, Jack; Kitroeff, Natalie; Gelles, David; and James Glanz (2019). *Boeing built deadly assumptions into 737 Max, blind to a late design change*. New York, NY, USA: The New York Times.

NVivo (1.0) (2020). [Computer software]. QSR International.

Park, Sun Young; Kuo, Pei-Yi; Barbarin, Andrea; Kaziunas, Elizabeth; Chow, Astrid; Singh, Karandeep; Wilcox, Lauren; and Walter S. Lasecki (2019). Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare. CSCW '19: *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing, Austin, TX, USA, 9 – 13 November 2019*. New York, NY, USA: Association for Computing Machinery, pp. 506–510.

Peeters, Gerben; Yayla, Gökay; Catoor, Tim; Van Baelen, Senne; Afzal, Muhammad Raheel; Christofakis, Christos; Storms, Stijn; Boonen, René; and Peter Slaets (2020b). An Inland Shore Control Centre for Monitoring or Controlling Unmanned Inland Cargo Vessels. *Journal of Marine Science and Engineering*, vol. 8, no. 10, September 2020.

Rahwan, Iyad; Cebrian, Manuel; Obradovich, Nick; Bongard, Josh; Bonnefon, Jean-François; Breazeal, Cynthia; Crandall, Jacob W.; Christakis, Nicholas A.; Couzin, Iain D.; Jackson, Matthew O.; Jennings, Nicholas R.; Kamar, Ece; Kloumann, Isabel M.; Larochelle, Hugo; Lazer, David; McElreath, Richard; Mislove, Alan; Parkes, David C.; Pentland, Alex 'Sandy'; … and Michael Wellman (2019). Machine behaviour. *Nature*, vol. 568, no. 7753, pp. 477–486.

Rasmussen, Jens (1986). *Information Processing and Human-machine Interaction: An Approach to Cognitive Engineering*. New York, NY, USA: Elsevier Science Publishing Co.

Reddy, Namireddy Praveen; Zadeh, Mehdi Karbalaye; Thieme, Christoph Alexander; Skjetne, Roger; Sorensen, Asgeir Johan; Aanondsen, Svein Aanond; Breivik, Morten; and Egil Eide (2019). Zero-Emission Autonomous Ferries for Urban Water Transport: Cheaper, Cleaner Alternative to Bridges and Manned Vessels. *IEEE Electrification Magazine*, vol 7, no. 4, November 2019, pp. 32–45.

Rødseth, Ørnulf Jan. (2017). *Definitions for Autonomous Merchant Ships*. Trondheim, Norway: Norwegian Forum for Autonomous Ships.

Rolls-Royce (2018). Rolls-Royce and Finferries Demonstrate World's First Fully Autonomous Ferry. Press release, 3 December 2018. https://www.rolls-royce.com/media/press-releases.aspx. Accessed 1 October 2020.

SAE International (2017). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*. Warrendale, PA, USA: Society of Automotive Engineers (SAE) International.

Shirado, Hirokazu; and Nicholas A. Christakis (2017). Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, vol. 545, no. 7654, pp. 370–374.

Shneiderman, Ben (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, vol. 36, no. 6, March 2020, pp. 495–504.

Stone, Peter; Brooks, Rodney; Brynjolfsson, Erik; Calo Ryan; Etzioni, Oren; Hager, Greg; Hirschberg, Julia; Kalyanakrishnan, Shivaram; Kamar, Ece; and Sarit Kraus (2016). *Artificial intelligence and life in 2030: The one hundred year study on artificial intelligence*. Stanford, CA: Stanford University.

Suchman, L. (1997). Centers of Coordination: A Case and Some Themes. In L. B. Resnick, R. Säljö, C. Pontecorvo, & B. Burge (eds): *Discourse, Tools and Reasoning: Essays on Situated Cognition*. Berlin, Germany: Springer, pp. 41–62.

Suchman, Lucy A. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions* (2nd Edition). Cambridge, UK: Cambridge University Press.

Tomasello, Michael; Hare, Brian; Lehmann, Hagen; Josep Call (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: The cooperative eye hypothesis. *Journal of Human Evolution*, vol. 52, no. 3, pp. 314–320.

Tschandl, Philipp; Rinner, Christoph; Apalla, Zoe; Argenziano, Giuseppe; Codella, Noel; Halpern, Allan; Janda, Monika; Lallas, Aimilios; Longo, Caterina; Malvehy, Josep; Paoli, John; Puig, Susana; Rosendahl, Cliff; Soyer, H. Peter; Zalaudek, Iris; and Harald Kittler (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, vol. 26, no. 8, pp. 1229–1234.

Utne, Ingrid Bouwer; Rokseth, Børge; Sørensen, Asgeir J.; and Jan Erik Vinnem (2020). Towards supervisory risk control of autonomous ships. *Reliability Engineering & System Safety*, vol. 196, pp. 106757.

Veitch, Erik; and Ole Andreas Alsos (2022). A systematic review of human-AI interaction in autonomous ship systems. *Safety Science*, vol. 152, pp. 105778.

Vicente, Kim J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Boca Raton, FL: CRC Press.

Voosen, Paul (2017). The AI detectives. *Science*, vol. 357, no. 6346, July 2017, pp. 22–27.

Wang, Wei; Gheneti, Banti; Mateos, Luis A.; Duarte, Fabio; Ratti, Carlo; and Daniela Rus (2019). Roboat: An Autonomous Surface Vehicle for Urban Waterways. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019*. New York: IEEE, pp. 6340–6347.

Waymo (2020). Waymo Safety Report. https://waymo.com/safety. Accessed 13 September 2021.

Whalen, Jack (1995). A technology of order production: Computer-aided dispatch in public safety communications. In P. Have, G. Psathas (eds): *Situated Order: Studies in the Social Organization of Talk and Embodied Activities*. Boston, MA, USA: International Institute for Ethnomethodology and Conversation Analysis, pp. 187–230.

Wu, Baiheng; Li, Guoyuan; Wang, Tongtong; Hildre, Hans Petter; and Houxiang Zhang (2021). Sailing status recognition to enhance safety awareness and path routing for a commuter ferry. *Ships and Offshore Structures*, vol. 16, no. 1, pp. 1–12.

Zimmerman, John; Forlizzi, Jodi; and Shelley Evenson (2007). Research through Design as a Method for Interaction Design Research in HCI. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *San Jose, CA, USA, 28 April – 3 May 2007*. New York, NY, USA; Association for Computing Machinery, pp. 493–502.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix C8: Academic contribution 8

Dybvik, H., Veitch, E., & Steinert, M. (2020). EXPLORING CHALLENGES WITH DESIGNING AND DEVELOPING SHORE CONTROL CENTERS (SCC) FOR AUTONOMOUS SHIPS. Proceedings of the Design Society: DESIGN Conference, 1, 847–856. https://doi.org/10/ggz7zb

C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16

INTERNATIONAL DESIGN CONFERENCE – DESIGN 2020
https://doi.org/10.1017/dsd.2020.131

EXPLORING CHALLENGES WITH DESIGNING AND DEVELOPING SHORE CONTROL CENTERS (SCC) FOR AUTONOMOUS SHIPS

H. Dybvik✉, E. Veitch and M. Steinert

Norwegian University of Science and Technology, Norway

✉ henrikke.dybvik@ntnu.no

Abstract

The concept of remotely operated, unmanned, and autonomous ships is creating increasing interest in the maritime domain, promising safety, increased efficiency and sustainability. Shore control centers (SCCs) have been proposed to operate such vessels and some industry projects are initiated. This paper aims at bringing knowledge about what a SCC is envisioned to be. It identifies and explores challenges related to designing and developing SCCs through semi-structured interviews with the research community and industry. We discuss tasks, functions and interactions between human and machine.

Keywords: autonomous ships, shore control center, design knowledge, information retrieval, early design phase

1. Introduction

Autonomous ships are creating increasing interest in the maritime domain, promising increased efficiency and sustainability. It is the goal of unmanned vessels to be at least as safe as manned vessels, in part because it moves seafarers away from potentially hazardous operations. An autonomous ship is a vessel with the possibility of operating on one or more Levels of Autonomy (LOA). This ranges from vessels with automated processes and decision support to remotely controlled and fully autonomous ships operating with or without seafarers onboard. Marine Autonomous Surface Ships (MASS) is the term suggested by the International Maritime Organization (IMO) to cover this broad range of automated and remotely controlled ships. Shore control centers (SCCs) have been proposed to operate such vessels (Levander, 2017; Lützhöft and Dekker, 2002; Rødseth et al., 2018) and to this date some industry projects have been initiated (MUNIN, 2015; Wilhelmsen, 2018).

Despite the recent interest in SCCs, there are currently no accepted guidelines for SCC design. As a result, some industry projects have approached the problem by replicating the ship's bridge onshore (see Figure. 1 for a prototype of an SCC and a traditional ship bridge). Research and systematic testing of alternative design approaches is lacking.

This paper aims at bringing forward knowledge about what an SCC is envisioned to be. It identifies and explores challenges related to designing and developing SCCs through semi-structured interviews with the research community and industry. We discuss expected tasks, functions, and interactions in the SCC and identify and describe associated design challenges.

Section 2 follows this introduction, providing background on why SCCs are needed and offers definitions for SCC and LOA. Section 3 covers methodology and Section 4 presents findings. Findings include how interviewees envisioned the SCC as well as what crucial interactions and design challenges might be encountered. Section 5 summarizes the work in a discussion; Section 6 presents a conclusion.

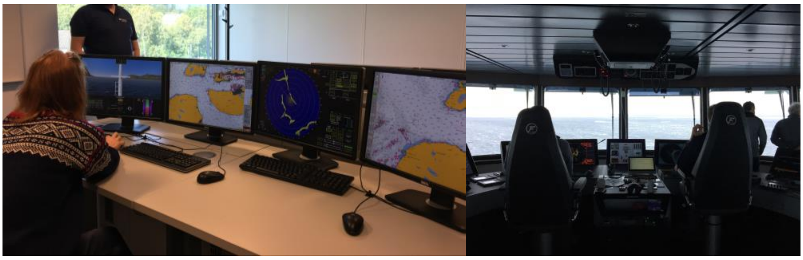


Figure 1. A prototype of an SCC (left) and a traditional ship bridge (right)

2. Background

2.1. The need for SCCs

The SCC is needed to monitor one or more autonomous ships remotely and to intervene in their navigation, if necessary. The word "autonomous" here does not describe a fully autonomous ship; rather, the moniker covers a range of autonomy levels onboard the ship's control system that stops before full autonomy. At Level 2 and Level 3 automation, according to the International Maritime Organization's (IMO) definition (IMO, 2018), the automation is not enough to allow the vessels to navigate without human supervision and the option to intervene via remote control is necessary. MUNIN (2015) represents the first study specifically undertaken to investigate design needs for SCCs. The project asserted that without a continuously manned SCC, design of an autonomous ship system would be very challenging. Leading design guidelines like the UK Code of Practice (Maritime UK, 2018) which to date only applies to MASS under 24 m in length, also highlight the importance of a "remote controller:" a person with whom responsibility of the vessel is assumed to lie and whose situational awareness and overview of the ship's mission is prerequisite for the operation's success. From the risk and reliability engineering perspective, too, have novel methods like Human-System Interaction in Autonomy (H-SIA) placed the SCC in a role of central importance as a mechanism for failure propagations of MASS scenarios (Ramos et al., 2019). Furthermore, autonomous and remotely operated ships operated by people onshore are expected to be safer, more efficient and cheaper to run (Ahvenjärvi, 2016; Levander, 2017), and reduce human error and workload (Lützhöft and Dekker, 2002). The risk of human errors due to fatigue is expected to reduce, leading to reduced risk of injuries to crew, ship, and cargo. Another advantage is the possibility to design these ships with a larger cargo capacity and reduced wind resistance. Certain features of today's ships, such as the deck house, the crew quarters, some ventilation, and sewage systems can be eliminated without crew onboard the ship. Despite advantages associated with introducing automation in the system, this also creates pathways to new technological and human errors, both of which have consequences (Lützhöft and Dekker, 2002).

Nevertheless, there is a general consensus that a fully autonomous vessel requiring no human input will likely not exist. Instead, a ship with highly automated control systems will require support from a human crew in the form of a Shore Control Center.

2.2. SCC definitions

While many definitions have been proffered by researchers and stakeholder organizations, there remains to be a widely accepted definition of the SCC. To date, even the terminology lacks unanimity, with the acronyms Shore Control Center (SCC), Remote Control Center (RCC), and Base Control Station (BCS) all appearing to describe what is by-and-large the same thing (Maritime UK, 2018; MUNIN, 2015; Rødseth et al., 2018; Rødseth and Nordahl, 2017). Remote Operating Center (ROC) has also been used in the research community (MTEC/ICMASS 2019, personal communication, 13-14. November 2019). In this work we use SCC. Rødseth and Nordahl (2017) defined the SCC simply as the "Owner's center for monitoring and control" and offered the following explanation of its function:

> *[The SCC] will be used partly as a backup in case the ship encounters unexpected events, partly to reduce the required complexity of on-board detection and control systems and partly to satisfy legal requirements that some human is in control of the ship.*

This definition positions the SCC as a backup and legal requirement, rather than the central part of a wider monitoring and control system. MUNIN (2015) hypothesized the SCC in more specific terms, describing a control room in which each operator monitors and controls six vessels at a workstation containing six screens. Furthermore, requirements in the MUNIN set-up for operators to quickly respond to unexpected hazards raised questions about difficulties the operator may experience as "automation supervisor," known as "Out-Of-The-Loop" phenomenon (Kaber and Endsley, 1997). Ethnographic research examining the future skills of SCC operators points to unreasonably high demands on human capabilities given the current vision for the SCC (Lutzhoft et al., 2019).

2.3. Levels of Autonomy (LOA)

The concept of Levels of Autonomy, or levels of automation (LOAs) in human-machine interaction has been used to define which functions should be managed by the autonomy and which ones should be managed by the human operator (Musić and Hirche, 2017). For instance, Sheridan and Verplank (1978) defined 10 discreet LOAs, ranging from no autonomy to full autonomy. Endsley (1987) developed a 5-level LOA taxonomy regarding how much decision support would supplement human decision making during a cognitive task, ranging from manual support, to decision support, consensual artificial intelligence (AI), monitored AI and finally, full automation with no operator interaction.

In the context of autonomous ships and remotely operated vessels, several definitions of LOAs have been proposed. Autonomy Level 4, therefore, does not refer to full autonomy, but rather to "self-controlled function," whereby the system will execute the operation, but a human is able to override the action. IMO has identified four degrees of autonomy as a scoping exercise (IMO, 2018). Lloyd's Register (2016) has also defined 5 levels of autonomy in a design guideline for autonomous and remotely operated ships in terms of cyber access, ranging from complete human control (no cyber access; Level 1) to cyber access for autonomous/remote monitoring and control (onboard override not possible; Level 5). A 5-level definition has also been proposed by DNV-GL (DNV GL, 2018). It describes stepwise degrees in automatic control, where the Lloyd's definition is framed as discrete cyber accessibility levels. Rødseth et al. (2018) suggested building a framework based on SAE J3016 standard for autonomous cars and defined five basic types of autonomy in merchant vessels.

Generally, the LOA is expected to vary during the voyage according to level of risk and complexity. Accordingly, during low-autonomy sailing, human operators are expected to be highly engaged; during high-autonomy sailing, the operators are mainly in a supervising role.

3. Method

For an explorative study, a qualitative, adaptive approach using semi-structured interviews was chosen (Yin, 2017). First, the maritime ecosystem was mapped using the methodology behind a Customer Value Chain Analysis (Donaldson et al., 2006) to identify relevant interviewees. A total of 8 interviewees included Subject Matter Experts (SMEs) in the field of autonomous shipping and represented academic institutions, research organizations, and private sector companies in Norway. We used convenience sampling and snowball sampling through our own network. The interviewees were invited to take part via email.

Each interview lasted from 45-60 minutes and was conducted both in-person and via video call. Two researchers attended, one led the interview by asking questions and one wrote notes. Oral consent was obtained before beginning the interview. Beforehand, an interview protocol and an interview guide were made with predefined main questions. The interview began with basic questions about interviewees' backgrounds and current work. The main questions were selected to shed light on the study's primary investigation areas; namely, defining SCC, outlining future design challenges, defining operations, and critical human-system interactions. Towards the end we focused on the critical interactions as defined by the interviewee. Then we asked them to argue for the need for an SCC and give their impression of how the SCC business model would function. The questions were open-ended due to the explorative nature of the study, allowing the interviewee free train of thought and enabling unexpected insights. Flexibility around the questions enabled the interviewer to follow up on a specific train of thought and adapt the protocol with more specific questions. It also enabled the participants and interviewer to reach a mutual understanding in the instance where a question needed clarification or rephrasing, thus leading to more accurate data (Dörnyei, 2007).

The data was analyzed using cross case-synthesis (Eisenhardt, 1989; Yin, 2017), treating each interview as if it were a separate case.

4. Findings

4.1. Envisioning the SCC - How is an SCC defined?

This section describes how the interviewees communicated their vision for the SCC. Comparisons with and analogies to other forms of control centers were frequently mentioned by interviewees when envisioning the SCC. This was especially the case when describing physical infrastructure, data sources, and functions the SCC would perform.

The main purpose of SCCs is to provide the ability to take control of autonomous vessels from a remote location, especially as means to avoid critical situations, collisions, and allisions that are outside the capability of the automatic navigation algorithms. Thus, the display of ship-board sensory information was important. The operator's work inside the SCC may be characterized primarily as passive monitoring, similar to that of current-day Vessel Traffic Services (VTS). VTS operators are provided with information and have inputs enabling them to manipulate ships indirectly. The SCC will have access to similar data sources, such as Electronic Chart Display and Information Systems (ECDIS), VHF Marine Radio, ship schedule and voyage plan information, etc. The key difference between SCC and VTS operators is that the former has the ability to directly influence a ship's navigation as if they were in the bridge of the ship, only from afar.

When envisioning the physical SCC infrastructure, a set-up involving a combination of large screen displays and personal workstations was described to simultaneously provide a generalized "big picture" overview of the ships, along with the ability to search for and view more detailed information. Today, we need a clean setup. The layout of the room was important, and should be designed to minimize disturbances, provide good sightlines to the large-screen displays, and generally promote a sense of professionalism and ownership to the operators. Information should be shown according to priority to not overload the operator. Some descriptions revealed a resemblance to modern-day full-mission ship bridge simulators, with the key difference that it navigated a real ship and contained upgraded ECDIS with the possibility to view and adjust tracked trajectories with the means of a drag-and-drop feature. The argument was that current operators are used to being in a ship and therefore the more you can replicate a real ship, the better. Still, those who envisioned an immersive simulator environment conceded that this would not be necessary to get an acceptable situational awareness in an SCC.

An important distinction emerged between the functions of pure remote control and that of an SCC. Some industry projects are essentially copying the bridge and positioning it on-shore with a telecommunication link - an instance of remote control. On the contrary, the primary function of SCCs is the ability to adapt to different types of ships from one center. SCCs can be characterized as one center for multiple vessels of different types. Purely remote control "base stations," on the other, are one-to-one. This paper is not concerned with pure remote control.

A theme of transition from traditional to autonomous shipping emerged from the interview findings. SCC technology would not be disruptive; rather, it would be incremental by necessity of a highly regulated industry. This applied especially to competence needs, which is outlined in Section 4.3.5, whereby the SCC operators' qualifications may transition from those of experienced seafarers to those of younger operators with only virtual ship handling experience, who, while fully capable in their assigned role, would not hold the marine certifications that their predecessors once did. No one specified how long the transition would last, but several interviewees described a starting phase where the SCC would primarily be used for testing, verification, and approval. Later, the SCC would settle into overall transportation logistics management and might eventually coordinate to a higher degree with other available resources, like VTS. As such, the SCC they expected to see in the near future was distinctly different from the SCC they envisioned in the distant future.

4.2. Crucial interactions in a future SCC

The human-machine interaction was said to be the most critical interaction in the future SCC. A "*good connection*" between the human and the automation system was needed. A prerequisite trust in the interaction system and automation was fundamental. The notion of trust includes buy-in from the industry, particularly from the shipowners' side in addition to trust from seafarers who will navigate ships alongside their unmanned counterparts. One interviewee said seafarers were sceptical of technologies that removed them from the location in which their work was done, since seafarers relied heavily upon their senses of sight, smell, sound, and proprioception. One example was given of an engine room operator's complaint, made during a participatory design process, about relocating a ship's engine control room far away from the engine room. The operator cited dissipation of vibration, sounds, and smells distant from the source as a potential hindrance to the operator's duties. Notwithstanding, the introduction of conditional monitoring software was accepted by the operator as making up for what may be lost by in-situ human monitoring, while having the benefit of placing the operator at a location more amenable to safe working conditions.

Researchers interviewed in this study were especially prone to citing human factor considerations as among the most important when designing the human-machine interactions and working environment in the SCC. Design of alarms was frequently mentioned as a method of communication that, if designed correctly, results in efficient detection of system prompts in order of priority.

4.3. SCC design challenges

4.3.1. Handling the old model of seafaring

Remote operation in the SCC was framed as a new paradigm in ship navigation. Yet, an old model of seafarers' traditional roles was acknowledged by many interviewees to dominate the emerging SCC mental model. Interviewees often described the imagined human-system interface in the SCC as the same that is found on-board contemporary ship bridges and VTS centers. They also commonly described the physical infrastructure of the SCC as a ship bridge moved to shore, albeit with incremental technology improvements. These assumptions appeared to bias some interviewees' reflections on SCC design challenges. For example, some described the challenges associated with moving the bridge to shore, confining the image of the SCC to that of ship's bridge with traditional bridge resource management. Some interviewees eschewed this assumption, instead framing the SCC as a new design space challenge. It appeared that copying the bridge onshore may not be the best solution and underscored the importance of separating the role of the SCC from that of the ship's bridge, including the competence needs of the operators occupying both spaces.

4.3.2. Information display design

With higher level technology and increasing automation there is an increasing access to a larger body of information, including different data sources and sensor information. Selecting which information to display and designing how it should be presented to provide the best possible overview of the situation is a challenge. One interviewee offered the model of top-down information search capability, in which the operator can effectively find specific information using a sequence of commands that narrows to the sub-system of interest. This top-down search should be complemented with a bottom-up information display model, in which specific information can quickly and effectively be placed within the larger context of the operation. An overarching task for the designer is to consider what information is displayed at what times, and how that information is presented in a way that does not overload operators.

4.3.3. The human-automation handover design

Designing the human-machine interface will be the most challenging part of SCC design, according to most interviewees. In particular, the handover; automation to human control will require special attention. During such a handover, some time will be required by the operator to gain situational awareness - time that is referred to as *"getting in the loop"*. It is believed that with increasing autonomy the SCC will have less to with the voyage of the vessel, requiring the operator to monitor the vessel continuously. Should something disrupt the autonomy and require the operator to get into the loop, this adjustment time may be safety critical. Lowering the amount of time to get into the loop was one of the most important challenges cited. Currently, there is no test criteria regarding handover from automation to human, though one interviewee believed a systematic test approach was needed to approve SCCs. Results of studies from autonomous cars suggest handover times of approximately seven second (Gold et al., 2013), but there is a lack of experiments to test how long a handover takes in a ship from an SCC. Several interviewees in this study suggested that allowable ship takeover times may be longer than for cars, since ships have a relatively slow cruising speed.

The opposite direction of control handover, namely from human to automation, was considered less of a challenge. It was suggested that this process would be gradual and controlled, whereby the operator would hand over control, possibly in a stepwise process, and subsequently monitor the system for some time to ensure a successful handover.

4.3.4. Communication

Effective communication was highlighted by several interviewees as an essential component of the SCC design. SCC operators will have to communicate with VTS, port authorities, and other vessels. If it is a cargo vessel and it is part of a wider logistics system, there must also be communication with crane operators. One interviewee pointed out that modern regulations stipulate continuously being tuned in to VHF-radio onboard each ship, which would make communication chaotic when operating several vessels simultaneously. This was also mentioned by several other interviewees; having more than one communication channel would be an issue. They questioned how many vessels one can communicate with.

It was noted that limited bandwidth is a potential barrier to effective communication. This issue appears to escalate the farther offshore the vessel is located. Specifically, mobile phone broadband does not cover waters farther than coastal areas, and satellites are known to have latency issues and prohibitive cost associated with video and image transfer.

4.3.5. Unknown skill requirements of SCC operators

The skill requirement of the operator is largely unknown. Although there is currently research underway for SCC curricula, the emphasis has been mainly on hypothesizing a learning framework rather than on developing a tested and verified training program. Some interviewees suggested that opportunities for testing will present themselves with the first large-scale SCC applications and not before.

It was acknowledged that the skills necessary for SCC operators would be distinctly different from those of traditional seafarers. Still, it was conceded by most interviewees that during the initial stages of development, a trained seafarer with the required certificates in navigation and watchkeeping would have to be employed. Only when legal and regulatory adjustments are made for competence needs of SCCs will operators be trained in the core skills required for the job; namely, remote monitoring and control.

4.3.6. Documentation and regulation

The maritime industry is one of the most regulated industries. Work on-board ships follow strict protocols, procedures, and documentation processes. Ironically, some tasks are apparently conducted only for the sake of the documentation requirements themselves. Audits and inspections for safety management systems are in large part the cause for this inefficient process. For work to function seamlessly in the SCC, efficient documentation was therefore highlighted as an organizational design challenge. Instead of using documentation as a tool for passing inspection, other methods could be used for safety management. Video surveillance was suggested as a substitute for textual documentation, since it may provide the potential to record work procedures and automatically collect documentation electronically.

Regulations for SCCs are currently underdeveloped and have the potential to limit their value. One example was number of allowable ships monitored by one operator; if the number is below a certain threshold - or, worse, is only one - then the business case for remote monitoring and control would limit the intended function of the SCC.

4.3.7. How much automation is automation?

There are several definitions of LOAs, none of which to our knowledge specifies exactly how much autonomy. Constrained autonomy was suggested as one way to clarify the boundaries of the automation system. It stems from the issue that maximizing intelligence may yield rigid boundaries in what the automation can and cannot do, resulting in abrupt handover exchanges with the operator. Constraining the automation such that self-limits its own functionality can circumvent this issue. If the automation detects, for instance, that the level of complexity in ship traffic has reached a certain level, rather than attempt to solve the situation alone, the system alerts the operators of an oncoming potential (and optional) handover. The operator is thus recruited to assess the situation and to judge whether human control is needed. This design has several additional benefits: improved communication of intent yields a higher degree of transparency and effectively stymies the Out-of-the-Loop phenomenon. It is a drawback, too; namely, increased alarms. With good design implementation, though, alarms are thought to help normalize workload away from a more disjointed peak-trough pattern, potentially serving to improve vigilance over time.

4.4. LOA task dependence

All 8 interviewees agreed that tasks in the SCC will vary according to the LOA. A picture of adaptive automation emerged, wherein the LOA would adjust according to the voyage demands. When prompted to provide details for which tasks related to which LOA, responses were more abstract and ambiguous. Uncertainty around the definition of LOA appeared to be partly the cause for the lack of clear answers. Most interviewees cited *"level 4 automation"* as the highest attainable LOA on-board an autonomous ship. This referred not full autonomy, but rather to the self-controlled functionality whereby the ship system execute operation and a human is able to override it from the SCC. Interviewees were therefore referring to a LOA definition more in line with that of DNV-GL (DNV GL, 2018) than the IMO definitions that describe level 4 as fully autonomous.

4.5. Challenging the assumption of if the SCC is needed

Interviewees were asked whether the SCC was really needed and thus encouraged to challenge the underlying assumption of the study. All interviews agreed that the SCC was needed, but added several key exceptions.

Firstly, a distinction was made between small-scale autonomous vessels and their larger counterparts. In the former, the requirements for an SCC was questioned due to the shorter distances, proximity to shore, and overall lower perceived risk. For data collection vessels especially, like in the case of seabed surveys and hyperspectral imagery, it was noted that software can read the data better than humans at an SCC. For larger vessels, like cargo ships, some interviewees believed it will take many more years for the SCCs to develop due to the higher risk they posed to other ships.

Secondly, demand in traditional seafaring jobs and a shift in the seafarers' work content underscored the need for SCCs. Seafaring jobs were perceived to be less desirable among young people. The option to work from shore may appeal to some recruits. Moreover, more of the work on-board ships was perceived to be automated and screen-based, suggesting that the evolution to remote control is going to be incremental. The notion of *"dirty, dull and dangerous"* tasks, words which are often used to describe types of work that justify automation, was said by several interviewees as being prevalent during ship navigation.

One interviewee noted that should full autonomy be developed in the future, the SCC would not be needed.

4.6. The business model of SCCs

Several interviewees had given considerable thought to the business case of the SCC and shared their ideas with us; the majority, however, did not. The most immediate value proposition of automation ships appeared to be removal of humans from the vessel, saving cost and freeing space associated with hotel infrastructure and operation. Especially for smaller work vessels, removal of the hotel and amenities removes a considerable portion of the overall vessel infrastructure. An SCC is thus more cost-effective for smaller vessels. The value proposition is compounded with the ability of monitoring and controlling several vessels simultaneously.

It was noted that to date, traditional ship owners are not interested in autonomous ships. Rather, logistic and cargo distributions companies are showing the most interest. The shorter crossing and slower speeds of autonomous ships seem to suit areas where cargo distribution is needed, and predictable scheduling and broader logistics benefits outweigh the higher CAPEX of autonomous vessels. In some cases, the reduction of truck transportation is a major cost-saving element. The electrification of automation ships also meets public and stakeholder pressure on sustainability targets. From this point of view, the autonomous ships and the SCC represent a broad transportation system, distinct from the tradition model of shipping.

The need for volume was critical to the business case. One operator must be allowed to monitor more than one vessel simultaneously. The more vessels one operator (or one team of operators) could effectively manage, the better it was for profitability.

It was noted that insurers are general positive towards the notion of autonomous shipping. International P&I clubs incur most costs due to payments associated with crew injuries.

4.7. Sketching potential solutions

Some interviewees challenged designers to frame the problem of SCC design with a higher degree of abstraction than simply moving the ship's bridge to shore. The process should start with the SCC and its unique requirements, not with how to solve the technical problems of a ship's remote control. Such a design strategy might result in simple and effective solutions.

Some specific solutions were offered by interviewees to address various components of the SCC design. For instance, regarding the limits of telecommunications (outlined in Section 4.3.4), one suggestion was to recreate digital imagery of objects detected in-situ, rather than continuously sending video and imagery to the SCC. In other words, the visual representation of detected obstacles, including their location in space, could appear as a simulation in the SCC, removing the need for expensive, limited, and delayed transfer of images.

Almost all interviewees highlighted the Out-of-the-Loop phenomenon as a major design challenge for SCCs. One interviewee suggested the "constrained autonomy" could be an effective mitigator (Section 4.3.7). The constrained autonomy model also improves transparency and may normalize workload and fit well into a larger framework of adaptive automation.

Finally, experimentation was highlighted as a key to development of SCCs. Currently many critical elements - like the number of vessels one operator can manage, for instance - are hypothesized. Systematic testing and verification are necessary for approval and acceptance.

5. Discussion

Many interviewees compared the SCC to control centers found in other domains and industries. VTS and air traffic controller (ATC) were most frequently mentioned. Power plants, space centers, centralized fish farming, train line monitoring, and autonomous cars were also mentioned as applications of remote control from which SCCs could adopt design elements.

In the ship navigation, rigid hierarchical roles define bridge management. One interviewee challenged the assumption that SCC operators will follow the same order of command, suggesting that the SCC may be a venue better suited for flatter organizational structure. In any case, SCC resource management will be an area for design consideration.

The concept of *"ship sense"* was occasionally discussed by interviewees, most often in the context of a set of distinct human qualities - be that on the bridge, engine room, or elsewhere - that operators rely upon to do their job well. These human qualities, explicitly mentioned by interviewees, included the senses of smell, touch, sight, and a sense of *"embodiment"* which the authors understand to mean proprioception. Interviewees expressed that removing the operators from the location of their work would effectively hamper ship sense, tacit knowledge required for operators to navigate ships effectively. While interviewees generally agreed on this definition, there emerged two school of thought regarding its role in the SCC. Some were in favour simulating ship sense in the SCC, with examples such as integrated motion beds, vibrating seats, and smell alarms. Others were not in favour of this concept, stating that a combination of condition monitoring sensors and alarms were more reliable than the human senses, and that distance from the sources of risk may enable safer and more effective decision making. Ahvenjärvi (2016) asks if the best of the human element will be lost by autonomous ship technology. We extend this question to how human factors can be applied in SCC design research to best adapt to human sensemaking, especially in safety critical situations.

All interviewees expressed a high degree of trust in current automation technology. This is interesting because it's remarkably different from other interview-based studies about human factors needs in the transition to automation shipping, which revealed trust issues as a major concern (Mallam et al., 2019). The difference may be because of the shift in interviewees' focus; interviewer envisioning operators in a monitoring and control role may place more trust in the autonomous systems compared to a third party, like another ship.

It was accepted that one operator (or one SCC team unit) is required to operate several vessels for it to be a viable business case. The questions of *how many* vessels remains unanswered. Many interviewees cited studies that showed that this number was 6 to 10 vessels. This refers to the MUNIN project, in which 6 vessels was set as the upper limit; however, MUNIN (2015) states that that this number is hypothesized and as of today still not verified. In any case, setting a fixed upper limit may be problematic, since manoeuvres like docking require more attention than open-water transiting. The number will also be dependent on additional variables, such as location, voyage, and LOA. The maximum number of ships one operator can manage is therefore likely to encompass a wide range rather than a single number.

Many interviewees were mentally fixated with the old model of seafaring. Such a design fixation (Moreno et al., 2015) might be a problem both during user testing and during design work, and must be overcome.

6. Conclusion

The concept of remotely operated, unmanned, and autonomous ships is creating increasing interest in the maritime domain, promising safety, increased efficiency and sustainability. SCCs have been proposed to operate such vessels and some industry projects are initiated. This paper aims at bringing knowledge about what a SCC is envisioned to be. It identifies and explores challenges related to designing and developing SCCs through semi-structured interviews with the research community and industry. We discuss tasks, functions and interactions between the human operation and autonomous ship control system. There are many unknowns and challenges regarding design, development and operations of SCCs. These include, but are not limited to, the human-automation handover, management of the transition from traditional shipping, design fixation, information display design, communications, and number of vessels an operator can handle. Currently, many critical elements - like the number of vessels one operator or one team can manage - are hypothesized and represent important research tasks. Experimentation and systematic testing may unlock solutions to SCC design challenges and will be necessary for regulatory approval and user acceptance. SCCs represent an emerging complex socio-technical system and introduces many new risks and unknown unknowns. Considering the consequences of failure stemming from poor implementation, design research is needed to address the challenges associated with incorporating the human in the SCC as operator of autonomous ships.

Acknowledgement

The authors would like to thank interviewees for their time and contribution. This research is supported by strategic funds from the department for Mechanical and Industrial Engineering (MTP) at NTNU.

References

Ahvenjärvi, S. (2016), "The Human Element and Autonomous Ships", *TransNav Int. J. Mar. Navig. Saf. Sea Transp.*, Vol. 10. https://doi.org/10.12716/1001.10.03.18

DNV GL. (2018), Class Guideline DNVGL-CG-0264 Autonomous and remotely operated ships.

Donaldson, K.M., Ishii, K. and Sheppard, S.D. (2006), "Customer Value Chain Analysis", *Res. Eng. Des.*, Vol. 16, pp. 174-183. https://doi.org/10.1007/s00163-006-0012-8

Dörnyei, Z. (2007), *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*, Oxford University Press, Oxford.

Eisenhardt, K.M. (1989), "Building Theories from Case Study Research", *Acad. Manage. Rev.*, Vol. 14, pp. 532-550. https://doi.org/10.2307/258557

Endsley, M.R. (1987), "The Application of Human Factors to the Development of Expert Systems for Advanced Cockpits", *Proc. Hum. Factors Soc. Annu. Meet.*, Vol. 31, pp. 1388-1392. https://doi.org/10.1177/154193128703101219

Gold, C. et al. (2013), "'Take over!' How long does it take to get the driver back into the loop? Proc. Hum. Factors Ergon", *Soc. Annu. Meet.*, Vol. 57, pp. 1938-1942. https://doi.org/10.1177/1541931213571433

IMO (2018), IMO takes first steps to address autonomous ships [WWW Document]. URL http://www.imo.org/en/MediaCentre/PressBriefings/Pages/08-MSC-99-MASS-scoping.aspx (accessed 11.24.19).

Kaber, D.B. and Endsley, M.R. (1997), "Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety", *Process Saf. Prog.*, Vol. 16, pp. 126-131. https://doi.org/10.1002/prs.680160304

Levander, O. (2017), "Autonomous ships on the high seas", *IEEE Spectr.*, Vol. 54, pp. 26-31, https://doi.org/10.1109/MSPEC.2017.7833502

Lloyd's Register (2016), Cyber safe for marine [WWW Document]. Lloyds Regist. URL https://www.lr.org/en/cyber-safe-for-marine/ (accessed 11.22.19).

Lutzhoft, M. et al. (2019), "Human-centred maritime autonomy - An ethnography of the future", *J. Phys. Conf. Ser.*, Vol. 1357, p. 012032. https://doi.org/10.1088/1742-6596/1357/1/012032

Lützhöft, M.H. and Dekker, S.W.A. (2002), "On Your Watch: Automation on the Bridge", *J. Navig.*, Vol. 55, pp. 83-96. https://doi.org/10.1017/S0373463301001588

Mallam, S.C., Nazir, S. and Sharma, A. (2019), "The human element in future Maritime Operations–perceived impact of autonomous shipping", *Ergonomics*, pp. 1-12.

Maritime UK (2018), Maritime Autonomous Surface Ships - UK Code of Practice | Maritime UK [WWW Document]. URL: https://www.maritimeuk.org/media-centre/publications/maritime-autonomous-surface-ships-uk-code-practice/ (accessed 11.25.19).

Moreno, D.P. et al. (2015), "A step beyond to overcome design fixation: a design-by-analogy approach", In: *Design Computing and Cognition '14*, Springer. pp. 607-624.

MUNIN (2015), D8.8: Final Report: Shore Control Centre – FP7 GA-No 314286 | All MUNIN Deliverables [WWW Document]. URL: http://www.unmanned-ship.org/munin/news-information/downloads-information-material/munin-deliverables/ (accessed 11.25.19).

Musić, S. and Hirche, S. (2017), "Control sharing in human-robot team interaction", *Annu. Rev. Control*, Vol. 44, pp. 342-354. https://doi.org/10.1016/j.arcontrol.2017.09.017

Ramos, M. et al. (2019), "Human-system concurrent task analysis for maritime autonomous surface ship operation and safety", *Reliab. Eng. Syst. Saf.*, p. 106697.

Rødseth, ØJ and Nordahl, H. (2017), "Definitions for autonomous merchant ships", *Presented at the Norwegian Forum for Unmanned Ships, Version*, pp. 2017-10.

Rødseth, ØJ, Nordahl, H. and Hoem, Å. (2018), Characterization of Autonomy in Merchant Ships, in: 2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO). Presented at *the 2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, pp. 1-7. https://doi.org/10.1109/OCEANSKOBE.2018.8559061

Sheridan, T.B. and Verplank, W.L. (1978), *Human and computer control of undersea teleoperators*, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

Wilhelmsen. (2018), Wilhelmsen and KONGSBERG establish world's first autonomous shipping company [WWW Document]. Wilhelmsen. URL: https://www.wilhelmsen.com/media-news-and-events/press-releases/2018/wilhelmsen-and-kongsberg-establish-worlds-first-autonomous-shipping-company/ (accessed 11.22.19).

Yin, R.K. (2017), *Case study research and applications: Design and methods*, Sage publications.

Appendix C9: Academic contribution 9

Dybvik, H., & Steinert, M. (Under review) Operationalized hypotheses build bridges between qualitative and quantitative design research. Journal of Mixed Methods Research.

This paper is awaiting publication and is not included in NTNU Open

C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16

Appendix C10: Academic contribution 10

Hatlem, L. A., Chen, J., Dybvik, H., & Steinert, M. (2020). A Modular Research Platform – Proof-of-Concept of a Flexible Experiment Setup Developed for Rapid Testing of Simulators, UIs and Human Physiology Sensors. Procedia CIRP, 91, 407–414. https://doi.org/10.1016/j.procir.2020.02.193

Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 91 (2020) 407–414

www.elsevier.com/locate/procedia

30th CIRP Design 2020 (CIRP Design 2020)

A Modular Research Platform – Proof-of-Concept of a Flexible Experiment Setup Developed for Rapid Testing of Simulators, UIs and Human Physiology Sensors

Leif Arne Hatlem[a], John Chen[a], Henrikke Dybvik[a]*, Martin Steinert[a]

[a]The Norwegian University of Science and Technology (NTNU), Richard Birkelands vei 2B, 7491 Trondheim, Norway

* Corresponding author. Tel.: +47 95 41 72 45. *E-mail address:* henrikke.dybvik@ntnu.no

Abstract

This work presents a modular research platform to design, test and run human-machine interaction (HMI) experiments. Traditionally, HMI experiments are time and resource consuming, particularly in the piloting phase. Furthermore, such experiment setups are often rigid and only fit to one particular hypothesis. Thus, significant time is needed to alter the setup to new hypotheses, if this is possible at all. The platform presented is a technical proof-of-concept of a highly flexible experiment setup, which can rapidly be adapted to alternative hypotheses. Examples of interchangeable modules include simulator software (context), user interface (independent variable) and human operator physiology sensors (dependent variable). An agile product development methodology, Wayfairing, was used to accomplish this.

Keywords: human-machine interaction, experiments, modularity, research platform, flexibility, UI, physiology sensors

1. Introduction

This paper presents development of a modular research platform intended for designing, testing and running human-machine interaction (HMI) experiments. The platform presented is a technical proof-of-concept of a highly flexible experiment setup. Examples of interchangeable modules include simulator software (context), user interface (independent variable) and human operator physiology sensors (dependent variable). The aim of this prototype is to be research-ready for HMI experiments, which means the researcher can rapidly investigate different hypotheses in a piloting phase [1] by changing abovementioned modules, prior to a full experimental run testing one or more selected hypotheses statistically.

Experiments and interaction studies are equipment intensive, time-consuming and labor intensive [2]. Availability of research platform, high cost and proprietary costs are additional challenges faced by academic researchers [3]. Furthermore, HMI experiments rarely run correct the first time [4], and although preparing all experimental parameters well in advance is ideal, this is not the case in practice. For novel interaction techniques in early stages of product development constructing a well-defined hypotheses if often very hard [5]. For example, multiple years were required to develop a humanoid research platform investigating human interaction [3]. Experimental setups like the car simulation of Ahn et al. [6] and autonomous car simulation of Gil et al. [7] are likely to be quicker to set up and pilot. However, they both feature a simple monitor, steering wheel and chair, thus lacking ecological validity. The lack of ecological validity is problematic as experimental findings are not generalizable to real-life settings.

HMI experiments such as the abovementioned investigate the human, how they interact, how they operate, how to and in what way include to the human user in the loop, etc. In such

10.1016/j.procir.2020.02.193

experiment setups there is an increasing interest in measuring the human's physiological response as a dependent variable. Examples of using physiology sensors in HMI research include assessing drivers' cognitive engagement under varying levels of automation in a driving simulator using fNIRS [8], investigating the effect of mental fatigue caused by sleep deprivation in driving using simultaneous fNIRS, EEG, and ECG [6], and investigating operators' mental state in ship navigation using ECG and GSR [9], [10]. When including physiology sensors in HMI research additional challenges, such as hardware integration, arise [11].

The fundamental nature of a classical experiment is to vary one or more independent variables, observing potential changes in one or more dependent variables [12]. HMI experiments has since the beginning embraced cognitive science and used psychology style experiments as a basis for usability testing [5]. Different designs or solutions are compared through controlled studies, often including baseline and comparisons tests in a within-subjects design (in which all subjects test all designs), measuring a range of variables [13]. Afterwards, testing for statistical differences in these variables are custom as there are no absolute values to compare with. As mentioned, the piloting phase is important in HMI since it increase the chance of experiment success [1], although it produces additional demands in terms of time, labor and cost simultaneously. Thus, time to develop an experiment and hypotheses can be greatly reduced by the possibility to quickly test different independent and dependent variables.

With these notions in mind the authors wanted to develop a modular research platform allowing for rapid changes of context, independent and dependent variables during piloting. The core feature would be the ability to easily and rapidly test different interface configurations, simulators and physiology sensors. The experiment setup should be flexible and produce an experiment with high ecological validity, reliability and reproducibility.

Thus, this paper presents a modular research platform. It describes design and development of the flexible experiment platform and demonstrate it by a proof-of concept experimental user test. This paper focuses on the physical aspect of the modular research platform, notably providing a prototype that is technically robust. A dedicated software was also developed as a part of the system, which is presented in brief, but an exhaustive description is not within the scope of the paper.

The paper is structured as follows: section 2 describes the method. Section 3 briefly describes the development process and proof-of-concept user test. Section 4 presents the modular research platform and describes how it can be adapted to new hypotheses. Section 5 is discussion and section 6 conclusion.

2. Method

Wayfaring is a flexible, physical product development methodology for early stage product development, which utilizes principles from agile software development [14], [15], [16]. Problem understanding is developed to such a degree that good concept choices and appropriate requirements can be made in an early phase thus preventing costly loop-backs later in the process. A special focus is placed on interactions between disciplines, leveraging diversity in teams to promote iterative learning cycles through rapid conceptual prototyping. Exploration is conducted through a journey of idea-probes, each probe consisting of designing, building and testing prototypes, aiming towards a vision of a problem solution. Critical functions of components should be tested in isolation. System integration occurs when these functions are fulfilled, to validate the continued fulfilment in the system context [14]. Integrating different disciplines can reveal inter-dependencies among disciplines, thus design changes in one domain can cause requirement adaptation in another domain. Unknown unknowns are uncovered early, while flexibility is high and cost of change low.

Wayfaring has been introduced as a development tool for human experiments in interaction design and engineering design science [17]. It is applicable in the early and ambiguous conceptualization and design of experiments, as well as cases where no obvious experiment precedes it and it must be built from scratch. Four main principles are particularly advocated: probing ideas, merging multidisciplinary, and retaining high speed and agility.

Prototypes are often used when developing products in engineering design and are important in fuzzy-front-end projects where wayfaring has been utilized, especially when developing products with a physical dimension [16], [18] [14], [17]. Prototypes are purposefully formed manifestations of design ideas built to traverse a design space. Such prototyping activity can create valuable knowledge of the final design [18]. In wayfaring, each prototype is built to test a specific idea and/or a system interaction [16].

Affective Engineering uses physiology sensors to capture and incorporate the human emotional dimension as a part of evaluating and identifying the better design of an interface, a process or product [19], [20]. Examples of physiology sensors are ECG, GSR, fMRI, fNIRS, EEG, PPG, EMG, pupil tracing devices, etc., [20].

3. Development

This section highlights certain aspects of the development process. It describes development of the physical infrastructure and user interfaces (independent variables), before pointing to rapid and relatively large changes between pilots. A description of a proof-of-concept experimental run follows.

3.1. Software development

A software, TrollSim, was developed using agile methodologies. Software requirements were mostly driven by needs and integration testing from the physical domain.

3.2. Developing experiment infrastructure

A small room was constructed using a timber-frame and MDF-sheets. This provided a stiff frame which internal structures could be attached to. Floor, walls and roof were easy to adapt and continue building upon. Flexibility for further development was continuously considered [16]. The internal infrastructure was designed and built to hold a classical experiment setup, aiming at a static, controlled physical environment. The simulator room was divided in two by a black curtain, with the majority dedicated to the participant. LED-lights were installed in the ceiling, as well as a curtain, limiting visual disturbances from the experimenter area and outside the box.

User interfaces was explored by developing Arduino based alarm systems and operator controllers. Initial development isolated the two systems until critical functions were stable, before they were integrated in one device. The alarm system first employed the modalities sound and smell, the first intended as a baseline and the second as a high novelty alternative. The sound alarm utilized a simple buzzer to play monotones and a 3D printed button to register reaction time. The container was laser-cut. The smell alarm utilized a servo motor to rotate a disk holding three different laser-cut cartridges containing variations of air fresheners (Little Trees/Wunderbaum), and a fan directing scent to the participant's nose.

A two-piece joystick served as a baseline controller, which we sought to test along with a high-novelty alternative. This drove the development of a sensor glove, an Arduino-based controller with a finger-actuated flex sensor controlling throttle and a gyroscope allowing hand movement to control pitch, roll and yaw.

3.3. Experiment piloting

The final experiment procedure (Fig. 4) was refined through piloting to a high level of detail. The piloting phase can be characterized by relatively large changes and rapid learning between pilots. This work had a development time of two months and was conducted by two master students.

The experiment was piloted eight times. From pilot 1 to 2 questionnaires were created and implemented in the procedure. From pilot 2 to 3, the setup was changed from three factors to two factors by removing controllers as an independent variable. The alarm system was redesigned, changing from sound and smell modalities to combinations of sound, light, and haptic feedback modality. A physical alarm control panel was added. Between 3 and 4 the method of achieving low and high workload intensity was redesigned from using different pre-programmed flight school tutorials in X-Plane 11 [21], to the same one coupled with secondary tasks. This was controlled by a newly created audio control panel. The alarm system parameters were also tweaked to better approximate reality. Between 5 and 6, all task training was redesigned from manual instruction to automated and a second screen was added to the participant infrastructure. This marked a point where the experiment ran fairly smooth, except for minor bugs in different modules. Between 7 and 8, preliminary data analysis of questionnaire answers was implemented in Google Sheets.

3.4. Proof-of-concept user test

This section describes the proof-of concept experimental run conducted after piloting, using the proposed setup as described in section 4. An external researcher conducted the experiment. They received instructions, which included a walkthrough of the procedure, a procedure checklist and a manuscript. The researcher had some experience with running HMI-experiments with physiology sensors, having previously completed one study. This researcher conducted one pilot with one of the developers (pilot nr 5), and two user tests after pilot 8. There were minor issues with the first test, whereas the second ran had no issues and marked the proof-of-concept experimental run. In total 10 runs ensured technical robustness and repeatability.

4. Proposed system – A descriptive of the modular research platform

This section describes a proposed experiment setup. This is the setup used during the proof-of-concept user test. Then, we give examples of changes that are easy to implement in the modular research platform to build a new experiment and investigate new hypotheses.

4.1. Software

A dedicated software, TrollSim, was developed. In general, TrollSim can collect data from a multitude of data sources, manipulate the data, log it and transmit it. TrollSim can gather, synchronize and log physiology data with data from scenario events, controller input and the simulator. A fronted allows the researcher to influence data through the UI, and a live plot with logdata visualization allows for a quick overview over a finished experiment. The software is intended to function as an API for its users, designed with the aim of exposing the user to minimal amount of dataflow logic, focusing on actual business logic. Further description of TrollSim is not within the scope of this paper.

4.2. Physical infrastructure

4.2.1. Proposed physical infrastructure

The following setup is designed around one experimenter, but can also be used by two experimenters. The physical infrastructure is shown in Fig. 1. It consists of (1) monitors , (2) laptop controlling TrollSim, (3) keyboard and mouse for main computer, (4) control panel for alarm device, (5) audio control panel for triggering secondary tasks, (6) experimenter check list, and (7) boxes for forms used in experiment (prepared for use are on the right, used and collected are on the left).

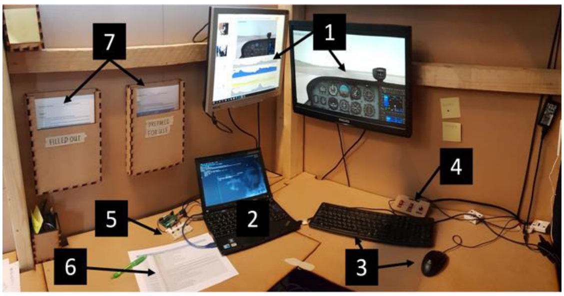


Fig. 1. Infrastructure for the experimenter

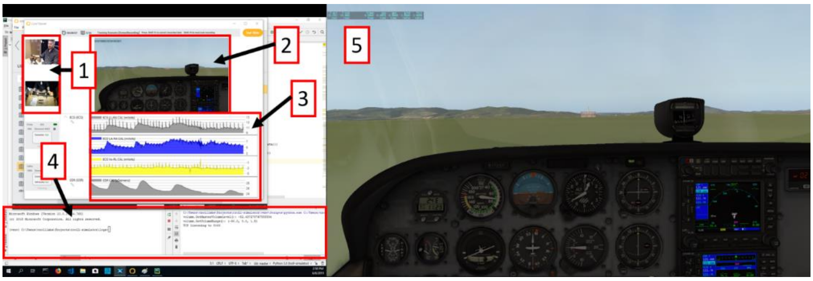


Fig. 2. Close-up of the experimenter screen setup

Fig. 2. shows a close-up of the experimenters' screen setup, which contains (1) the camera feeds capturing the participants' face and desk, (2) a video feed of the participants main screen, (3) live plots of physiology sensors, (4) pyCharm, showing which audio tracks are triggered and played for the participant and (5) participant screen, used to monitor flying tasks as well as initiate X-Plane scenarios.

The participant is separated from the experimenter by a curtain. They enter thought the rearmost part, mimicking entering a cockpit. Participant interface is shown in Fig. 3. and consists of (1) 2 web cameras capturing participant face and desk, (2) 2 screens, instrument zoom (left), main screen (right) (3) ear protection w/auxiliary input from PC for playback of audio alarms and secondary task instructions, (4) joystick: throttle (left), main stick (right), (5) keyboard (FN-buttons and numpad restricted), (6) pink note with participant number and experiment group, (7) alarm device, (8) secondary task sheet, (9) mouse, and (10) printout with names of cockpit instruments.

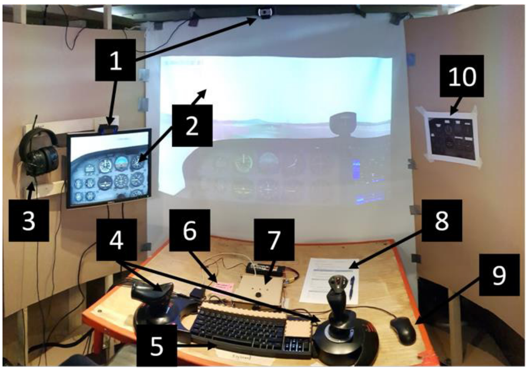


Fig. 3. Participant interface

4.2.2. Changing the physical infrastructure

The wooden structure allows for fastening objects anywhere in the experiment room. All physical objects can be moved to accommodate re-configurations. The main screen, a single, short-throw projector can for example be swapped for one or more monitors, different backdrop and projector. It could be interesting to use different monitor types as an independent variable, observing what effect different configurations impose on the participants. Arduino based devices or other devices can be swapped or added by simply connecting USB cables.

4.3. Procedure

4.3.1. Proposed procedure

Initial instructions were given by the experimenter after having greeted the participant, collecting consent form and attaching physiology sensors. Afterwards, instructions were automated to minimize variation. Exceptions were briefing and debriefing. Automatic instructions were made by combining iMotions [22] and Google Slides. iMotions also read physiology measurements and recorded video and audio.

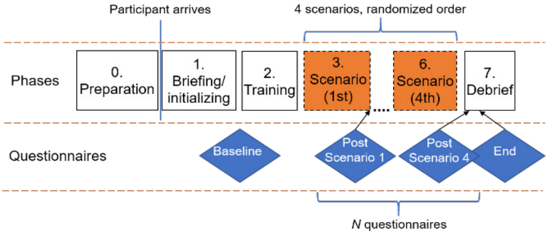


Fig. 4. Experiment procedure.

Primary task training was conducted through pre-programmed flight school tutorials in X-Plane 11 [21]. A Google Slides presentation provided secondary task training, allowing the participant to navigate back and forth at their own pace.

The primary task was operating a Cessna 172 airplane, the participants were tasked with completing a "Traffic Pattern" tutorial. Four different scenarios were created by combining two independent variables, alarm system and workload intensity:

- Traffic pattern. High-workload. Alarm system A.
- Traffic pattern. Low-workload. Alarm system B.
- Traffic pattern. High-workload. Alarm system B.
- Traffic pattern. Low-workload. Alarm system A.

To avoid learning effects, the order was randomized.

4.3.2. Changing procedure

The experiment procedure can be adapted to new setups as described above by altering the number and order of scenarios. Experiment instructions are easy to modify. The combination of iMotions with Google Slides in training allowed the participant to navigate selected instructions at their own pace, while the experimenter retains overall control.

4.4. Independent variables

4.4.1. Proposed independent variables

A two factor, two-level experiment design was achieved with the two independent variables: alarm system and workload intensity. Each has two levels: alarm system A and B, and high and low workload.

The alarm device triggers two different alarm systems with different combinations of the following modalities; visual (blinking LEDs), auditory (alarming sounds) and tactile (haptic force feedback). Alarm system A consist of light and sound. Alarm system B consists of light, sound and haptic vibration in the right-hand controller. Three levels of danger (high, moderate, low) approximated airplanes centralized warning system to achieve ecological validity. Alarms are triggered at random points in time, a pre-defined number of times. The participant is tasked with pressing a button when they register an alarm going off. Both the alarm going off and the button being pressed is logged, measuring response time.

To alter the perceived workload intensity under different flight scenarios, secondary tasks were implemented in the high workload scenario. These tasks were pre-recorded questions, triggered by the experimenter using an audio control panel at pre-defined times during the scenario. The participant would answer the question using information from X-Plane in writing on a dedicated secondary task sheet. To ensure ecologically validity the questions were developed in collaboration with a pilot. Three tasks resulted: estimation of remaining flight time based on remaining fuel, transponder code changing, and reading and logging of time and heading. The low workload scenario had no secondary tasks.

4.4.2. Changing independent variables

Interchangeability of peripheral programs such as X-Plane 11 and iMotions is a feature as they merely are examples of what TrollSim can communicate with. Switching X-Plane for a different simulator software, for example Ship Simulator Extremes [23] is relatively simple. One aspect to keep in mind, X-Plane provides an API for developers to communicate with and send commands through, which allows for e.g. reading and overwriting internal states. If the alternative software doesn't have an API these features are lost. Ship Simulator Extremes does not have an API, therefore reading or writing internal states is not possible, neither is sending commands. If data for controllers are still an important aspect, there are two possibilities here. Ship Simulator accepts input from keyboard, which we can emulate in two ways. The first is with physical microcontrollers such as Arduino. The second alternative is to create software running on the host side sending keyboard signals to the host OS. This shows that we can still create custom controllers, similar to the sensor glove prototype, where we get full access to data emitted. Full access to controller data means that we can log it as we would with X-Plane.

Sub-variables of the current alarm system are easy to change, e.g. the number of alarms, volume, alarm sound or haptic intensity.

A custom Arduino-based controller was developed (the sensor glove), but not included in the final experiment due to reducing from 3 to 2 independent variables. However, this means it is trivial to add custom controllers, as it is implemented in the same manner as the alarm device. Arduino devices are logged in TrollSim and the same protocols and dataref-control can be used for any Arduino-based controller. Possibilities for Aruduino based devices are only limited by available sensors and imagination.

Scenario nature can be altered by reducing or increasing the number, selecting different pre-programmed tutorials or flights, or custom-make flights in the software. Using pre-programmed tutorials in itself offers very little flexibility, aside from selecting an appropriate tutorial. Building highly customized scenarios in X-Plane is an option, which requires more time, but incudes many options and more freedom. Options include triggering hundreds of failure modes such as engine failure, rudder failure, fuel leakage etc., creating custom weather and plan routes for the participant to fly.

Scenario intensity can be changed by altering workload. Workload can be increased by triggering secondary tasks at more stress-full points in time, such as during a turn or landing. The three tasks could be made harder by increasing the arithmetic difficulty, creating a larger table to look up values in, and requiring the participant to log additional instrument readings. Furthermore, introducing more stimuli, such as white noise in headphones, makes it harder to interpret sounds and voice commands. Removing the second screen requires the participant to more actively zoom in/out to read instruments. For decreased scenario intensity analogue adaptations to the stimuli can be made. Generally, cognitive workload can be adjusted by adjusting primary and secondary tasks in combination with other stimuli.

4.5. Dependent variables

4.5.1. Proposed dependent variables

Two types of physiology data were collected, electrocardiography (ECG) and galvanic skin response (GSR). ECG measures the electrical potential difference across the heart, which can be translated to heart rate and heart rate variability (HRV) among other measures. Physical and mental states can be interpreted from these measures. The Shimmer3 ECG unit [24] was used with five electrodes and a sampling rate of 512 Hz. Galvanic skin response (GSR) measures skin conductance, which increase with physical activity and/or emotional arousal/alertness [20]. To detect GSR changes due to emotional stimuli the experiment must be conducted under very controlled physical conditions [25], which the setup exhibits. The Shimmer3 GSR+ Unit [26] was used with two electrodes and 128 Hz sampling rate.

Performance is measured by alarm response time, answers to secondary tasks, and flight performance data which is generated by X-Plane 11's 'report card'.

Subjective measures were collected by questionnaires in Google Forms. They included The Affect Grid [27], level of stress, NASA-RTLX [28], Overall Workload [29] and scenario specific questions, such as alarm preference.

4.5.2. Changing dependent variables

Alternative physiological sensors are possible through use of code wrappers and TrollSim. One example are open source physiology sensors in Arduino.

The current alarm system has one pushbutton recording the participants response time. To measure e.g. alarm recognition instead of reaction time is simple. Three additional buttons corresponding to the three different alarms can be added. This requires rewiring a few electronic circuits, slightly altering an Arduino code by adding variables, 3D printing extra buttons and a simple laser cut (this was done in an earlier prototype). As explained, any Arduino based sensor or device integrates with the setup.

Constructs other than stress, affective state and workload, can be teste by changing the questionnaires accordingly.

4.6. Example hypotheses

The proposed setup enable investigation of e.g. the following hypothesis:

- Reaction time decrease with alarm system B, both during high and low workload scenario.

By changing one or more of the independent variables, dependent variables or experiment procedure a new hypothesis can be investigated. One example of an alternative setup is as follows. A custom flight is made in while several accidents or near accidents occur, which trigger several different types of failure modes. The operator's action in response to the failure can be categorized as correct, intermediate and wrong reaction (based on domain knowledge). Physiology sensors are changed and EEG/fNIRS are used to measure e.g. cognitive workload. The alarm systems stay the same. This new setup enables investigation of e.g. the following hypotheses:

- Alarm system B significantly increase the number of correct failure reactions, and
- significantly decrease reaction time, and
- the operator's cognitive workload significantly decreases.

Here, the time to alter the setup to new hypotheses is significantly less than in traditional HMI experiments setups. Due to our platforms focus onto flexibility, researchers can thus rapidly iterate and test example alternative senarios, interface configurations or sensors which in turn enables them to rapidly investigate multiple new hypotheses.

5. Discussion

5.1. System development

Free software was prioritized to minimize system cost and to reduce entry barriers to testing the system. Using python and Arduino, both open source gave maximum flexibility. iMotions [22], a paid closed source software was used for experiment execution and initial sensor connection due to not finding an equivalent open source software. G-Suite products (Forms, Sheets, Slides) was used in combination with iMotions, which countered lacking flexibility in iMotions and provided an integral experiment execution. This illustrates that many issues can be solved without specifically written software, but instead hacking together existing solutions, provided integration does not lead to a greater total challenge. By basing many of the modules where flexibility was particularly important on inherently flexible, open source products and platforms we argue our system has inherited much of the same flexibility.

Throughout the development process, the system architecture considered TrollSim as a central hub, which was built to effectively handle data from a broad range of sources. Modularity was achieved by creating a standard code wrapper. This allowed development of custom modules without the need to extensively consider how they would affect overall system architecture and how it should be integrated. The Arduino based alarm device is one example of such a custom module. It was prototyped and tested on a module level first, until it functioned as intended by itself. The code wrapper was then added, identification defined and integrated in the code. We believe this strategy allowed for more testing since each prototype was built to test only one specific idea and/or system interaction, removing anything superfluous, which is consistent with Leifer and Steinert [16]. This simplified module-code made both debugging and modifications of code easier at the early stages of development. Furthermore, this modularity allowed a non-developer (mechanical engineer) to develop custom modules with relatively little code, which was wrapped and then functioned with the overall system. Through all pilot studies and two subsequent user tests, we have had no crashes or lost logs in TrollSim, which we argue is a showcase of its robustness.

The strategy of adding on modules transferred to the physical domain. The alarm control panel and the audio control panel are both examples of parts of the experiment system added when the need emerged, without requiring any noteworthy environment adaptation for system integration.

Throughout the development process, high flexibility was the main consideration driving development choices. This is illustrated by the choice of a modular system structure, versatile logging functions in TrollSim, and Arduino. Arduino is an opensource hardware platform for mechatronics development. Having developed and integrated one such device made all subsequent integration of Arduino based devices trivial. The choice to not solder any connections when prototyping mechatronics greatly sped up and simplified all iterations of these subsystems. Except for dynamically stressed connections, only breadboard or screw connectors were used.

Flexibility during development was further aided by how the physical infrastructure was constructed. The timber and MDF construction created a room in which all walls, floors and the roof could be modified and built upon. This gave great design freedom. A wall-to-wall scrum board inside the room and strategically placed storage for equipment contributed to adding many minor features, which overall contributed substantially to the end result. There were no penalties and low barriers to prototyping different setups and configurations, consistent with findings on key factors of spaces that facilitate change [16].

5.2. System cost

The system is developed based on the lowest resolution possible (but not lower) in every area, keeping overall cost low. Overall costs can be split in two. A one-time cost of 4200 € for developing the system, and a yearly subscription cost of 2077 € for iMotions. Approximately 50% of the total one-time costs are general hardware a university or research institution might already have available, such as main computer, monitors, keyboards, mouse, cables etc. That aside the main costs are the projector, the Shimmer sensors and iMotions software. These were available and chosen over cheaper alternatives to secure medical-grade equipment and reliable data capture. If requirements were different and iMotions was not required it would be possible to replace the projector with a monitor at < 50% of the cost and exchange the Shimmer sensors for Arduino ECG and GSR sensors priced below 50 €. However, we assume the quality of these are different. The sum of these changes would reduce the total one-time cost of the setup by approximately 1/3 and remove the subscription cost.

It should be noted that iMotions has an initial cost, dependent on your setup and their proposal.

5.3. Limitations

The flexible experiment setup is a proof-of-concept of technical robustness. A full experimental run is out of the scope of this paper. Pilot experiments were conducted with student participants. The current TrollSim is a proof-of-concept and has not been tested extensively.

6. Conclusion

This paper presents a modular research platform to design, test and run HMI experiments. Traditional HMI experiments are time and resource consuming, often rigid and only fit to one particular hypotheses. Thus, significant time is needed to alter the setup to new hypotheses, if this is possible at all. The platform presented is a proof-of-concept of a highly flexible experiment setup, which can rapidly be adapted to alternative hypotheses by simply changing one or more independent and dependent variables. Examples of interchangeable modules include simulator software (context), user interface and human operator physiology sensors. The time to alter this setup to new hypotheses is significantly less than traditional HMI experiments. Wayfairing, an agile product development methodology, was used to accomplish this.

Acknowledgements

This research is supported by strategic funds from the department for Mechanical and Industrial Engineering (MTP) at NTNU.

References

[1] E. R. Van Teijlingen and V. Hundley, "The importance of pilot studies," 2001.

[2] K. Dautenhahn, "Methodology & Themes of Human-Robot Interaction: A Growing Research Field," *International Journal of Advanced Robotic Systems*, vol. 4, no. 1, p. 15, Mar. 2007, doi: 10.5772/5702.

[3] K. Nishiwaki, J. Kuffner, S. Kagami, M. Inaba, and H. Inoue, "The Experimental Humanoid Robot H7: A Research Platform for Autonomous Behaviour," *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1850, pp. 79–107, 2007.

[4] H. C. Purchase, *Experimental human-computer interaction: a practical guide with visual examples*. Cambridge University Press, 2012.

[5] J. Lazar, J. H. Feng, and H. Hochheiser, *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

[6] S. Ahn, T. Nguyen, H. Jang, J. G. Kim, and S. C. Jun, "Exploring Neuro-Physiological Correlates of Drivers' Mental Fatigue Caused by Sleep Deprivation Using Simultaneous EEG, ECG, and fNIRS Data," *Front. Hum. Neurosci.*, vol. 10, 2016, doi: 10.3389/fnhum.2016.00219.

[7] M. Gil, M. Albert, J. Fons, and V. Pelechano, "Designing human-in-the-loop autonomous Cyber-Physical Systems," *International Journal of Human-Computer Studies*, vol. 130, pp. 21–39, Oct. 2019, doi: 10.1016/j.ijhcs.2019.04.006.

[8] S. Sibi, S. Balters, B. K. Mok, M. Steinert, and W. Ju, "Assessing driver cortical activity under varying levels of automation with functional near infrared spectroscopy," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 1509–1516, doi: 10.1109/IVS.2017.7995923.

[9] A. S. Wulvik, H. Dybvik, and M. Steinert, "Investigating the relationship between mental state (workload and affect) and physiology in a control room setting (ship bridge simulator)," *Cogn Tech Work*, Apr. 2019, doi: 10.1007/s10111-019-00553-8.

[10] H. Dybvik, A. Wulvik, and M. Steinert, "STEERING A SHIP - INVESTIGATING AFFECTIVE STATE AND WORKLOAD IN SHIP SIMULATIONS," presented at the 15th International Design Conference, 2018, pp. 2003–2014, doi: https://doi.org/10.21278/idc.2018.0459.

[11] A. Sears and J. A. Jacko, *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. CRC press, 2007.

[12] L. T. Blessing and A. Chakrabarti, *DRM, a design research methodology*. Springer Science & Business Media, 2009.

[13] J. A. Jacko, *The human-computer interaction handbook : fundamentals, evolving technologies, and emerging applications*, 3rd ed. Boca Raton, Fla: CRC Press, 2012.

[14] A. Gerstenberg, H. Sjöman, T. Reime, P. Abrahamsson, and M. Steinert, "A Simultaneous, Multidisciplinary Development and Design Journey – Reflections on Prototyping," in *Entertainment Computing - ICEC 2015*, 2015, pp. 409–416, doi: 10.1007/978-3-319-24589-8_33.

[15] M. Steinert and L. J. Leifer, "'Finding One's Way': Re-Discovering a Hunter-Gatherer Model based on Wayfaring," *International Journal of Engineering Education*, vol. 28, no. 2, p. 251, 2012.

[16] L. J. Leifer and M. Steinert, "Dancing with ambiguity: Causality behavior, design thinking, and triple-loop-learning," *Information Knowledge Systems Management*, vol. 10, no. 1–4, pp. 151–173, Jan. 2011, doi: 10.3233/IKS-2012-0191.

[17] K. K. Leikanger, S. Balters, M. Steinert, and others, "Introducing the wayfaring approach for the development of human experiments in interaction design and engineering design science," in *DS 84: Proceedings of the DESIGN 2016 14th International Design Conference*, 2016, pp. 1751–1762.

[18] Y.-K. Lim, E. Stolterman, and J. Tenenberg, "The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 15, no. 2, p. 7, 2008.

[19] S. Balters and M. Steinert, "Decision-making in engineering-a call for affective engineering dimensions in applied engineering design and design sciences," presented at the Innovative Design and Manufacturing (ICIDM), Proceedings of the 2014 International Conference on, 2014, pp. 11–15.

[20] S. Balters and M. Steinert, "Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices," *Journal of Intelligent Manufacturing*, pp. 1–23, 2015.

[21] *X-Plane 11 Flight Simulator*. 2017.

[22] "iMotions: Biometric Research, Simplified," *iMotions*, 12-Jul-2017. [Online]. Available: https://imotions.com/. [Accessed: 07-Dec-2017].

[23] *Ship Simulator Extremes*. 2010.

[24] *Shimmer3 ECG/EMG Unit*. Dublin, Ireland: Shimmersense, 2017.

[25] M. van Dooren, J. J. G. (Gert-J. de Vries, and J. H. Janssen, "Emotional sweating across the body: Comparing 16 different skin conductance measurement locations," *Physiology & Behavior*, vol. 106, no. 2, pp. 298–304, May 2012, doi: 10.1016/j.physbeh.2012.01.020.

[26] *Shimmer3 GSR+ Unit*. Dublin, Ireland: Shimmersense, 2017.

[27] J. A. Russel, A. Weiss, and G. A. Mendelsohn, "Affect grid: A single-item scale of pleasure and arousal," *Journal of Personality and Social Psychology*, vol. 57, no. 3, pp. 493–502, 1989.

[28] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," presented at the Proceedings of the human factors and ergonomics society annual meeting, 2006, vol. 50, pp. 904–908.

[29] M. A. Vidulich and P. S. Tsang, "Absolute Magnitude Estimation and Relative Judgement Approaches to Subjective Workload Assessment," *Proceedings of the Human Factors Society Annual Meeting*, vol. 31, no. 9, pp. 1057–1061, Sep. 1987, doi: 10.1177/154193128703100930.

C1

C2

C3

C4

C5

C6

C7

C8

C9

C10

C11

C12

C13

C14

C15

C16

Appendix C11: Academic contribution 11

Erichsen, C. K., Dybvik, H., & Steinert, M. (2020). Integration of low-cost, dry-comb EEG-electrodes with a standard electrode cap for multimodal signal acquisition during human experiments. DS 101: Proceedings of NordDesign 2020, Lyngby, Denmark, 12th - 14th August 2020, 1–12. https://doi.org/10.35199/NORDDESIGN2020.19

NordDesign 2020
August 11-14, 2020
Kgs. Lyngby, Denmark

Integration of low-cost, dry-comb EEG-electrodes with a standard electrode cap for multimodal signal acquisition during human experiments

Christian Kuster Erichsen[1], Henrikke Dybvik[2], Martin Steinert[3]

[1]The Norwegian University of Science and Technology (NTNU), Norway
cnerich@stud.ntnu.no
[2]The Norwegian University of Science and Technology (NTNU), Norway
henrikke.dybvik@ntnu.no
[3]The Norwegian University of Science and Technology (NTNU), Norway
martin.steinert@ntnu.no

Abstract

This paper describes the development of a convenient and fast integration of low-cost, spring-loaded, dry electroencephalography (EEG) electrodes with a research-grade sensor cap ensuring electrode positioning according to the five percent system. Measuring brain activity is of increasing interest in fields outside of psychology and neuroscience, such as engineering. Human error often occurs due to lapses of attention, an inability to fully understand consequences or inadequacies in interface design. Effective design solutions incorporating and recognizing human behavior and various types of responses are needed to mitigate human error. Physiology sensors can be used to better evaluate which design meets user needs in the best way. Brain activity sensors have been applied within the brain-computer interface (BCI) community for decades. EEG is a highly popular modality due to its non-invasive nature and high temporal resolution. Prior research demonstrates increased performance of experimental results when using multimodal measurements compared to unimodal in prediction and classification tasks. Thus, we wanted to integrate EEG with an existing experimental setup which included functional Near-Infrared Spectroscopy (fNIRS). An integration was developed by means of rapid prototypes in cycles of design-build-test. The proposed setup increases available electrode positions, compared to currently available low-cost equipment, and constitutes a pragmatic, low-cost approach for integrating EEG-measurements in combination with other brain activity sensors, such as fNIRS. A proof-of-concept test of the signal quality was performed by means of two tasks shown to reveal easily detectable changes in the EEG-signal: closing the eyes and eye-blinking. Closing the eyes gave an increase in peak amplitude in the alpha range, an effect that was reversed once the eyes were opened. Deliberately blinking the eyes in specific intervals produced characteristic electrooculographic (EOG) artifacts in the signal. Both responses agree with the literature. The proposed solution aims to lower the barrier to including EEG as an additional modality in existing experimental setups, and thus increase the performance of experimental results.

Keywords: EEG, fNIRS, prototyping, human-centered design, experimentation

1 Introduction

1.1 Background and motivation

1.1.1 Affective engineering and its importance to interface design

Measuring brain activity is becoming of increasing relevance in fields outside of psychology and neuroscience, such as engineering. The use of sensors that measure brain activity, especially electroencephalography (EEG), has been widely applied within the brain-computer interface (BCI) community for decades, see for instance Vidal (1973). The BCI community concerns itself with technology that enables control of computers and machines through commands generated by cognitive activity in the brain. However, the emotional response, also referred to as affective response, in humans when interacting with technological systems should be of obvious interest for a wider range of professionals, including product developers and interaction designers. Accident statistics suggest that a majority of accidents in a wide range of industries are caused by human error (Hansen, 2007; Wiegmann & Shappell, 2017). Human error often occurs due to lapses of attention, an inability to understand the full consequence of an accident, or inadequacies in interface design. Effective design solutions that incorporate and recognizes human behavior and various types of responses are needed to mitigate human error. The increased awareness of the importance of the human response in interaction with technical systems has led to the call for integrating human affect when designing human-machine interactions (HMI) (Balters & Steinert, 2017). A multitude of physiology sensors can be used to better evaluate which product solution or interface design meets user needs in the best way (Balters & Steinert, 2014, 2017).

1.1.2 Increasing hardware availability

With the increased interest and research on BCI, there has been an improvement in the cost and accessibility of the necessary hardware to measure brain activity. EEG is a highly popular modality due to its non-invasive nature and high temporal resolution. Readily available, low-cost, consumer-grade EEG systems from several suppliers such as Emotiv, NeuroSky, and OpenBCI are being applied in a wide range of research today (Sawangjai, Hompoonsup, Leelaarporn, Kongwudhikunakorn, & Wilaiprasitporn, 2019). Although these devices do not possess the same certifications as the medical-grade equipment, studies have shown that they can be comparable in terms of quality (Frey, 2016). Frey (2016) compared the consumer-grade OpenBCI Cyton with the medical-grade g.USBamp EEG amplifier, and found almost identical performance in the P300 speller task and workload monitoring during the n-back task.

1.1.3 Advantages of multimodal data-capture

EEG is a neuroimaging tool measuring electrical brain activity, notably the local current flow caused by neural activation. The sum of all activity of neurons in close vicinity constitutes an EEG signal, which is measured by two or more electrodes (Balters & Steinert, 2017). Functional Near-Infrared Spectroscopy (fNIRS) is another neuroimaging technique that optically measures the hemodynamic response in brain tissue, i.e. concentration changes of oxygenated and deoxygenated hemoglobin following neural activation (Pinti et al., 2018). Two or more optodes, i.e. a minimum of one source and one detector must be used to measure cerebral blood flow. EEG has a high temporal resolution (in the millisecond range) and low spatial resolution (Balters & Steinert, 2017; Pinti et al., 2018). EEG is also susceptible to noise (Al-Shargie et al., 2016). fNIRS offers better spatial resolution (Al-Shargie et al., 2016; Pinti et al., 2018), is more robust to motion, but has lower temporal resolution, in addition to a ~5 second delay from stimuli onset to response peak due to the nature of the hemodynamic response (Pinti et al., 2018). Both EEG and fNIRS offer several advantages; they are non-invasive, portable, less expensive than other brain-imaging techniques (such as MRI and PET) and safe for longer-term monitoring (Al-Shargie et al., 2016; Pinti et al., 2018). They have been reported to be a good combination (Al-Shargie et al., 2016). Prior research has found that multimodal signal analysis can improve performance in predicting human reactivity (Cisler, Greenwood, Roberts, McKendrick, & Baldwin, 2019) and signal classification (Fazli et al., 2012; Lee, Fazli, Mehnert, & Lee, 2015). Fazi et al. (2012) found that combining fNIRS with EEG data significantly increased the accuracy in classifying visualized movements. Al-Shargie et al. (2016) found EEG and fNIRS measurements to improve the classification accuracy of mental stress, compared to EEG only and fNIRS only.

Multimodal integration of EEG and fNIRS is still relatively new (Ahn & Jun, 2017), which means there are limited options for simple multimodal data capture or integration. Current EEG-fNIRS integrations include; custom built fNIRS-system to fit an existing EEG-system with the need for two desktop computers for simultaneous data recording (Ahn, Nguyen, Jang, Kim, & Jun, 2016), or needing to purchase two separate systems, record data independently and having to fuse the data post-experiment (Al-Shargie et al., 2016; Fazli et al., 2012).

1.2 Objective and scope of the paper

Based on existing access to one brain activity sensor, namely fNIRS, we wanted to integrate EEG measurements to improve experimental results. Thus, we wanted a low-cost EEG integration to an existing fNIRS system to enable multimodal measurements during HMI experiments. It should be compatible with a standard electrode cap and optode placement, with the ability to quickly adapt to new setups, including different caps and sensor montages.

This paper presents the result, a functional prototype of an EEG integration. A low-cost, adaptive integration of dry-comb EEG-electrodes in a standard electrode cap, fitted with fNIRS optodes, has been developed. This paper demonstrates the development process, the final prototype and initial tests demonstrating a proof-of-concept. Following this introduction is a section describing the hardware basis used in the following development process. The methodology, the development process, and the test protocol are described in section three. Section four presents the resulting prototype, both the EEG adaption and system integration before showing the results of the proof-of-concept tests. Discussion and conclusion follow.

2 Hardware spesifications

The paper covers the integration of the Cyton biosensing board (*OpenBCI Inc.*, 2019) in combination with their spring-loaded, dry EEG electrodes (Figure 1) provided as part of the Ultracortex Mark IV EEG headset (*OpenBCI Inc.*, 2019). OpenBCI provides three different dry electrodes: flat, spikey or 5 mm combs respectively. The developed adapters are compatible with all the available dry electrodes OpenBCI provides. Among the low-cost EEG-systems available, OpenBCI was chosen due to minimalistic form factor, low weight, small size, and open-source nature, lending itself to be modified and hacked to meet the researchers' requirements. The performance of the Cyton board has also been found to be comparable to medical-grade EEG-amplifiers (Frey, 2016), which makes it an interesting option for non-clinical research. The key specifications of the OpenBCI Cyton are listed in Table 1.

Table 1. Key specifications OpenBCI Cyton

Channels	8
Compatible electrodes	Active & Passive
Data resolution	24-bit
Programmable gain	1, 2, 4, 6, 8, 12, 24
Operating voltage	3.3V Digital / +/- 2.5V analog
Amplifier	Texas Instruments ADS1299 ADC
Microcontroller	PIC32MX250F128B

OpenBCI provides spring-loaded, dry electrodes intended for use with their 3D-printed headset. Although this system might be well suited for simple, low-cost EEG applications, it has several drawbacks rendering it impractical for more elaborate experiments. Most notably, the headset only accommodates 35 electrode positions and does not lend itself to be used in combination with other sensors such as fNIRS. Thus, the authors needed to develop a set of custom adapters to integrate the EEG dry electrodes with a standard fNIRS cap. The goal was to integrate the spring-loaded dry EEG electrodes (Figure 1) provided by OpenBCI Inc. into a standard electrode cap, namely the EASYCAP AC-128-X1-C-58, which has a 128-channel layout according to the five percent system (Oostenveld & Praamstra, 2001). This system is a part of the international EEG system, a standardized method for consistent description of electrode placement on the scalp, to enable greater experiment replicability.

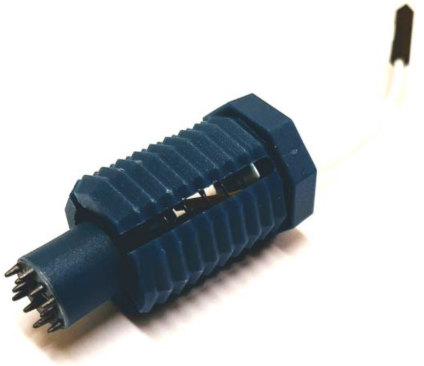


Figure 1. Spring-loaded dry electrode from OpenBCI.

3 Method and development

3.1 Development methodology

The guiding development methodology was Wayfaring (Steinert & Leifer, 2012). Rapid prototypes in cycles of design-build-test (Gerstenberg, Sjöman, Reime, Abrahamsson, & Steinert, 2015; Leikanger, Balters, & Steinert, 2016) were made based on initial hardware specifications and requirements for EEG integration. Design-build-test cycles are argued to be effective for problem-solving in product development projects (Wheelwright & Clark, 1994), suited for a dynamic environment (Gerstenberg et al., 2015).

3.2 Development process

The development process consisted of rapid iterations of several prototypes during design-build-test cycles. Since the product was small in size, prototypes could be produced quickly through 3D-printing. All prototypes were printed on a consumer-grade desktop 3D-printer, with the parameters specified in Table 2.

Table 2. Printing parameters.

Printer	Prusa MK3
Material	PLA
Nozzle diameter	0.4 mm
Extruder temperature	210 °C (first layer:215 °C)
Bed temperature	60 °C
Layer height	0.15 mm
Perimeters vertical shells	2
Solid layers top	7
Solid layers bottom	5

3.2.1 Iteration 1 – Mounting

A simple prototype with circular geometry was made based on manual measurements of the electrode, electrode housing and the optode-mounts that were supplied with the cap originally. The slots enabling the electrode to slide back and forth in the housing also make the housing flexible (Figure 1). The first prototype (Figure 2) was designed to test whether or not a simple snap-lock mechanism would be feasible due to the flexible properties of the electrode housing.

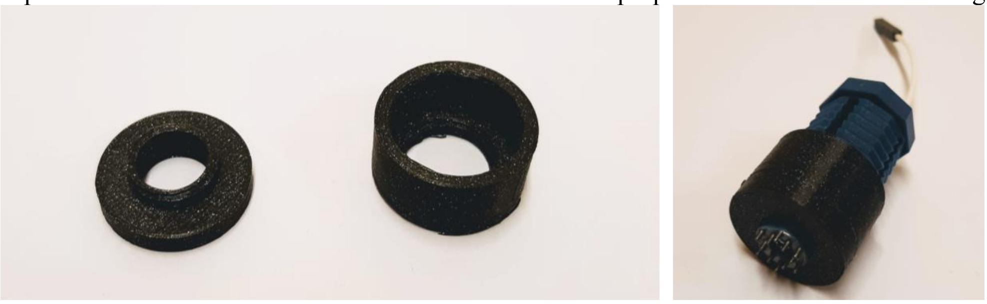


Figure 2. First iteration prototype. Bottom and top part of the adapter (left). Assembled adapter attached to the dry electrode (right).

Initial testing of the prototype showed that the snap-lock did not provide enough grip on the electrode housing when wearing the cap. The threads on the electrode housing also made the electrodes prone to tilting when interacting with the adapter. Additionally, the outer diameter of the bottom (scalp-facing) part was found to be too large, causing poor flexibility in the cap, if mounted closely together.

3.2.2 Iteration 2 – Threaded mount

Addressing the issues discovered by the first prototype, a threaded connection between the adapter and the electrode housing was implemented. The threads on the housing are not standard threads, which complicate precision in modeling. However, due to the open-source nature of the OpenBCI products, STL-files of all components of the Ultracortex Mark IV headset are freely available under the GNU General Public License on Github (*OpenBCI/Ultracortex*, 2015/2020). By importing the geometry of the threads into the design of the adapters, a good connection between the dry electrodes and the adapters was achieved.

The diameter of the bottom (scalp-facing) part of the adapter was also reduced to improve flexibility in the cap when several adapters are mounted in close proximity.

3.2.3 *Iteration 3 – Improving handling*

To improve the handling of the adapters during assembly and ensure better fit with the electrode housing, outer geometry was made octagonal. This made connecting the top and bottom part of the adapter easier when installing in the cap. As a secondary benefit, it reduced the tendency of the electrode housing to unscrew itself when mounting the cap on test-subjects. A 3D model of the final design is illustrated in Figure 3, while Figure 7 depicts the exterior and interior view of a fully mounted electrode.

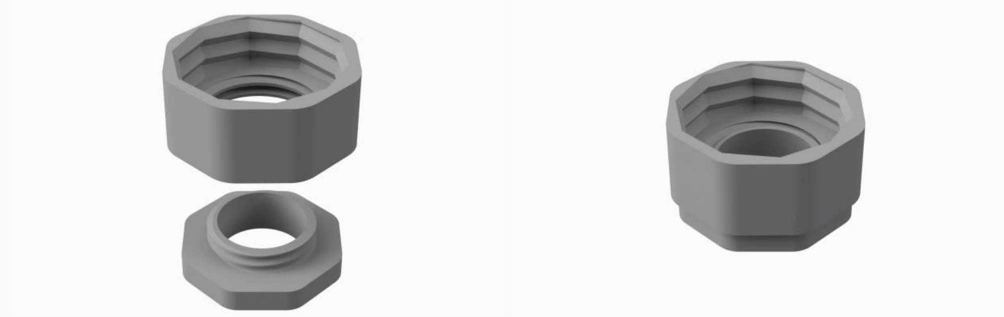
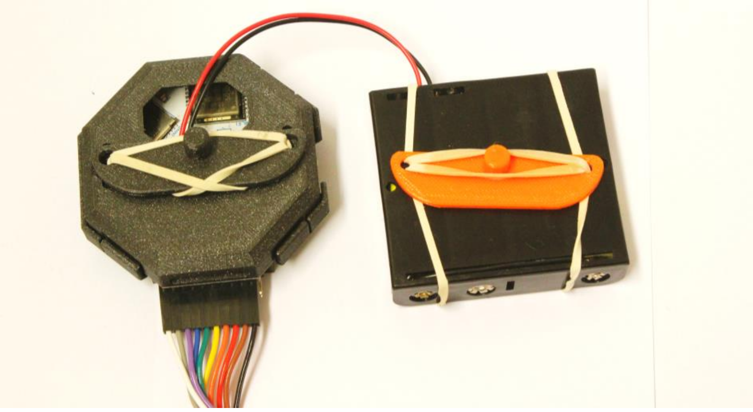


Figure 3. 3D-model of the final prototype. Exploded view (left) and assembled (right).

3.2.4 *Mounting the Cyton board and battery pack*

As the final step in development, fixtures to mount the Cyton board and the battery pack to the cap were prototyped, see Figure 4 and Figure 5. Although the board and battery pack does not necessarily have to be mounted on the subjects head, we deemed it beneficial to mount the Cyton board close to the electrodes. Placing the Cyton board elsewhere would require a longer wire, increasing electrical resistance in the system and making it more vulnerable to electrical noise. Furthermore, placing the board and battery on the subject's head imposes fewer restrictions on the subject's seating position or movement. This is advantageous for in-situ experiment applications. To fixate them to the cap, the existing fNIRS optode holders were exploited. Due to the low weight of the Cyton board and battery pack, a simple cylindrical feature with a tight-fit was sufficient to keep the components in place during stationary testing. The mounts can be assembled by means of screws, glue, or as demonstrated in Figure 4, with standard rubber bands. Due to the weight of the battery pack, the placement should be close to the top of the head (position Cz), to reduce the risk of inducing movement of the cap. Alternatively, a light-weight battery pack can be used. If this is not feasible with the desired



electrode montage, we propose using a battery pack with longer wires that enable fixating the battery pack to the subject's body instead.

Figure 4. Mounting fixture attached to the battery pack and Cyton board with housing, by means of rubber bands.

3.3 Testing signal quality – protocol for a proof-of-concept test

As a preliminary test of the signal quality, we performed two tasks that have been shown to reveal easily detectable changes in the EEG-signal: closing the eyes, and eye-blinking. Since formal EEG analysis is not within the scope of this paper, we relied on the output from the OpenBCI GUI software for these initial tests. EEG signals are often separated in distinct frequency bands during analysis (i.e. Delta (1-4 Hz), Theta (4-8 Hz), Alpha (8-13 Hz) and Beta (13-30 Hz)) and used to examine emotional states (Al-Shargie et al., 2016). Signal artifacts can be identified through visual inspection (Krishnaveni, Jayaraman, Aravind, Hariharasudhan, & Ramadoss, 2006) and used to assess if the EEG signal is correct. Three OpenBCI dry-electrodes were mounted in position Fp1 (channel 1), Fp2 (channel 2), and O2 (channel 8), while the reference electrodes were mounted in position A1, and A2. The locations Fp1 and Fp2 cover the prefrontal cortex, while position O2 covers the visual cortex. Positions A1 and A2 are situated on the left and right earlobe respectively (Figure 5).

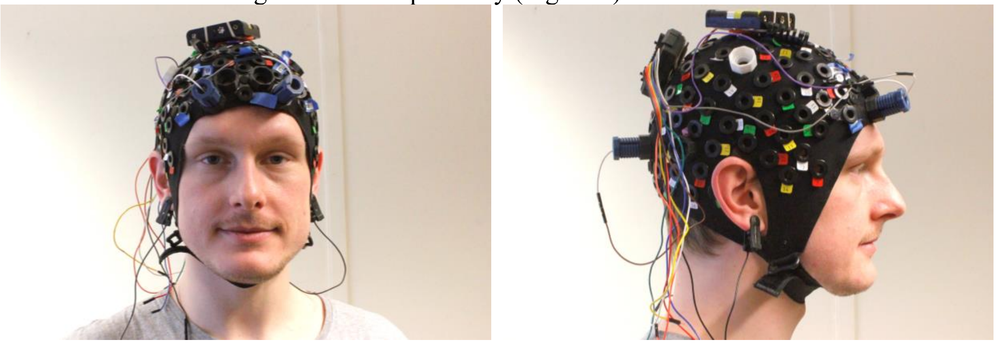


Figure 5. Setup during initial testing.

3.3.1 Induce increase in the alpha range

According to Mulholland (1995), closing the eyes should cause a visual increase in the alpha range (8-13 Hz) amplitude. Thus, we measured the EEG response with open eyes, closed eyes, and then after opening the eyes again. The data was captured and displayed by the OpenBCI GUI software (*OpenBCI GUI*, 2019) and visually inspected. The purpose of this test was to provide proof-of-concept.

3.3.2 Produce ocular artifacts

Electrooculographic (EOG) artifacts, also referred to as ocular artifacts (OA), are one of the most common and well-known disturbances in EEG signals (Krishnaveni et al., 2006; Vidal, 1973; Zeng, Song, Yan, & Qin, 2013). EOG artifacts are caused by interference from the electrical field that is induced when the eyeballs rotate in their sockets. Since these electrical potentials are very large compared to brain induced potentials, failure to identify EOGs would imply unacceptable signal quality. A simple test was conducted to verify the ability to detect EOGs. The subject wearing the headset would deliberately blink their eyes in intervals of roughly five seconds. Between blinks, efforts were made to restrain from blinking and movement of the eyes. Subsequent inspection of the signal response displayed in the time-series plot in the OpenBCI GUI was performed.

4 Results

This section first describes the proposed setup, e.i, the final prototype. An initial test was made as a proof-of-concept to ensure the final prototype functioned as intended. The results of these tests are described in section 4.2.

4.1 Resulting prototype

4.1.1 EEG integration

The final prototype consists of two octagonal parts with a circular, central hole for insertion of the EEG electrodes (Figure 3); The bottom part (height = 3 mm, outer diameter = 21.6 mm, inner diameter = 11.2 mm) is mounted on the inside of the cap, extending through the pre-cut holes in the cap and connected to the top part by means of a threaded connection (Figure 6). The top part (height = 9.35 mm, outer diameter = 25.6 mm) is threaded to allow fast fixation of the spring-loaded dry electrodes. Figure 7 illustrates a fully assembled adapter with a mounted electrode. The adapters can be mounted on any of the precut holes in the cap. The setup has been tested with EASYCAP AC-128-X1-C-58 (see Figure 6). The costs associated with the setup (excluding fNIRS device and cap) is limited to $849 for the hardware supplied by OpenBCI (excluding shipping), and roughly $1 in filament cost, assuming free access to a 3D-printer.

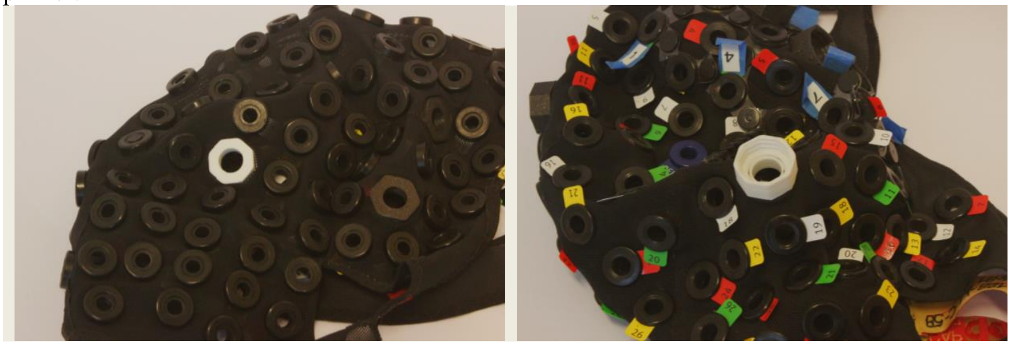


Figure 6. Mounted adapter, interior view (left), and exterior view (right).

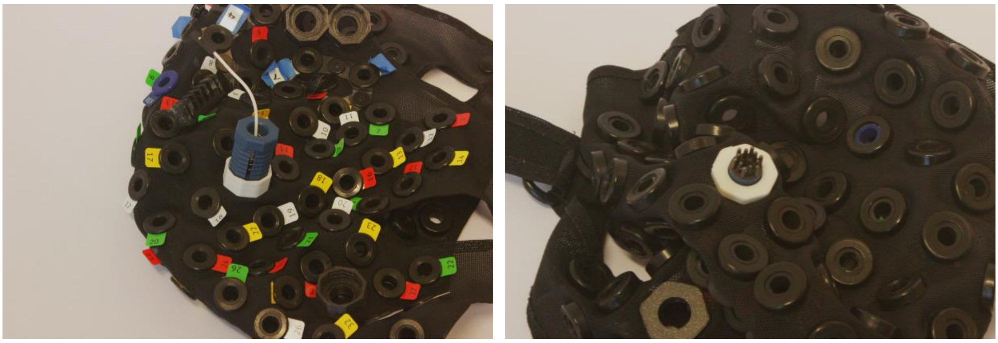


Figure 7. Dry EEG electrode mounted. Exterior view (left), interior view (right).

4.1.2 Simultaneous recording of fNIRS and EEG

Figure 8 shows the final setup including fNIRS. A pilot experiment utilizing the setup is currently being conducted which intends to verify the ability to record high-quality data of both modalities simultaneously. This experiment nor its data analysis is in the scope of this paper, so reservations must be made regarding the signal quality of the final setup. However, the results from the test described in section 4.2 indicate the feasibility of acquiring acceptable measurements.

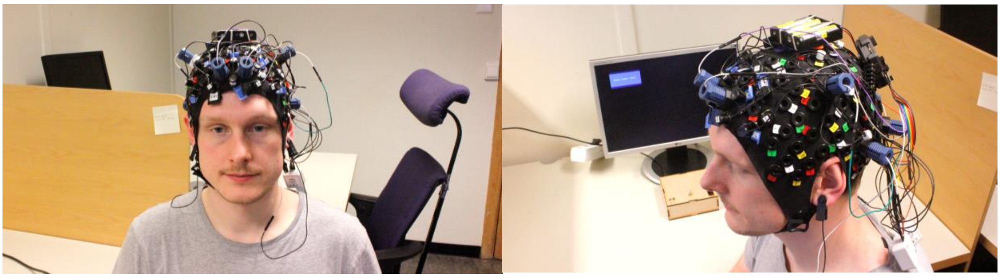


Figure 8. Eight channels EEG integrated and 20/40 channel fNIRS fully integrated.

4.1.3 Software integration

One of the advantages of the OpenBCI platform is that the raw data is readily available and accessible. If only EEG-recordings are of interest, raw data can be recorded with the free OpenBCI GUI application. For multimodal measurements, raw data can also easily be streamed using lab streaming layer, which enables synchronization of multiple sensors and forwarding to third-party software, see Figure 9. The proposed setup was tested by streaming EEG and fNIRS data simultaneously to iMotions (*IMotions*, 2020) successfully.

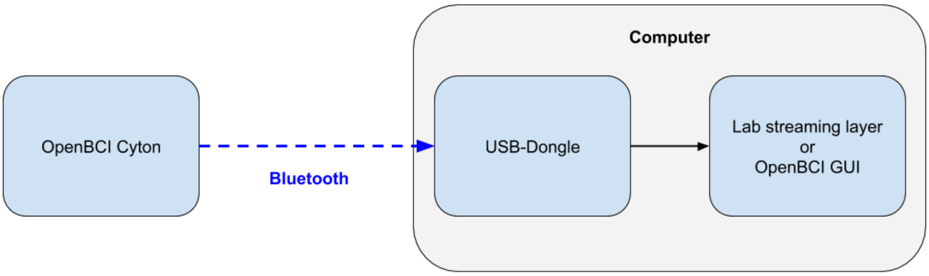


Figure 9. Schematic overview of raw data flow.

4.2 Proof-of-concept test

The results described in this section are from the initial tests made with EEG-electrodes integrated into EASYCAP using the final prototype, measuring EEG only.

4.2.1 Eyes closed

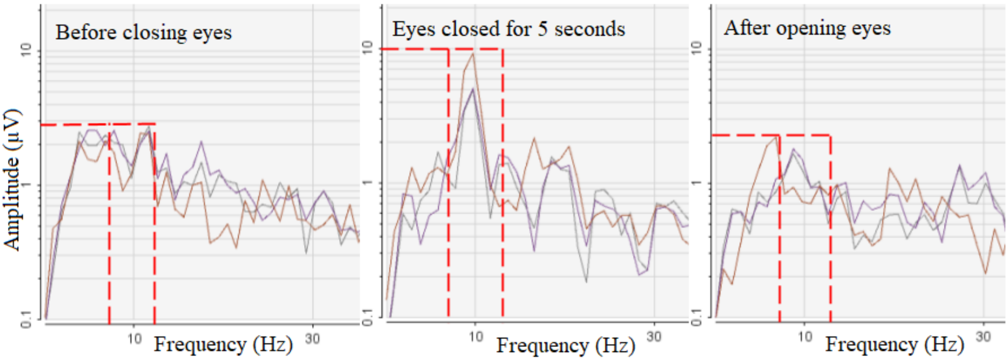
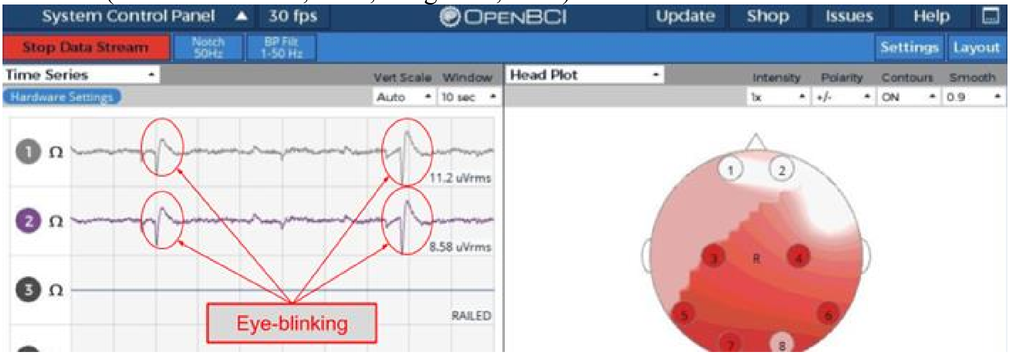


Figure 10. OpenBCI GUI.

Figure 10 shows the power spectrum (EEG-response) prior to closing the eyes, approximately five seconds after closing the eyes, and shortly after opening the eyes again. This clearly indicates that the peak amplitude in the alpha range increase during the period with eyes closed. Furthermore, this effect is reversed once the eyes are opened again. This response was as expected and in accordance with Mulholland (1995).

4.2.2 *Eye blinking*

Figure 11 shows the results of the test of controlled eye-blinking. The time-series plot of the recorded signals displays characteristic artifacts in agreement with what is reported in the literature (Krishnaveni et al., 2006; Zeng et al., 2013).



Vert Scale: Auto
Window: 10 sec
Intensity: 1x
Polarity: +/-
Contours: ON
Smooth: 0.9

Figure 11. OpenBCI GUI. EOGs clearly visible in channels one and two, positioned at Fp1 and Fp2 respectively.

5 Discussion

When used exclusively with EEG electrodes, the ability to detect EOGs and an increase in alpha-range frequencies was demonstrated using the developed EEG electrode integration. The results are in accordance with the expected behavior (Mulholland, 1995; Vidal, 1973; Zeng et al., 2013). The purpose of these tests was to provide proof-of-concept. Although the tests lack scientific rigor, it indicates that a minimal acceptable performance is achieved. Evaluation of electrode performance is a complicated issue in its own regard (Lopez-Gordo, Sanchez-Morillo, & Valle, 2014; Tăuţan, Serdijn, Mihajlović, Grundlehner, & Penders, 2013) and is considered beyond the scope of this paper. A larger experiment utilizing both modalities is currently being piloted to ensure sufficient data quality and system performance. Initial results are promising, however not within the scope of this paper. The setup has only been tested with EASYCAP AC-128-X1-C-58. The proposed solution should be applicable to other soft-fabric sensor caps. However, small modifications might be necessary. During preliminary testing of the setup on multiple subjects some challenges acquiring acceptable signals were encountered. This might be explained by differences in hair-thickness and general skin conductance. Skin conductance properties are highly individual, in addition to that some individuals do not exhibit sufficient levels of electrodermal activity for good measurements. Since the electrodes require skin contact, acquiring high-quality EEG-signals in hair-covered regions remains an inherent challenge and might require additional efforts in some cases. A conductive gel can be applied to the electrodes to better secure skin contact.

6 Conclusion

This paper describes the design and development of an adaptive EEG integration to a research-grade sensor cap ensuring electrode positioning according to the five percent system. The proposed setup enables a low-cost solution to supplement fNIRS-measurements with EEG data.

The proposed setup enables convenient and fast integration of low-cost, spring-loaded, dry electrodes, and it increases the available electrode positions compared to currently available equipment. It constitutes a pragmatic approach for integrating EEG-measurements with other brain activity sensors and requires minimal investment. The proposed solution aims to lower the barrier to include EEG as an additional modality in existing experimental setups, which can increase the quality of experimental results.

Acknowledgment

This research is supported by strategic funds from the department for Mechanical and Industrial Engineering (MTP) at NTNU.

References

Ahn, S., & Jun, S. C. (2017). Multi-Modal Integration of EEG-fNIRS for Brain-Computer Interfaces – Current Limitations and Future Directions. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00503

Ahn, S., Nguyen, T., Jang, H., Kim, J. G., & Jun, S. C. (2016). Exploring Neuro-Physiological Correlates of Drivers' Mental Fatigue Caused by Sleep Deprivation Using Simultaneous EEG, ECG, and fNIRS Data. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00219

Al-Shargie, F., Kiguchi, M., Badruddin, N., Dass, S. C., Hani, A. F. M., & Tang, T. B. (2016). Mental stress assessment using simultaneous measurement of EEG and fNIRS. *Biomedical Optics Express*, *7*(10), 3882. https://doi.org/10.1364/BOE.7.003882

Balters, S., & Steinert, M. (2014). *Decision-making in engineering-a call for affective engineering dimensions in applied engineering design and design sciences*. 11–15. IEEE.

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

Cisler, D. S., Greenwood, P. M., Roberts, D. M., McKendrick, R., & Baldwin, C. L. (2019). Comparing the Relative Strengths of EEG and Low-Cost Physiological Devices in Modeling Attention Allocation in Semiautonomous Vehicles. *Front. Hum. Neurosci.* https://doi.org/10.3389/fnhum.2019.00109

Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K.-R., & Blankertz, B. (2012). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *NeuroImage*, *59*(1), 519–529. https://doi.org/10.1016/j.neuroimage.2011.07.084

Frey, J. (2016). Comparison of an open-hardware electroencephalography amplifier with medical grade device in brain-computer interface applications. *ArXiv:1606.02438 [Cs]*. Retrieved from http://arxiv.org/abs/1606.02438

Gerstenberg, A., Sjöman, H., Reime, T., Abrahamsson, P., & Steinert, M. (2015). A Simultaneous, Multidisciplinary Development and Design Journey – Reflections on Prototyping. *Entertainment Computing - ICEC 2015*, 409–416. https://doi.org/10.1007/978-3-319-24589-8_33

Hansen, F. D. (2007). Human Error: A Concept Analysis. *Journal of Air Transportation*, *11*(3), 61–77.

IMotions (Version 8.1). (2020). Retrieved from https://imotions.com/

Krishnaveni, V., Jayaraman, S., Aravind, S., Hariharasudhan, V., & Ramadoss, K. (2006). Automatic Identification and Removal of Ocular Artifacts from EEG using Wavelet Transform. *MEASUREMENT SCIENCE REVIEW*, *6*(4), 13.

Lee, M.-H., Fazli, S., Mehnert, J., & Lee, S.-W. (2015). Subject-dependent classification for robust idle state detection using multi-modal neuroimaging and data-fusion techniques in BCI. *Pattern Recognition*, *48*(8), 2725–2737. https://doi.org/10.1016/j.patcog.2015.03.010

Leikanger, K. K., Balters, S., & Steinert, M. (2016). *Introducing the wayfaring approach for the development of human experiments in interaction design and engineering design science*. Presented at the DS 84: Proceedings of the DESIGN 2016 14th International Design Conference.

Lopez-Gordo, M. A., Sanchez-Morillo, D., & Valle, F. P. (2014). Dry EEG Electrodes. *Sensors (Basel, Switzerland)*, *14*(7), 12847–12870. https://doi.org/10.3390/s140712847

Mulholland, T. (1995). Human EEG, behavioral stillness and biofeedback. *International Journal of Psychophysiology*, *19*(3), 263–279. https://doi.org/10.1016/0167-8760(95)00019-O

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, *112*(4), 713–719. https://doi.org/10.1016/S1388-2457(00)00527-7

OpenBCI GUI (Version 4.1.7-beta.2). (2019). Retrieved from https://github.com/OpenBCI/OpenBCI_GUI

OpenBCI Inc. (2019). Retrieved from https://openbci.com/

OpenBCI/Ultracortex. (2020). Retrieved from https://github.com/OpenBCI/Ultracortex (Original work published 2015)

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2018). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, *n/a*(n/a). https://doi.org/10.1111/nyas.13948

Sawangjai, P., Hompoonsup, S., Leelaarporn, P., Kongwudhikunakorn, S., & Wilaiprasitporn, T. (2019). Consumer grade EEG Measuring Sensors as Research Tools: A Review. *IEEE Sensors Journal*, 1–1. https://doi.org/10.1109/JSEN.2019.2962874

Steinert, M., & Leifer, L. J. (2012). "Finding One's Way": Re-Discovering a Hunter-Gatherer Model based on Wayfaring. *International Journal of Engineering Education*, *28*(2), 251.

Tăuţan, A.-M., Serdijn, W., Mihajlović, V., Grundlehner, B., & Penders, J. (2013). Framework for evaluating EEG signal quality of dry electrode recordings. *2013 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 186–189. https://doi.org/10.1109/BioCAS.2013.6679670

Vidal, J. J. (1973). Toward Direct Brain-Computer Communication. *Annual Review of Biophysics and Bioengineering*, *2*(1), 157–180. https://doi.org/10.1146/annurev.bb.02.060173.001105

Wheelwright, S. C., & Clark, K. B. (1994). Accelerating the Design-build-test Cycle for Effective Product Development. *International Marketing Review*, *11*(1), 32–46. https://doi.org/10.1108/02651339410057509

Wiegmann, D. A., & Shappell, S. A. (2017). *A human error approach to aviation accident analysis: The human factors analysis and classification system*. Routledge.

Zeng, H., Song, A., Yan, R., & Qin, H. (2013). EOG Artifact Correction from EEG Recording Using Stationary Subspace Analysis and Empirical Mode Decomposition. *Sensors (Basel, Switzerland)*, *13*(11), 14839–14859. https://doi.org/10.3390/s131114839

Appendix C12: Academic contribution 12

Dybvik, H., & Steinert, M. (2021). Real-World fNIRS Brain Activity Measurements during Ashtanga Vinyasa Yoga. Brain Sciences, 11(6), 742. https://doi.org/10.3390/brainsci11060742

C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16

brain sciences

MDPI

Article

Real-World fNIRS Brain Activity Measurements during Ashtanga Vinyasa Yoga

Henrikke Dybvik * and **Martin Steinert**

TrollLABS, Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway; martin.steinert@ntnu.no

* Correspondence: henrikke.dybvik@ntnu.no; Tel.: +47-95-41-72-45

Abstract: Functional near-infrared spectroscopy (fNIRS) is often praised for its portability and robustness towards motion artifacts. While an increasing body of fNIRS research in real-world environments is emerging, most fNIRS studies are still conducted in laboratories, and do not incorporate larger movements performed by participants. This study extends fNIRS applications in real-world environments by conducting a single-subject observational study of a yoga practice with considerable movement (Ashtanga Vinyasa Yoga) in a participant's natural environment (their apartment). The results show differences in cognitive load (prefrontal cortex activation) when comparing technically complex postures to relatively simple ones, but also some contrasts with surprisingly little difference. This study explores the boundaries of real-world cognitive load measurements, and contributes to the empirical knowledge base of using fNIRS in realistic settings. To the best of our knowledge, this is the first demonstration of fNIRS brain imaging recorded during any moving yoga practice. Future work with fNIRS should take advantage of this by accomplishing studies with considerable real-world movement.

Keywords: fNIRS; cognitive load; human cognition; real-world; in situ; ecological validity; ashtanga; yoga

check for updates

Citation: Dybvik, H.; Steinert, M. Real-World fNIRS Brain Activity Measurements during Ashtanga Vinyasa Yoga. *Brain Sci.* **2021**, *11*, 742. https://doi.org/10.3390/brainsci11060742

Received: 29 April 2021
Accepted: 31 May 2021
Published: 3 June 2021

1. Introduction

Functional near-infrared spectroscopy (fNIRS) is a non-invasive, lightweight, and portable neuroimaging technique which measures cortical brain activity [1,2]. fNIRS uses optical fibers to emit near-infrared light into a region of the brain, and detect changes in blood flow oxygenation (oxygenated (ΔHbO)) and deoxygenated hemoglobin (ΔHbR), caused by neural activation [1]. The light of different wavelengths in the near-infrared (NIR) spectrum penetrates the scalp and travels through different layers of the head, before reaching neuronal tissue. Inside the tissue, NIR light is absorbed differently in hemoglobin depending on the oxygen saturation state. Non-absorbed light scatter components are detected, and ΔHbO and ΔHbR are calculated by the modified Beer-Lambert Law. Neural activity induces changes in local hemodynamics, causing an increase in HbO concentration in the activated region, and a decreased concentration of HbR [1–3] (although this is not always the case [4]). This is used to measure cognitive states and cognitive load [5–8].

For cerebral hemodynamics, fNIRS can act as a surrogate for functional magnetic resonance imaging (fMRI) [9–11]. FNIRS is not limited to the restrictive fMRI environment, and since it is relatively robust against motion artifacts, the technique allows for freely moving participants in contexts with high ecological validity and in the real world (or in situ) [1,2,12]. Examples of such studies include, but are not limited to, outdoor activities, such as riding a bike [13] and walking [14]; farm workers at individual farm locations [15]; driving a car on an expressway [16]; setting a table [17]; radiologists interpreting MRI and CT images [18]; exposure therapy of arachnophobia [19]; performing penalty kicks in soccer [20]; playing table tennis, playing the piano, and human interaction during a violin duo [12]. Such studies are important since they may help us understand how the brain functions in real-life situations. They may also allow us to detect brain activity that can only be detected during movement. For example, one study revealed that the cortical activation from conducting an everyday task was not detected during an imitation of the same task [21]. To investigate and understand brain activity during *any* activity or task, it is, therefore, best to measure it directly in the environment where it naturally occurs (in situ). In situ studies may, for example, be beneficial when analyzing sports or strenuous exercise, social interaction in natural environments [13], operators at work (air traffic controllers [22], captains [23], and drivers [24,25]), and walks in nature. Moreover, when studying populations who may not be able to come to the lab (e.g., severe Alzheimer's patients [26]), or when coming to a lab would be counterproductive to the topic of interest (e.g., during physical therapy and rehabilitation [27,28]), it may be necessary to conduct the study in the participant's own environment.

FNIRS is often praised for its portability and robustness towards motion artifacts. However, most fNIRS studies are still conducted in laboratories today. An increasing body of research uses fNIRS in real-world environments at moderate levels of motion; indeed, several of the studies mentioned above include moderate levels of motion, e.g., [12], but none incorporate considerable or vigorous movements. We believe that fNIRS can be applied in real situations to a greater extent than it is currently. There is a need for in situ fNIRS studies with considerable movement.

Thus, the aim of this study was to extend fNIRS applications in real-world environments by recording fNIRS during a moving yoga practice in a participant's natural environment. Ashtanga Vinyasa Yoga is a practice with considerable movement and complex postures, which may have some effect on cognitive functions. Therefore, we explored changes in brain activity (prefrontal cortex activation) within postures in the Ashtanga primary series. The research objectives were as follows: (1) Test the feasibility of fNIRS recordings during a yoga practice with considerable movement. (2) Test if different yoga postures have different cognitive loads. To this end, a single-subject observational study was adopted, in which one participant practiced Ashtanga with a wearable fNIRS in their own apartment for a total of seven times. The results show differences in cognitive load when comparing technically complex postures to relatively simple ones, but also some contrasts with little difference, although a greater difference was hypothesized. The fNIRS measurements taken during Ashtanga Vinyasa Yoga deepen our understanding of the effect of yoga postures and thus contributes to the scientific foundation of yoga. This study explores the boundaries of cognitive load measurements in the real-world, and contributes to the empirical knowledge base of using fNIRS in realistic settings.

2. Background

2.1. Yoga

The term "Yoga" denotes a group of physical, mental, and spiritual practices originating in ancient India [29–31]. Today, modern schools of yoga and thus styles of yoga each have a distinct relative content of ethics (*yama* and *niyama*), physical postures and exercises (*asanas*), breathing techniques (*pranayama*), and meditation practices, which aims to cultivate awareness; unite the mind, body, and spirit, alleviating suffering; and ultimately obtain profound states of consciousness [29–32]. Meditation practices include sensory withdrawal (*pratyahara*), concentration (*dharana*), meditation (*dhyana*), and a deep level of concentration (or absorption) described as self-transcendence (*samadhi*) [30].

2.2. Ashtanga Vinyasa Yoga

Ashtanga Vinyasa Yoga (Ashtanga for short) is a popular and physically demanding yoga style [33–35]. It is known for its vigorous flow, which may be why some adaptations of the practice are known as power yoga [36]. In Ashtanga, physical postures (*asanas*) are linked by flowing movements (*vinyasas*) and synchronous breathing techniques (*pranayama*) [35–37]. An Ashtanga session begins with sun salutations as a warmup, followed by a predefined sequence of postures, and a closing sequence. A total of six series exists, each with different sequences. The primary series is often called yoga therapy or yoga for health. It focuses on health healing effects, the release of trapped emotions, and raising and overcoming emotional and other unhealthy habitual patterns [37,38]. Ashtanga focuses on the coordination of posture, breath, and gaze [34]. These components form the *Tristana*, which is unique to Ashtanga [34]. A strong focus on physical embodiment is necessary since the postures are technically complex, and each movement is coordinated with an inhale or exhale, while the postures are held for five breaths [34,39]. The breathing technique is called *Ujjayi Breathing*, or victorious breath [39]. Each posture and movement has a specific gaze point intended to reduce external distractions and induce concentration (e.g., navel-gazing or *omphaloskepsis*, which is defined as the "contemplation of one's navel as an aid to meditation" [40]) [34]. The repetitive practice is intended to move practitioners towards a control of mental activity that enables true self-realization [37,38]. Due to this highly focused attention during bodily movements, yoga is often called "meditation in motion" [30]. The rigid adherence to a standardized and documented posture series makes Ashtanga a strong candidate for scientific study [35].

2.3. Existing Research on Yoga

An increasing body of research shows positive effects from yoga practices and interventions on physical and psychological health [29,30]. Symptoms of depression, PTSD, epilepsy, ADHD, stress, and anxiety have been alleviated with yoga-based therapies. A reduction in stress and anxiety symptoms is also found in healthy individuals [30,41]. Yoga practitioners report increased psychological wellbeing, life satisfaction, happiness, motivation, and relaxation [30,36,42]. Reduced levels of galvanic skin response and blood lactate have been measured [41], along with improvements in physical fitness, and reduced sympathetic nervous system activity [33,39]. Some can also reduce their heart rate voluntarily without external cues [32]. Ashtanga practitioners specifically show improvements in muscular strength, endurance, flexibility, health perception, diastolic blood pressure, perceived stress [39], cardiac and respiratory fitness [33], and self-transcendence [43]. The results of Ashtanga intervention studies show significant improvements in psychological wellbeing, self-esteem, assertiveness, attention to one's needs, and capacity to connect [34,44].

Several yoga techniques claim to enhance cognitive and executive functions, such as attention/awareness, concentration, emotion regulation, and cognitive control. Studies of such found greater gray matter volume, increased functional connectivity, improved cognitive performance, the strengthening of interoceptive and executive/control networks for yoga practitioners, and decreased glucose metabolism (which is linked to the improved regulation of negative emotions) [45]. Both elderly and adolescent practitioners have significantly improved cognitive performance [46,47]. A functional near-infrared spectroscopy (fNIRS) study found increased blood flow (measured by HbO concentration) to the dorsolateral prefrontal cortex during a yoga breathing technique [48] and increased bilateral blood flow to the prefrontal cortex in yoga practitioners compared to non-practitioners during sustained attention [49]. Electroencephalograms (EEGs) have also been conducted during various yoga practices with little movement [50]. In [30], the findings are consistent with the notion that yoga can improve cognitive regulation, to the point of offsetting the age-related decline in fluid intelligence in practitioners. They further suggest that this may be explained by the increased availability of neural resources, and postulate that neuronal interactions occurring during yoga practice include the cortical regions (i.e., the dorsolateral prefrontal cortex (DLPFC), anterior cingulate cortex (ACC), and orbitofrontal cortex) [30].

2.4. The Prefrontal Cortex

Executive and cognitive functions, such as attention/awareness, working memory, cognitive flexibility, and cognitive control and planning, are performed by the prefrontal cortex (PFC) [51–54]. The PFC synthesizes diverse information related to a given goal; it is responsible for planning and selecting complex cognitive behavior; and it is crucial for higher order processing [52]. Further studies relate cognitive control to activity in dorsolateral PFC (DLPFC) [55]. DLPFC plays an important role in the anticipatory organization of action and effortful tasks [56,57]. The mid-dorsolateral PFC (mDLPFC) aids in planning action sequences (organization external/internal action), i.e., mental conception and evaluation of behavioral sequences and associated outcomes before execution [56]. The frontopolar cortex (FPC) is involved in mind wandering, planning, abstract reasoning, multitasking, and cognitive branching, which require switching away from an ongoing behavioral option, considering multiple behavioral options, and/or exploring new ones [58]. Thus, the FPC is suggested to make a crucial contribution to the exploration and rapid acquisition of novel behavioral options [59]. The medial FPC governs undirected exploration, i.e., monitoring the current goal for possibly redistributing cognitive resources to other potential goals. The lateral (right/left) FPC cortex governs directed exploration, i.e., monitoring a few alternative tasks/goals for possibly re-engaging one as a replacement of the current task/goal [58].

3. Materials and Methods

Given that there are cognitive benefits of yoga practices, and that executive and cognitive functions are governed by the PFC, our aim was to determine whether we could measure any of them. Specifically, we investigated whether there were differences in cognitive load and cognitive state within postures in the Ashtanga primary series. The features of Ashtanga makes its practice a suitable context for demonstrating that brain activity can be measured during vigorous movement. To our knowledge, while intervention and laboratory studies of various yoga practices have been conducted, there is no study that incorporates neuroscientific measurements during a moving yoga session, nor in a real-world setting. We thus believe that this study is the first to record fNIRS during a moving yoga session.

3.1. Single-Subject Observational Study

A single-subject real-world (in situ) study of the half primary series in Ashtanga was conducted in the participant's own living room. The participant was fitted with an fNIRS sensor cap aided by another person. The signal quality check, and the start and stop of data collection, was performed by the participant.

A video of the full primary series performed and led by Ty Landrum was used for instructional purposes [60]. In this video, the yogi practices the sequence and gives voice-over instructions with postures and queues. The yogi also performs the opening and closing mantra (or chant). The participant listened to the chants in Mountain Pose with their hands in prayer position. This video was used in all sessions. The practice was adapted to accommodate the head-mounted sensors, i.e., the participant refrained from postures requiring the head to be placed on the ground. Table A1 in Appendix A includes a list of the postures that were a part of the practice, and if they were adapted or not conducted. The postures were held for five long breaths. Repeated measures over time were made corresponding to the repeated postures over several full practices, as explained in the introduction.

3.2. Participant

The participant was 26–27 years old, female, right-handed, had corrected-to-normal vison, and had a good general physical fitness level. The participant had 3 months of practicing the primary series in Ashtanga once a week and was, therefore, considered a novice in Ashtanga. She had two years of experience practicing other yoga types at the time of recording. The participant was obtained by convenience sampling and had worn fNIRS ahead of this study.

3.3. Data Collection

fNIRS data were sampled at 7.81 Hz by NIRSport (NIRx Medical Technologies, Berlin, Germany) with 8 sources and 8 detectors at two wavelengths (760 and 850 nm). Optodes were placed on the PFC, per montage by NIRx, as illustrated in Figure 1. The sources fNIRS data were sampled at 7.81 Hz by NIRSport (NIRx Medical Technologies, Berlin, Germany) with 8 sources and 8 detectors at two wavelengths (760 and 850 nm). Optodes were placed on the PFC, per montage by NIRx, as illustrated in Figure. 1. The sources (denoted Sx) were placed as follows: S1: F3; S2: AF7; S3: AF3; S4: Fz; S5: Fpz; S6: AF4; S7: F4; S8: AF8. The detectors (denoted Dx) were placed as follows: D1: F5; D2 F1; D3 Fp1; D4 AFz; D5 F2; D6: Fp2; D7: F6. We used an EASYCAP AC-128-X1-C-58 (EASYCAP GmbH, Herrsching, Germany) with a 128-channel layout following the 10–5 system [61]. This montage covers the anterior frontal lobe, more specifically anterior regions of the right, left, and mid-dorsolateral PFC (l/r/mDLPFC), and right, left, and medial FPC (r/l/mFPC). A sensitivity profile of the montage/probe was generated with AtlasViewer [62] and is illustrated in Figure 1c.

In the bilateral PFC, HbO activation increases linearly with increasing cognitive load. In the PFC, an increase in functional connectivity between hemispheres and across hemispheres is associated with increasing cognitive load. Moreover, functional connectivity is different for different cognitive states [51].

(denoted Sx) were placed as follows: S1: F3; S2: AF7; S3: AF3; S4: Fz; S5: Fpz; S6: AF4; S7: F4; S8: AF8. The detectors (denoted Dx) were placed as follows: D1: F5; D2 F1; D3 Fp1; D4 AFz; D5 F2; D6: Fp2; D7: F6. We used an EASYCAP AC-128-X1-C-58 (EASYCAP GmbH, Herrsching, Germany) with a 128-channel layout following the 10–5 system [61]. This montage covers the anterior frontal lobe, more specifically anterior regions of the right, left, and mid-dorsolateral PFC (l/r/mDLPFC), and right, left, and medial FPC (r/l/mFPC). A sensitivity profile of the montage/probe was generated with AtlasViewer [62] and is illustrated in Figure 1c.

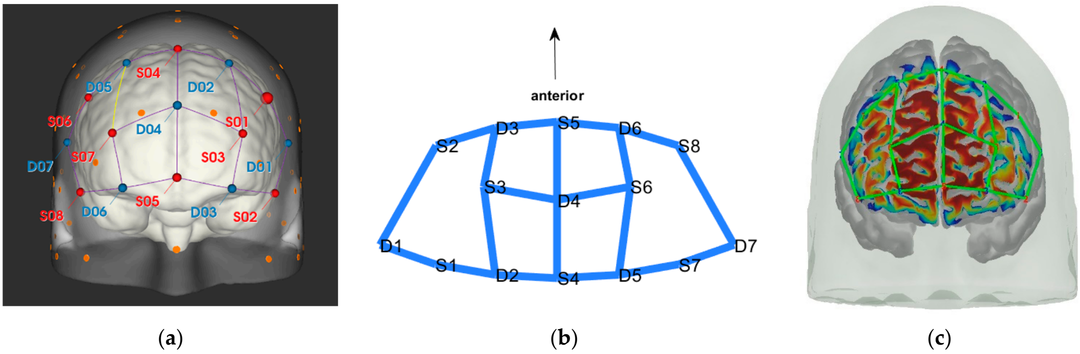


Figure 1. (**a**) The montage rendered onto the ICBM 152 Nonlinear atlases version 2009 [63,64], created with NIRSite 2020.7 (NIRx Medical Technologies). (**b**) The channels from above in an anterior orientation, created with NIRS toolbox [65]. (**c**) The sensitivity profile of the probe, created with AtlasViewer [62].

Figure 1. (**a**) The montage rendered onto the ICBM 152 Nonlinear atlases version 2009 [63,64], created with NIRSite 2020.7 (NIRx Medical Technologies). (**b**) The channels from above in an anterior orientation, created with NIRS toolbox [65]. (**c**) The sensitivity profile of the probe, created with AtlasViewer [62].

NIRSport was connected via cables to a Dell Latitude 7490 laptop (with Microsoft Windows 10 Education, Intel(R) Core(TM) i7-8650U CPU @ 1.90 GHz 2.11 GHz processor, 32.0 GB RAM installed, 64-bit operating system, x64-based processor, and a 500 GB SSD harddrive). Nirstar 15.2 Acquisition Software (NIRx Medical Technologies) [66] was used to forward a continuous data stream through LabStreamingLayer [67] to iMotions 8.1 [68], which synchronized physiology data and video recordings. Two video recordings were made with a laptop-integrated web camera and an additional external web camera (Logitech HD Pro Webcam C920, Newark, NJ, USA).

In total, seven (*N* = 7) yoga sessions were recorded from April to September 2020. Session 1 was a pilot session, lasting only 50 mins due to software issues; therefore, there are no data from the last part of this practice. There was a problem with the external webcam in sessions 1 and 4, and so this footage is incomplete.

In the bilateral PFC, HbO activation increases linearly with increasing cognitive load. In the PFC, an increase in functional connectivity between hemispheres and across hemispheres is associated with increasing cognitive load. Moreover, functional connectivity is different for different cognitive states [51].

NIRSport was connected via cables to a Dell Latitude 7490 laptop (with Microsoft Windows 10 Education, Intel(R) Core(TM) i7-8650U CPU @ 1.90 GHz 2.11 GHz processor, 32.0 GB RAM installed, 64-bit operating system, x64-based processor, and a 500 GB SSD harddrive). Nirstar 15.2 Acquisition Software (NIRx Medical Technologies) [66] was used to forward a continuous data stream through LabStreamingLayer [67] to iMotions 8.1 [68], which synchronized physiology data and video recordings. Two video recordings were made with a laptop-integrated web camera and an additional external web camera (Logitech HD Pro Webcam C920, Newark, NJ, USA).

In total, seven (*N* = 7) yoga sessions were recorded from April to September 2020. Session 1 was a pilot session, lasting only 50 mins due to software issues; therefore, there are no data from the last part of this practice. There was a problem with the external webcam in sessions 1 and 4, and so this footage is incomplete.

3.4. Data Analysis

3.4.1. Video Coding

Physiology data were labeled with names of postures (English translations of Sanskrit) from half primary series in Ashtanga post recording by video coding in BORIS [69]. A total of 57 postures were used in the analysis; see Table A2 in Appendix B for a list of these postures. Information on head position (up, down, side, and uptilted; side and down-tilted; or changing), whether the pose required bilateral, or unilateral muscle activation (left and right sides, respectively), its order in the sequence and which part of the sequence it belonged to (warm-up, standing, seated, or finishing) were added as metadata.

3.4.2. fNIRS Analysis

FNIRS data were prepared in python and analyzed in MATLAB R2020a with NIRS toolbox [65]. The standard pre-processing included conversion from raw data to optical density, then conversion to hemoglobin concentration using the modified Beer–Lambert Law with extinction coefficient from [70], and a partial pathlength factor (PPF) of 0.1. Statistical analysis used a first-level general linear model regression, which used an autoregressive pre-whitening method with iteratively reweighted least-squares (AR-IRLS) [65,71] for motion artifact correction that controls type-I errors. For second-level (group) statistics, a linear mixed-effects model was used, which tested for the main effect of conditions (the yoga postures constitute the conditions). A mixed-effects model was selected since it more effectively accounts for design imbalances and missing values. Then, we ran all permutations of condition contrasts with *t*-tests; i.e., each posture was compared to all other postures. To correct for multiple comparisons, the Benjamini–Hochberg procedure was used, and the corrected *p*-value is denoted as the q-value [72]. The statistical significance level was set at q < 0.05. We refer to [65,71,73] for further details on analysis techniques.

3.4.3. Selection of Contrasts

A number of $P(57,2) = 3192$ permutations were tested. After removing duplicate contrasts, a total of 1309 contrasts had one or more statistically different channels. We removed postures where the top of the head was facing down or changing position during the posture as this causes increased or unstable blood flow to PFC due to gravity. This returned 939 contrasts. Then, we made sure to only compare postures with the same head position, which resulted in 371 significantly different contrasts. These contrasts were sorted based on the number of significantly different channels. Thereafter, the ten contrasts with the most significantly different channels, the ten contrasts with the least significant different channels, and the ten contrasts with a mid-range number of significantly different channels were selected for visual inspection. The visual inspection of the 30 contrasts determined the postures presented in the results.

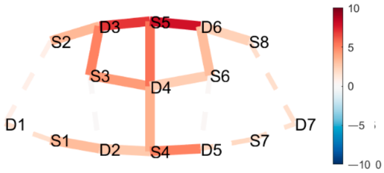
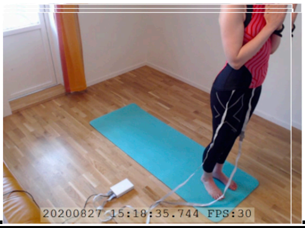
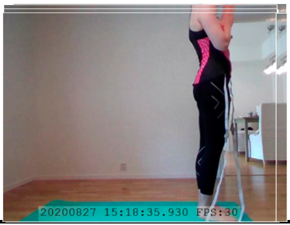
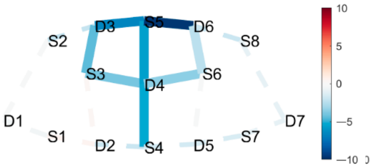
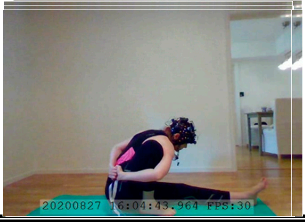
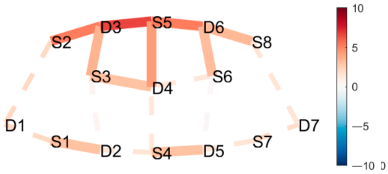
4. Results

First, we present combinations of postures yielding the greatest number of channels with expected significantly different brain activation or cognitive load. Second, we present posture combinations that show some unexpected differences in activation. Finally, we also highlight some interesting combinations with little significant difference, i.e., postures where we hypothesized differences but were unable to measure any greater significant difference.

4.1. Posture Combinations with Expected Greater Statistical Difference

Boat–Mountain Pose Start. There was a significant increase in HbO in the Boat compared to the Mountain Pose Start, in 21 channels. Activations occurred along the medial line and parts of r/lPFC, and mDLPFC; see Table 1, row 1, for Hbo visualization, and Table A3 (Appendix C) for statistics. Increased HbO in the PFC, and hemisphere connectivity indicate that the Boat Pose has a higher cognitive load compared to the Mountain Pose. This is an expected finding. Mountain Pose Start is the opening posture of the sequence, by some described as a resting pose. Instructions usually include clearing the head and preparing for the practice. In contrast, Boat is physically demanding in that muscle activation, posture, and balance require presence and concentration. Boat is strenuous, but is claimed to aid in developing concentration stamina, focus, internal awareness, and emotional calmness. Activation of the mFPC may indicate a mental exploration of behavioral changes (e.g., changes in posture and muscle activation) to make the posture easier. The activation predominantly in the mFPC, which is responsible for undirected exploration, may suggest that the practitioner was seeking a new, unknown alternative (as opposed to directed exploration of a known alternative). Activation in the mDLPFC, which plans upcoming behavioral actions, may further suggest that the practitioner had begun to think about and plan upcoming postures. It is possible that there was an attempt to divert thoughts from the strenuous Boat posture.

Table 1. Contrasts postures with greater statistical differences. Contrasts are shown as *t*-tests of Posture 1 minus Posture 2 [1].

Posture 1	Contrast Statistics Visualized on Brain Montage	Posture 2
Boat	HbO	Mountain Pose Start
		
Mountain Pose Start	HbO	MA Right Knee Bent Bind
		
Boat	HbO	Lotus
		

[1] Second-level analysis of HbO and HbR changes for selected contrasts; *t*-statistic maps with significantly different brain activation (*q* < 0.05) in solid lines; color bar represents the *t*-statistic (scaled to range [−10, 10]) with red/blue indicating statistical increase/decrease in HbO concentration between the postures.

4.2. Posture Combinations with Some Unexpected Differences (Mid-Range)

[illegible] series of four postures (A, B, C, D) to show the difference [illegible] of the posture added given the name [illegible] (translation from Sanskrit) Right Knee Bent Bind. There was [illegible] HbO in Mountain posture compared to the MA Right Knee Bent Bind [illegible] activation occurred along the medial line, mPFC, and partly the r/l PFC. See Table 1, row 2 for HbO visualization and Table A4 (Appendix C) for statistics. As expected, Mountain pose is thought [illegible] cognitively demanding than the MA Right Knee Bent Bind. This seated pose stretches the hamstrings and back by performing a forward fold and binds the hands behind the back to open the chest. Medial line and dl/r/l PFC deactivation suggest mental exploration, perhaps for an improved performance of the posture. Moreover, insignificant DLPFC activity suggests that the practitioner did not divert their attention from this posture to seek alternatives (as opposed to the Boat—Mountain Pose Start contrast above). The MA Right Knee Bent Bind offers a higher cognitive load than Mountain Pose, but it does not seem so intense that the practitioner wants to exit the pose as quickly as in Boat.

Head to Knee Pose B, sit on right heel—Extended Hand to Right Big Toe, hold: There was a significant increase in HbO in mPFC for Head to Knee Pose B compared to Extended Hand to Right Big Toe Hold. See Table 2, row 1 for HbO visualization, and Table A6 (Appendix C) for statistics. Head to Knee Pose B poses a greater cognitive demand than the Extended Hand to Right Big Toe Hold. This is interesting since we expected the opposite effect. Extended Hand to Right Big Toe imposes a great demand on leg and thigh muscles to statically hold the leg out and up in a straight line while maintaining balance. Additionally, the practitioner was not strong enough to hold their leg straight out and up, but worked hard to not let their foot fall further. Head to Knee Pose B is a seated position where practitioners sit on their heel, folding forward, stretching the hamstring and the back of the leg. The posture is generally considered uncomfortable due to the pressure of the heel on the perineum, especially for novice practitioners. This pain may have caused the unexpected finding. Head to Knee Pose (Sanskrit: Janu Sirsasana) is a hip opener, hamstring stretch, and a slight torso twist, with three variations (A, B, and C) on both the left and right sides, which become incrementally more difficult. A recurring pattern of HbO recruitment to mPFC (as depicted in Table 1, row 1) appeared when comparing the Extended Hand to Right Big Toe Hold to both the left and right sides of variation A and B, and the left side of variation C. However, the right side of variation C only recruited D3-

Boat–Lotus. There was significant increase in HbO in Boat compared to Lotus in 19 channels. Activations occurred in mFPC, some r/lFPC lateralization prevailing in the left hemisphere, and mDLPFC. This indicates a clear difference in cognitive load and state. See Table 1, row 3 for HbO visualization, and Table A5 (Appendix C) for statistics. This result was also expected since Boat was compared to the traditional meditation pose Lotus, which aims to clear the mind. Interestingly, there was less significant activation along the medial line as compared to the Boat–Mountain Pose Start contrast. Activation was predominantly at the montage's front (mFPC) and back (mDLPFC). This may be due to the greater similarity of these postures, i.e., they are both seated postures, whereas the two former contrasts compare standing and seated postures. We conclude that standing and seated postures have differing requirements.

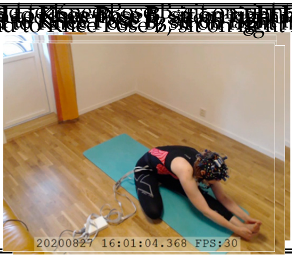
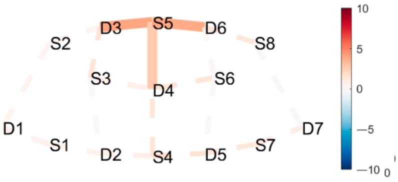
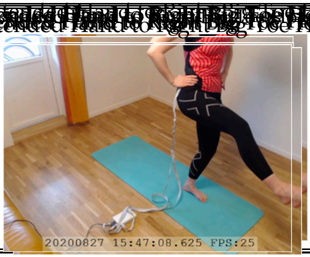
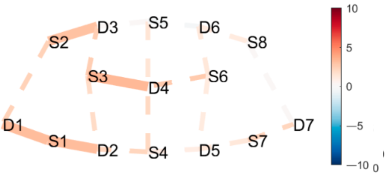
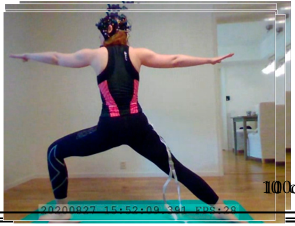
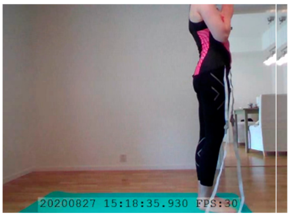
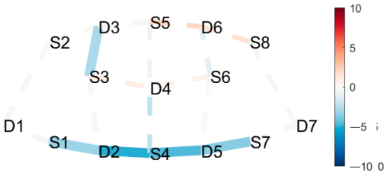
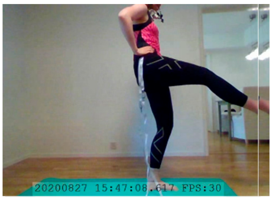
4.2. Postures Combinations with Some Unexpected Differences (Mid-Range)

Head to Knee Pose B, sit on right heel–Extended Hand to Right Big Toe, hold. There was a significant increase in HbO in mFPC for Head to Knee Pose B compared to Extended Hand to Right Big Toe Hold. See Table 2, row 1 for HbO visualization, and Table A6 (Appendix C) for statistics. Head to Knee Pose B poses a greater cognitive demand than the Extended Hand to Right Big Toe Hold. This is interesting since we expected the opposite effect. Extended Hand to Right Big Toe imposes a great demand on leg and thigh muscles to statically hold the leg out and up in a straight line while maintaining balance. Additionally, the practitioner was not strong enough to hold their leg straight out and up, but worked hard to not let their foot fall further. Head to Knee Pose B is a seated position where practitioners sit on their heel, folding forward, stretching the hamstring and the back of the leg. The posture is generally considered uncomfortable due to the pressure of the heel on the perineum, especially for novice practitioners. This pain may have caused the unexpected finding. Head to Knee Pose (Sanskrit: Janu Sirsasana) is a hip opener, hamstring stretch, and a slight torso twist, with three variations (A, B, and C) on both the left and right sides, which become incrementally more difficult. A recurring patten of HbO recruitment to mFPC (as depicted in Table 1, row 1) appeared when comparing the Extended Hand to Right Big Toe Hold to both the left and right sides of variation A and B, and the left side of variation C. However, the right side of variation C only recruited D3-S6 and S5-D6, but not S5-D4 (along the medial line), as prior variations did. Since this is the final side in the last variation of this pose, this may suggest the anticipation of and preparation for the next pose, and thus a decreased activation in mFPC overall possibly due to decreasing focus and attention to the posture.

Lotus Uplifting–Warrior 2 Left. There was a significant increase in HbO for Lotus Uplifting compared to Warrior 2 Left, lateralized to left hemisphere, in lFPC, lDLPFC. Lotus Uplifting is more cognitively demanding than Warrior 2 Left. See Table 2, row 2 for HbO visualization, and Table A7 (Appendix C) for statistics. In Lotus Uplifting practitioners lift their legs off the ground while maintaining the classical Lotus formation of the legs. It is interesting to see increased activity only in the left hemisphere, suggesting a lateralization of cognitive functions in this contrast. Moreover, when we compared Lotus Uplifting to Warrior 2 Right (same posture performed on the right side), there was no statistical significance in any channels.

Mountain Pose Start–Extended Hand to Right Big Toe Hold. There was a significant decrease in HbR in Mountain Pose compared to Extended Hand to Right Big Toe Hold in mDLPFC, and some in lFPC. See Table 2, row 3 for HbR visualization, and Table A8 (Appendix C) for statistics. Decreased HbR in mDLPFC indicates neural activity in this region, which indicates increased cognitive load in Mountain Pose compared to Extended Hand to Right Big Toe Hold. This is an unexpected finding since Extended Hand to Right Big Toe Hold is, as previously described, demanding. Since activity is localized to DLPFC, which activates in anticipation of difficult tasks, it may suggest that the practitioner is mentally preparing themselves for the practice by mentally mapping out coming postures. Therefore, we speculate that there is less need for planning ahead when in Extended Hand mDLPFC, and some in lFPC. See Table 2, row 3 for HbR visualization, and Table A8 (Appendix C) for statistics. Decreased HbR in DLPFC indicates neural activity in this region, which indicates increased cognitive load in Mountain Pose compared to Extended Hand to Right Big Toe Hold. This is an unexpected finding since Extended Hand to Right Big Toe Hold is, as previously described, demanding. Since activity is localized to DLPFC, which activates in anticipation of difficult tasks, it may suggest that the practitioner is mentally preparing themselves for the practice by mentally mapping out coming postures. Therefore, we speculate that there is less need for planning ahead when in Extended Hand to Right Big Toe Hold, due to greater need for concentrating on proper performance during the posture.

Table 2. Contrasts postures with middle significant differences. Contrasts are shown as t-tests of Posture 1 minus Posture 2 [1].

Posture 1	Contrast Statistics Visualization Probe Montage	Posture 2
Head to Knee Pose B, sit on right heel.	HbO	Extended Hand to Right Big Toe Hold
		
Lotus Uplifting	HbO	Warrior 2 Left
		
Mountain Pose Start	HbR	Extended Hand to Right Big Toe Hold
		

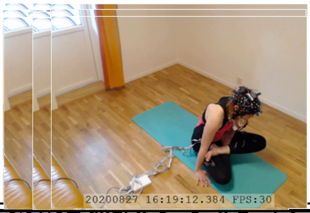
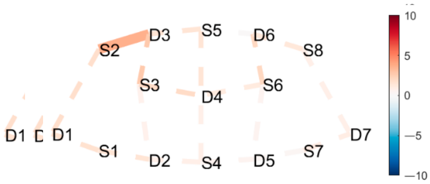
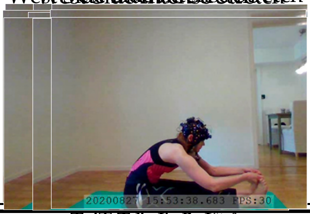
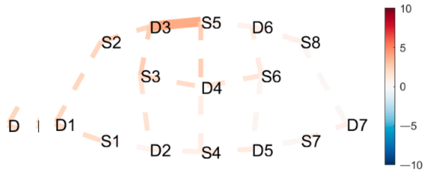
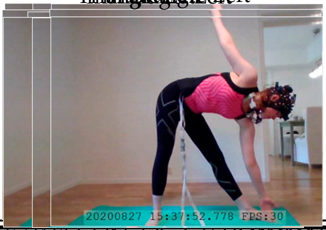
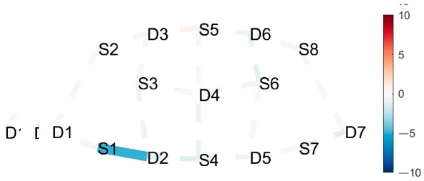
[1] Second-level analysis of HbO and HbR changes for selected contrasts; t-statistic maps with significantly different brain activation (q < 0.05) in solid lines; color bar represents the t-statistic (scaled to range [−10, 10]) with red/blue indicating statistical increase/decrease in HbO concentration between the postures.

4.3. Posture Combinations with Unexpected Little Difference

Lotus Uplifting–Lotus. There was one significant HbO channel for Lotus Uplifting compared to Lotus in lFPC. Lotus Uplifting is scarcely more cognitively demanding than Lotus. See Table 3, row 1 for HbO visualization, and Table A9 (Appendix C) for statistics. We initially thought that Lotus Uplifting might pose a much higher cognitive demand than Lotus due to the strenuous muscle engagement and coordination required to lift crossed legs from the ground. It is, therefore, interesting that we observed a significant difference in only one channel. We speculate that the participant might have been able to direct their attention to something other than the physical demand, but not so much so that they started planning other action sequences and engaging DLPFC. This may be attributed to their breathing pattern and given cues regarding where to direct their focus; however, we still find this interesting.

West Side Intense Stretch A and B–Warrior 2 Right. There was a significant increase in HbO in West Side Intense Stretch A and B compared to Warrior 2 Left, in one channel in m/rFPC. See Table 3, row 2 for HbO visualization, and Table A10 (Appendix C) for statistics. Since Warrior 2 Right is a standing posture requiring balance, hip opening, and a straight back and arms, we initially thought that it would be more demanding than a "simple" seated stretching pose. However, the two variations (A and B) of West Side Intense Stretch had increased HbO in a small region in m/rFPC, and was thus more cognitively demanding than Warrior 2 Left. This may be why its translation from Sanskrit includes the word "intense".

Table 3. Contrasts postures with little significant difference. Contrasts are shown as *t*-tests of Posture 1 minus Posture 2 [1].

Posture 1	Contrast Statistics Visualized on Probe Montage	Posture 2
Lotus Uplifting 	HbO  S2 D3 S5 D6 S8 S3 D4 S6 D1 S1 D2 S4 D5 S7 D7 10 5 0 −5 −10	Lotus
Wide Side Intense Stretch 	HbO  S2 D3 S5 D6 S8 S3 D4 S6 D1 S1 D2 S4 D5 S7 D7 10 5 0 −5 −10	Warrior 2 Right
Triangle Left 	HbO  S2 D3 S5 D6 S8 S3 D4 S6 D1 S1 D2 S4 D5 S7 D7 10 5 0 −5 −10	Triangle Right

[1] red and blue indicating statistical increase/decrease in HbO concentration between the postures.

5. Discussion

Wide Side Intense Stretch A and B–Warrior 2 Right. There was a significant increase in HbO in Wide Side Intense Stretch A and B compared to Warrior 2 Left, in one channel in m/rFPC. See Table 3, row 2 for HbO visualization, and Table A10 (Appendix C) for statistics.

rFPC. Triangle Left imposed a slightly lower cognitive load than Triangle Right. See Table 3, row 3 for HbO visualization, and Table A11 (Appendix C) for statistics. A common statement heard in yoga classes is that

5. Discussion

5.1. Lateralization

An overall observation of contrasts in Table 3 shows that all the lateralization of activation occurs in the left hemisphere. This observation also holds when inspecting Tables 1 and 2. As found by [51], load-dependent HbO activation yielded stronger activation in the left hemisphere in bilateral DLPFC. This may explain why we observed more activation in the left hemisphere, as it is an indication of increasing cognitive load. Moreover, changes in cognitive state changes functional connectivity in adjacent frontal lobe regions (i.e., FPC and DLPC) measured by HbO [51], which supports the notion that HbO changes found in this study stem from changes in cognitive states.

5.2. Reflections on Motion Artifacts

We observed systemic motion artifacts in these data. The practice's transitions between postures are fast paced. This slightly reduces the optodes pressure on the scalp during the movement across all channels causing spikes in fNIRS data. The practice also includes postures with the head positioned upside-down, e.g., Downward Facing Dog, which causes increased blood flow to the head due to gravity and the head's position relative to the heart, which adds a baseline shift to the fNIRS data. Postures with the head upside-down can of course not be compared with most of the other postures in which the head is upright. Therefore, we found it helpful to add information on the head position for each posture during the video coding. We did not see any major problems in the fNIRS data for postures with the head positioned on the side, with different tilts or rotations of the head.

As mentioned by [8], head movement, heartbeat, and respiration artifacts may be corrected with filtering, which helps to reduce noise in fNIRS data. The AR-IRLS filter used to process these data is designed for correcting slippage of optodes, motion, and physiological noise by designing optimal pre-whitening filters using autoregressive models and iteratively reweighted least squares [71]. As described in [71], AR-IRLS removes serially correlated errors, and by doing so reduces the false positive rate to 5–9%, which, compared to 37% of ordinary least squares (OLS) with no motion correction, is impressive. We also tested a Temporal Derivative Distribution Repair (TDDR) [74] motion correction method, which allowed us to visually inspect time-series fNIRS data without the visual clutter of artifacts, which was helpful in distilling the insights above (but we did not use it for our analyses here). As outlined by [74], future work on fNIRS motion correction should include whether combinations of two or more correction methods yield improved performance, given the growing number of fNIRS motion correction methods with different strengths and limitations. We look forward to seeing results from these efforts in the coming years.

Despite excellent motion correction methods, for future studies, we obviously recommend keeping gravity where it usually is, but larger changes in, for example, posture and activity performed by participants can be integrated. FNIRS measurement can distinguish between various stimuli within similar contexts, despite noise from the real-world environment and activity.

6. Conclusions

This study obtained fNIRS brain activity measurements from seven (N = 7) sessions of an Ashtanga Vinyasa Yoga practice conducted in a real-world environment. The results show differences in cognitive load when comparing technically complex postures to relatively simple ones, but also some contrasts with little difference, although a greater difference was initially hypothesized. We now know more about cortical brain activity during a yoga practice. Despite motion artifacts and real-world noise, we can distill cognitive load from applications with considerable motion in the real world, and conclude that it is feasible to obtain neuroimaging measurements in such settings. To the best of our knowledge, this is the first demonstration of fNIRS neuroimaging recorded during any moving yoga practice. It exemplifies that we now have the technologies available for neuroimaging measurements in the real world. This study explores the boundaries of cognitive load measurements in the real world, and contributes to the empirical knowledge base of using fNIRS in realistic settings. Future work with fNIRS should take advantage of this by accomplishing studies with considerable movement in the real world.

Author Contributions: Conceptualization, H.D. and M.S.; methodology, H.D.; software, H.D.; validation, H.D. and M.S.; formal analysis, H.D.; investigation, H.D.; resources, M.S. and H.D.; data curation, H.D.; writing—original draft preparation, H.D.; writing—review and editing, H.D. and M.S.; visualization, H.D.; supervision, M.S.; funding acquisition, M.S. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by strategic funds from the Department of Mechanical and Industrial Engineering (MTP) at NTNU.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Norwegian Centre for Research Data (NSD) (reference: 124330).

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We would like to thank Hendrik Santosa and Theodore J. Huppert for their helpful input to fNIRS analysis.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1 includes the list of the postures that were a part of the practice. Postures were held for 5 breaths unless otherwise stated.

Table A1. Practice sheet/list of postures.

Sanskrit	English	Comment
	Warm-up	
Samasthitih	Mountain Pose	Conducted. Participant listened to chant in video.
Surya Namaskara A x5	Sun Salutation A	Conducted.
Surya Namaskara B x5	Sun Salutation A	Conducted.
	Standing postures	
Padangusthasana	Big Toe Pose	Conducted.
Pada Hastasana	Hand to Foot Pose/hands under feet	Conducted.
Utthita Trikonasana	Triangle	Conducted.
Parivṛtta Trikonasana	Revolved Triangle	Conducted.
Utthita Parsvakonasana	Extended Side Angle	Conducted.
Parivṛtta Parsvakonasana	Revolved Side Angle	Conducted.
Prasarita Padottanasana, A, B, C and D	Wide Leg Forward Fold A, B, C, and D	Conducted, but adapted to accommodate the sensors. The head was not placed on the ground.
Parsvottanasana	Side Intense Stretch	Conducted.
Utthita Hasta Padangusthasana	Extended Hand to Big Toe Pose	Conducted.
Utthita Parsvasahita	Extended Hand to Big Toe Side and Hold Pose	
Ardha Baddha Padmottanasana	Half Bound Lotus Standing Forward Bend	Conducted.

Table A1. *Cont.*

Sanskrit	English	Comment
Utkatasana	Chair Pose	Conducted.
Virabhadrasana I/A	Warrior 1/A	Conducted.
Virabhadrasana II/B	Warrior 2/B	Conducted.
	Seated postures	
Dandasana	Staff Pose	Conducted.
Pascimottanasana A, B	West Intense Stretch	Conducted.
Purvottanasana	East Intense Stretch	Conducted.
Ardha Baddha Padma Pascimottanasana	Half Bound Lotus Forward Fold	Conducted.
Triyang Mukha Eka Pada Pascimottanasana	One Leg Folded Back, Forward Fold	Conducted.
Janu Sirsasana A, B and C	Head to Knee Pose A, B, and C	Conducted.
Maricyasana A, B C and D	Marichi's Pose. A seated pose with twist variations.	Conducted.
Navasana x5	Boat	Conducted.
	The first half of the Primary Series is finished	
Baddha Konasana A and B	Bound Angle Pose A (upright) and B (fold)	Conducted sometimes.
	Finishing sequence	
Urdhva Dhanurasana	Wheel Pose	Not conducted.
Pascimottanasana—10 breaths	Seated Forward Fold/Bend	Conducted.
Salamba Sarvangasana—10 breaths	Shoulderstand	Not conducted.
Halasana	Plow	Not conducted.
Karna Pidasana	Ear Pressure Pose	Not conducted.
Urdhva Padmasana	Upward Lotus Pose	Not conducted.
Pindasana	Embryo	Not conducted.
Matsyasana	Fish Pose	Not conducted.
Uttana Padasana	Raised Leg Pose	Not conducted.
Sirsasana A and B—10 breaths	Headstand A and B	Not conducted.
Balasana	Child's Pose	Conducted. The head was not placed on the ground.
Baddha Padmasana (–10 breaths)	Bound Lotus and Bow	Conducted.
Padmasana	Lotus position	Conducted
Utplutih	Scale Pose/Lotus Uplifting	Conducted
Samasthitih	Mountain Pose	Conducted. Participant listened to chant in video.
Savasana	Corpse Pose	Not conducted.

Appendix B

The labels used for video coding in BORIS can be found in Table A2. These are the postures used in the analysis.

Table A2. Video coded postures/list of postures used in the analysis.

Sanskrit	English
Samasthitih (start of practice)	Mountain_Pose_Start
Surya Namaskar A	Sun_Salutation_A
Surya Namaskar B	Sun_Salutation_B
Adho Mukha Svanasana	Downward_Facing_Dog
Padangusthasana	Big_toe_pose
Pada Hastasana	Hands_under_feet
Utthita Trikonasana (right foot forward)	Trikonasana_Right
Utthita Trikonasana (left foot forward)	Trikonasana_Left
Parivrtta Trikonasana (right foot forward)	Revolved_triangle_Right
Parivrtta Trikonasana (left foot forward)	Revolved_triangle_Left
Utthita Parsvakonasana (right foot forward)	Extended_side_angle_Right
Utthita Parsvakonasana (left foot forward)	Extended_side_angle_Left
Parivṛtta Parsvakonasana (right foot forward)	Revolved_side_angle_Right
Parivṛtta Parsvakonasana (left foot forward)	Revolved_side_angle_Left
Prasarita Padottanasana	Wide_leg_forward_fold
Parsvottanasana (right foot forward)	Side_intense_stretch_Right
Parsvottanasana (left foot forward)	Side_intense_stretch_Left
Utthita Hasta Padangusthasana (right toe)	Extended_hand_to_Right_big_toe
Utthita Parsvasahita (right foot uplifted)	Extended_hand_to_Right_big_toe_hold
Utthita Hasta Padangusthasana (left toe)	Extended_hand_to_Left_big_toe
Utthita Parsvasahita (left foot uplifted)	Extended_hand_to_Left_big_toe_hold
Ardha Baddha Padmottanasana (right foot bound)	Right_foot_in_half_bound_lotus
Ardha Baddha Padmottanasana (left foot bound)	Left_foot_in_half_bound_lotus
Utkatasana	Chair_Pose
Virabhadrasana I (right foot forward)	Warrier_1_Right
Virabhadrasana I (left foot forward)	Warrier_1_Left
Virabhadrasana II (right foot forward)	Warrier_2_Left
Virabhadrasana II (left foot forward)	Warrier_2_Right
Dandasana	Staff_pose
Pascimottanasana A, B	West_intense_stretch_A_B
Purvottanasana	East_intense_stretch
Ardha Baddha Padma Pascimottanasana (right foot)	Right_foot_in_half_bound_lotus_forward_fold
Ardha Baddha Padma Pascimottanasana (left foot)	Left_foot_in_half_bound_lotus_forward_fold
Triyang Mukha Eka Pada Pascimottanasana (right leg)	Right_leg_folded_back_forward_fold
Triyang Mukha Eka Pada Pascimottanasana (left leg)	Left_leg_folded_back_forward_fold
Janu Sirsasana A (right leg)	Head_to_knee_pose_A_Right_Leg_folded
Janu Sirsasana A (left leg)	Head_to_knee_pose_A_Left_Leg_folded
Janu Sirsasana B (right heel)	Head_to_knee_pose_B_sit_on_Right_heel
Janu Sirsasana B (left heel)	Head_to_knee_pose_B_sit_on_Left_heel
Janu Sirsasana C (right toe)	Head_to_knee_pose_C_Right_toe_stretch
Janu Sirsasana C (left toe)	Head_to_knee_pose_C_Left_toe_stretch
Maricyasana A (right knee bent up)	MA_Right_knee_bent_bind

Table A2. *Cont.*

Sanskrit	English
Maricyasana A (left knee bent up)	MA_Left_knee_bent_bind
Maricyasana B (left leg Lotus)	MB_Left_leg_lotus_Right_knee_bent_bind
Maricyasana B (right leg Lotus)	MB_Right_leg_lotus_Left_knee_bent_bind
Maricyasana C (right knee bent up)	MC_Right_knee_bent_twist_bind
Maricyasana C (left knee bent up)	MC_Left_knee_bent_twist_bind
Maricyasana D (left leg Lotus)	MD_Left_leg_lotus_Right_knee_bent_twist_bind
Maricyasana D (right leg Lotus)	MD_Right_leg_lotus_Left_knee_bent_twist_bind
Navasana	Boat
Baddha Konasana	Bound_angle_upright_and_fold
Pascimottanasana	Forward_fold_end
Balasana	Childs_pose
Baddha Padmasana	Bound_lotus_bow
Padmasana	Lotus
Utplutih	Lotus_uplifting
Samasthitih (end of practice)	Mountain_Pose_Stop

Appendix C

Table A3. *t*-test statistics for each channel (source–detector pair). Boat–Mountain Pose Start.

Source	Detector	Type	Beta	Se	Tstat	Dfe	Q	Power
1	2	hbo	197.46	60.05	3.29	321	0.00	0.79
1	2	hbr	95.15	20.27	4.69	321	0.00	0.99
2	3	hbo	313.03	90.63	3.45	321	0.00	0.83
3	3	hbo	313.97	60.74	5.17	321	0.00	1.00
3	3	hbr	49.19	13.82	3.56	321	0.00	0.86
3	4	hbo	253.23	58.59	4.32	321	0.00	0.97
4	2	hbo	192.11	71.50	2.69	321	0.02	0.58
4	2	hbr	153.95	21.13	7.29	321	0.00	1.00
4	4	hbo	337.58	92.19	3.66	321	0.00	0.88
4	4	hbr	94.87	30.08	3.15	321	0.00	0.75
4	5	hbo	205.60	40.26	5.11	321	0.00	1.00
4	5	hbr	99.92	14.15	7.06	321	0.00	1.00
5	3	hbo	485.21	68.56	7.08	321	0.00	1.00
5	3	hbr	87.51	24.00	3.65	321	0.00	0.88
5	4	hbo	322.83	57.88	5.58	321	0.00	1.00
5	6	hbo	691.47	87.80	7.88	321	0.00	1.00
5	6	hbr	58.16	19.36	3.00	321	0.01	0.70
6	4	hbo	179.67	71.04	2.53	321	0.02	0.51
6	6	hbo	272.10	85.00	3.20	321	0.00	0.76
7	5	hbr	115.30	21.91	5.26	321	0.00	1.00
8	6	hbo	225.62	94.04	2.40	321	0.03	0.46

Table A4. *t*-test statistics for each channel (source–detector pair). Mountain Pose Start–MA Right Knee Bent Bind.

Source	Detector	Type	Beta	Se	Tstat	Dfe	Q	Power
2	3	hbo	−346.99	103.30	−3.36	321	0.00	0.81
2	3	hbr	−109.30	26.27	−4.16	321	0.00	0.95
3	3	hbo	−281.68	75.93	−3.71	321	0.00	0.89
3	3	hbr	−49.84	17.72	−2.81	321	0.01	0.63
3	4	hbo	−215.96	69.01	−3.13	321	0.00	0.74
3	4	hbr	−44.20	15.20	−2.91	321	0.01	0.66
4	4	hbo	−354.04	101.43	−3.49	321	0.00	0.84
4	4	hbr	−107.66	37.51	−2.87	321	0.01	0.65
5	3	hbo	−558.02	80.00	−6.98	321	0.00	1.00
5	3	hbr	−197.23	28.89	−6.83	321	0.00	1.00
5	4	hbo	−346.45	69.69	−4.97	321	0.00	0.99
5	4	hbr	−77.77	20.22	−3.85	321	0.00	0.91
5	6	hbo	−818.60	99.86	−8.20	321	0.00	1.00
5	6	hbr	−101.37	22.93	−4.42	321	0.00	0.97
6	4	hbo	−274.98	82.53	−3.33	321	0.00	0.80
6	4	hbr	−81.21	27.58	−2.94	321	0.01	0.67
6	6	hbo	−351.87	104.43	−3.37	321	0.00	0.81
6	6	hbr	−89.55	25.90	−3.46	321	0.00	0.83
7	5	hbr	−61.91	26.04	−2.38	321	0.03	0.45
8	6	hbo	−462.97	105.39	−4.39	321	0.00	0.97
8	6	hbr	−112.37	32.59	−3.45	321	0.00	0.83

Table A5. *t*-test statistics for each channel (source–detector pair). Boat–Lotus.

Source	Detector	Type	Beta	Se	Tstat	Dfe	Q	Power
1	2	hbo	145.86	47.68	3.06	321	0.01	0.71
2	3	hbo	366.15	69.53	5.27	321	0.00	1.00
2	3	hbr	65.47	18.41	3.56	321	0.00	0.86
3	3	hbo	175.83	50.31	3.49	321	0.00	0.84
3	4	hbo	149.83	48.33	3.10	321	0.01	0.73
3	4	hbr	34.79	10.21	3.41	321	0.00	0.82
4	4	hbr	75.27	25.77	2.92	321	0.01	0.67
4	5	hbo	101.00	33.64	3.00	321	0.01	0.69
4	5	hbr	38.79	12.10	3.21	321	0.00	0.76
5	3	hbo	374.76	56.25	6.66	321	0.00	1.00
5	3	hbr	61.96	19.22	3.22	321	0.00	0.77
5	4	hbo	202.77	47.24	4.29	321	0.00	0.96
5	4	hbr	45.67	13.59	3.36	321	0.00	0.81
5	6	hbo	406.25	76.05	5.34	321	0.00	1.00
5	6	hbr	80.07	15.84	5.05	321	0.00	0.99
6	6	hbo	222.71	68.88	3.23	321	0.00	0.77

Table A5. *Cont.*

Source	Detector	Type	Beta	Se	Tstat	Dfe	Q	Power
1	2	hbo	145.86	47.68	3.06	321	0.01	0.71
2	3	hbo	366.15	69.53	5.27	321	0.00	1.00
2	3	hbr	65.47	18.41	3.56	321	0.00	0.86
3	3	hbo	175.83	50.31	3.49	321	0.00	0.84
3	4	hbo	149.83	48.33	3.10	321	0.01	0.73
3	4	hbr	34.79	10.21	3.41	321	0.00	0.82
4	4	hbr	75.27	25.77	2.92	321	0.01	0.67
4	5	hbo	101.00	33.64	3.00	321	0.01	0.69
4	5	hbr	38.79	12.10	3.21	321	0.00	0.76
5	3	hbo	374.76	56.25	6.66	321	0.00	1.00
5	3	hbr	61.96	19.22	3.22	321	0.00	0.77
5	4	hbo	202.77	47.24	4.29	321	0.00	0.96
5	4	hbr	45.67	13.59	3.36	321	0.00	0.81
5	6	hbo	406.25	76.05	5.34	321	0.00	1.00
5	6	hbr	80.07	15.84	5.05	321	0.00	0.99
6	6	hbo	222.71	68.88	3.23	321	0.00	0.77
6	6	hbr	58.18	17.87	3.26	321	0.00	0.78
7	7	hbr	47.33	20.23	2.34	321	0.04	0.44
8	6	hbo	263.54	76.20	3.46	321	0.00	0.83
1	2	hbo	145.86	47.68	3.06	321	0.01	0.71
2	3	hbo	366.15	69.53	5.27	321	0.00	1.00

Table A6. *t*-test statistics for each channel (source–detector pair). Head to Knee Pose B, sit on right heel–Extended Hand to Right Big Toe Hold.

Source	Detector	Type	Beta	Se	Tstat	Dfe	Q	Power
5	3	hbo	453.27	111.13	4.08	321	0.00	0.94
5	3	hbr	116.32	36.45	3.19	321	0.00	0.76
5	4	hbo	256.29	91.15	2.81	321	0.01	0.63
5	6	hbo	568.80	145.42	3.91	321	0.00	0.92
5	6	hbr	160.80	31.36	5.13	321	0.00	1.00
5	3	hbo	453.27	111.13	4.08	321	0.00	0.94

Table A7. *t*-test statistics for each channel (source–detector pair). Warrier 2 Left–Lotus Uplifting.

Source	Detector	Type	Beta	Se	Tstat	Dfe	Q	Power
1	1	hbo	−715.08	219.45	−3.26	321	0.02	0.78
1	2	hbo	−295.74	92.86	−3.18	321	0.02	0.76
2	3	hbo	−414.64	136.13	−3.05	321	0.03	0.71
3	4	hbo	−292.00	85.87	−3.40	321	0.02	0.82
4	5	hbr	65.32	22.85	2.86	321	0.04	0.64
1	1	hbo	−715.08	219.45	−3.26	321	0.02	0.78

Table A8. *t*-test statistics for each channel (source–detector pair). Mountain Pose Start–Extended Hand to Right Big Toe Hold.

Source	Detector	Type	Beta	Se	Tstat	Dfe	Q	Power
1	2	hbr	−86.34	26.26	−3.29	321	0.01	0.79
3	3	hbr	−53.18	18.16	−2.93	321	0.03	0.67
4	2	hbr	−141.34	27.43	−5.15	321	0.00	1.00
4	5	hbr	−88.99	19.47	−4.57	321	0.00	0.98
7	5	hbr	−99.69	27.36	−3.64	321	0.00	0.87
1	2	hbr	−86.34	26.26	−3.29	321	0.01	0.79

Table A9. *t*-test statistics for each channel (source–detector pair). Lotus Uplifting–Lotus.

Source	Detector	Type	Beta	Se	Tstat	Dfe	Q	Power
2	3	hbo	359.06	97.59	3.68	321	0.01	0.88

Table A10. *t*-test statistics for each channel (source–detector pair). West Intense Stretch A and B–Warrior 2 Right.

Source	Detector	Type	Beta	Se	Tstat	Dfe	Q	Power
5	3	hbo	363.78	95.76	3.80	321	0.01	0.90

Table A11. *t*-test statistics for each channel (source–detector pair). Triangle Left–Triangle Right.

Source	Detector	Type	Beta	Se	Tstat	Dfe	Q	Power
1	2	hbo	−393.35	78.26	−5.03	321	0.00	0.99

References

1. Pinti, P.; Tachtsidis, I.; Hamilton, A.; Hirsch, J.; Aichelburg, C.; Gilbert, S.; Burgess, P.W. The Present and Future Use of Functional Near-infrared Spectroscopy (FNIRS) for Cognitive Neuroscience. *Ann. N. Y. Acad. Sci.* **2020**, *1464*, 5–29. [CrossRef] [PubMed]
2. Herold, F.; Wiegel, P.; Scholkmann, F.; Müller, N.G. Applications of Functional Near-Infrared Spectroscopy (FNIRS) Neuroimaging in Exercise–Cognition Science: A Systematic, Methodology-Focused Review. *J. Clin. Med.* **2018**, *7*, 466. [CrossRef]
3. Paszkiel, S.; Szpulak, P. Methods of Acquisition, Archiving and Biomedical Data Analysis of Brain Functioning. In *Proceedings of the Biomedical Engineering and Neuroscience*; Hunek, W.P., Paszkiel, S., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 158–171.
4. Zohdi, H.; Scholkmann, F.; Wolf, U. Individual Differences in Hemodynamic Responses Measured on the Head Due to a Long-Term Stimulation Involving Colored Light Exposure and a Cognitive Task: A SPA-FNIRS Study. *Brain Sci.* **2021**, *11*, 54. [CrossRef]
5. Maior, H.A.; Wilson, M.L.; Sharples, S. Workload Alerts—Using Physiological Measures of Mental Workload to Provide Feedback During Tasks. *ACM Trans. Comput.-Hum. Interact.* **2018**, *25*, 1–30. [CrossRef]
6. Pike, M.F.; Maior, H.A.; Porcheron, M.; Sharples, S.C.; Wilson, M.L. Measuring the Effect of Think Aloud Protocols on Workload Using FNIRS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2014; pp. 3807–3816.
7. Solovey, E.T.; Girouard, A.; Chauncey, K.; Hirshfield, L.M.; Sassaroli, A.; Zheng, F.; Fantini, S.; Jacob, R.J.K. Using FNIRS Brain Sensing in Realistic HCI Settings: Experiments and Guidelines. In Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, Victoria, BC, Canada, 4–7 October 2009; Association for Computing Machinery: New York, NY, USA, 2009; pp. 157–166.
8. Treacy Solovey, E.; Afergan, D.; Peck, E.M.; Hincks, S.W.; Jacob, R.J.K. Designing Implicit Interfaces for Physiological Computing: Guidelines and Lessons Learned Using FNIRS. *ACM Trans. Comput.-Hum. Interact.* **2015**, *21*, 35:1–35:27. [CrossRef]
9. Cui, X.; Bray, S.; Bryant, D.M.; Glover, G.H.; Reiss, A.L. A Quantitative Comparison of NIRS and FMRI across Multiple Cognitive Tasks. *NeuroImage* **2011**, *54*, 2808–2821. [CrossRef] [PubMed]

10. Eggebrecht, A.T.; White, B.R.; Ferradal, S.L.; Chen, C.; Zhan, Y.; Snyder, A.Z.; Dehghani, H.; Culver, J.P. A Quantitative Spatial Comparison of High-Density Diffuse Optical Tomography and FMRI Cortical Mapping. *NeuroImage* **2012**, *61*, 1120–1128. [CrossRef]
11. Huppert, T.J.; Hoge, R.D.; Diamond, S.G.; Franceschini, M.A.; Boas, D.A. A Temporal Comparison of BOLD, ASL, and NIRS Hemodynamic Responses to Motor Stimuli in Adult Humans. *NeuroImage* **2006**, *29*, 368–382. [CrossRef] [PubMed]
12. Balardin, J.B.; Zimeo Morais, G.A.; Furucho, R.A.; Trambaiolli, L.; Vanzella, P.; Biazoli, C.J.; Sato, J.R. Imaging Brain Function with Functional Near-Infrared Spectroscopy in Unconstrained Environments. *Front. Hum. Neurosci.* **2017**, *11*. [CrossRef]
13. Piper, S.K.; Krueger, A.; Koch, S.P.; Mehnert, J.; Habermehl, C.; Steinbrink, J.; Obrig, H.; Schmitz, C.H. A Wearable Multi-Channel FNIRS System for Brain Imaging in Freely Moving Subjects. *Neuroimage* **2014**, *85*. [CrossRef]
14. Pinti, P.; Aichelburg, C.; Lind, F.; Power, S.; Swingler, E.; Merla, A.; Hamilton, A.; Gilbert, S.; Burgess, P.; Tachtsidis, I. Using Fiberless, Wearable FNIRS to Monitor Brain Activity in Real-World Cognitive Tasks. *JoVE (J. Vis. Exp.)* **2015**, e53336. [CrossRef]
15. Baker, J.M.; Rojas-Valverde, D.; Gutiérrez, R.; Winkler, M.; Fuhrimann, S.; Eskenazi, B.; Reiss, A.L.; Mora Ana, M. Portable Functional Neuroimaging as an Environmental Epidemiology Tool: A How-To Guide for the Use of FNIRS in Field Studies. *Environ. Health Perspect.* **2017**, *125*, 094502. [CrossRef] [PubMed]
16. Yoshino, K.; Oka, N.; Yamamoto, K.; Takahashi, H.; Kato, T. Functional Brain Imaging Using Near-Infrared Spectroscopy during Actual Driving on an Expressway. *Front. Hum. Neurosci.* **2013**, *7*. [CrossRef] [PubMed]
17. Sun, P.-P.; Tan, F.-L.; Zhang, Z.; Jiang, Y.-H.; Zhao, Y.; Zhu, C.-Z. Feasibility of Functional Near-Infrared Spectroscopy (FNIRS) to Investigate the Mirror Neuron System: An Experimental Study in a Real-Life Situation. *Front. Hum. Neurosci.* **2018**, *12*. [CrossRef]
18. Nihashi, T.; Ishigaki, T.; Satake, H.; Ito, S.; Kaii, O.; Mori, Y.; Shimamoto, K.; Fukushima, H.; Suzuki, K.; Umakoshi, H.; et al. Monitoring of Fatigue in Radiologists during Prolonged Image Interpretation Using FNIRS. *Jpn. J. Radiol.* **2019**, *37*, 437–448. [CrossRef] [PubMed]
19. Rosenbaum, D.; Leehr, E.J.; Rubel, J.; Maier, M.J.; Pagliaro, V.; Deutsch, K.; Hudak, J.; Metzger, F.G.; Fallgatter, A.J.; Ehlis, A.-C. Cortical Oxygenation during Exposure Therapy – in Situ FNIRS Measurements in Arachnophobia. *NeuroImage Clin.* **2020**, *26*, 102219. [CrossRef]
20. Slutter, M.W.J.; Thammasan, N.; Poel, M. Exploring the Brain Activity Related to Missing Penalty Kicks: An FNIRS Study. *Front. Comput. Sci.* **2021**, *3*. [CrossRef]
21. Okamoto, M.; Dan, H.; Shimizu, K.; Takeo, K.; Amita, T.; Oda, I.; Konishi, I.; Sakamoto, K.; Isobe, S.; Suzuki, T.; et al. Multimodal Assessment of Cortical Activation during Apple Peeling by NIRS and FMRI. *NeuroImage* **2004**, *21*, 1275–1288. [CrossRef]
22. Ayaz, H.; Onaral, B.; Izzetoglu, K.; Shewokis, P.A.; McKendrick, R.; Parasuraman, R. Continuous Monitoring of Brain Dynamics with Functional near Infrared Spectroscopy as a Tool for Neuroergonomic Research: Empirical Examples and a Technological Development. *Front. Hum. Neurosci.* **2013**, *7*. [CrossRef]
23. Causse, M.; Chua, Z.; Peysakhovich, V.; Del Campo, N.; Matton, N. Mental Workload and Neural Efficiency Quantified in the Prefrontal Cortex Using FNIRS. *Sci. Rep.* **2017**, *7*, 5222. [CrossRef]
24. Ahn, S.; Nguyen, T.; Jang, H.; Kim, J.G.; Jun, S.C. Exploring Neuro-Physiological Correlates of Drivers' Mental Fatigue Caused by Sleep Deprivation Using Simultaneous EEG, ECG, and FNIRS Data. *Front. Hum. Neurosci.* **2016**, *10*. [CrossRef]
25. Sibi, S.; Balters, S.; Mok, B.K.; Steinert, M.; Ju, W. Assessing Driver Cortical Activity under Varying Levels of Automation with Functional near Infrared Spectroscopy. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; IEEE: Los Angeles, CA, USA, 2017; pp. 1509–1516.
26. Perpetuini, D.; Chiarelli, A.M.; Cardone, D.; Filippini, C.; Bucco, R.; Zito, M.; Merla, A. Complexity of Frontal Cortex FNIRS Can Support Alzheimer Disease Diagnosis in Memory and Visuo-Spatial Tests. *Entropy* **2019**, *21*, 26. [CrossRef]
27. Lin, C.-C.; Barker, J.W.; Sparto, P.J.; Furman, J.M.; Huppert, T.J. Functional Near-Infrared Spectroscopy (FNIRS) Brain Imaging of Multi-Sensory Integration during Computerized Dynamic Posturography in Middle-Aged and Older Adults. *Exp. Brain Res.* **2017**, *235*, 1247–1256. [CrossRef] [PubMed]
28. de Campos, A.C.; Sukal-Moulton, T.; Huppert, T.; Alter, K.; Damiano, D.L. Brain Activation Patterns Underlying Upper Limb Bilateral Motor Coordination in Unilateral Cerebral Palsy: An FNIRS Study. *Dev. Med. Child Neurol.* **2020**, *62*, 625–632. [CrossRef]
29. Büssing, A.; Michalsen, A.; Khalsa, S.B.S.; Telles, S.; Sherman, K.J. Effects of Yoga on Mental and Physical Health: A Short Summary of Reviews. *Evid.-Based Complement. Altern. Med.* **2012**, *2012*, 165410. [CrossRef]
30. Gard, T.; Noggle, J.J.; Park, C.L.; Vago, D.R.; Wilson, A. Potential Self-Regulatory Mechanisms of Yoga for Psychological Health. *Front. Hum. Neurosci.* **2014**, *8*. [CrossRef] [PubMed]
31. Pascoe, M.C.; Bauer, I.E. A Systematic Review of Randomised Control Trials on the Effects of Yoga on Stress Measures and Mood. *J. Psychiatr. Res.* **2015**, *68*, 270–282. [CrossRef] [PubMed]
32. Telles, S.; Joshi, M.; Dash, M.; Raghuraj, P.; Naveen, K.V.; Nagendra, H.R. An Evaluation of the Ability to Voluntarily Reduce the Heart Rate after a Month of Yoga Practice. *Integr. Psych. Behav.* **2004**, *39*, 119–125. [CrossRef] [PubMed]
33. Cowen, V.S.; Adams, T.B. Heart Rate in Yoga Asana Practice: A Comparison of Styles. *J. Bodyw. Mov. Ther.* **2007**, *11*, 91–95. [CrossRef]
34. Jarry, J.L.; Chang, F.M.; La Civita, L. Ashtanga Yoga for Psychological Well-Being: Initial Effectiveness Study. *Mindfulness* **2017**, *8*, 1269–1279. [CrossRef]
35. Mikkonen, J.; Pedersen, P.; McCarthy, P.W. A Survey of Musculoskeletal Injury among Ashtanga Vinyasa Yoga Practitioners. *Int. J. Yoga Therap.* **2008**, *18*, 59–64. [CrossRef]

36. Sorosky, S.; Stilp, S.; Akuthota, V. Yoga and Pilates in the Management of Low Back Pain. *Curr. Rev. Musculoskelet. Med.* **2008**, *1*, 39–47. [CrossRef]
37. Smith, B.R. Body, Mind and Spirit? Towards an Analysis of the Practice of Yoga. *Body Soc.* **2007**, *13*, 25–46. [CrossRef]
38. Smith, B.R. *Adjusting the Quotidian: Ashtanga Yoga as Everyday Practice*; Citeseer: Princeton, NJ, USA, 2004.
39. Cowen, V.S.; Adams, T.B. Physical and Perceptual Benefits of Yoga Asana Practice: Results of a Pilot Study. *J. Bodyw. Mov. Ther.* **2005**, *9*, 211–219. [CrossRef]
40. Merriam-Webster Omphaloskepsis. Available online: https://www.merriam-webster.com/dictionary/omphaloskepsis (accessed on 10 November 2020).
41. Streeter, C.C.; Gerbarg, P.L.; Saper, R.B.; Ciraulo, D.A.; Brown, R.P. Effects of Yoga on the Autonomic Nervous System, Gamma-Aminobutyric-Acid, and Allostasis in Epilepsy, Depression, and Post-Traumatic Stress Disorder. *Med. Hypotheses* **2012**, *78*, 571–579. [CrossRef]
42. Cho, H.K.; Moon, W.; Kim, J. Effects of Yoga on Stress and Inflammatory Factors in Patients with Chronic Low Back Pain: A Non-Randomized Controlled Study. *Eur. J. Integr. Med.* **2015**, *7*, 118–123. [CrossRef]
43. Fiori, F.; David, N.; Aglioti, S.M. Processing of Proprioceptive and Vestibular Body Signals and Self-Transcendence in Ashtanga Yoga Practitioners. *Front. Hum. Neurosci.* **2014**, *8*. [CrossRef]
44. Benavides, S.; Caballero, J. Ashtanga Yoga for Children and Adolescents for Weight Management and Psychological Well Being: An Uncontrolled Open Pilot Study. *Complement. Ther. Clin. Pract.* **2009**, *15*, 110–114. [CrossRef]
45. van Aalst, J.; Ceccarini, J.; Schramm, G.; Van Weehaeghe, D.; Rezaei, A.; Demyttenaere, K.; Sunaert, S.; Van Laere, K. Long-Term Ashtanga Yoga Practice Decreases Medial Temporal and Brainstem Glucose Metabolism in Relation to Years of Experience. *EJNMMI Res.* **2020**, *10*. [CrossRef] [PubMed]
46. Talwadkar, S.; Jagannathan, A.; Raghuram, N. Effect of Trataka on Cognitive Functions in the Elderly. *Int. J. Yoga* **2014**, *7*, 96. [CrossRef] [PubMed]
47. Vhavle, S.P.; Rao, R.M.; Manjunath, N.K. Comparison of Yoga versus Physical Exercise on Executive Function, Attention, and Working Memory in Adolescent Schoolchildren: A Randomized Controlled Trial. *Int. J. Yoga* **2019**, *12*, 172. [CrossRef]
48. Bhargav, H.; Bn, G.; Raghuram, N.; Hr, N. Frontal Hemodynamic Responses to High Frequency Yoga Breathing in Schizophrenia: A Functional Near-Infrared Spectroscopy Study. *Front. Psychiatry* **2014**, *5*. [CrossRef]
49. Dev, P.; Lancet, R.S.; Saurav, S.; Seshadri, N.P.G.; Kumar Singh, B.; Jha, M. Effect of Yoga on Hemodynamic Changes at Prefrontal Cortex during Sustained Attention Task. In Proceedings of the 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS), Coimbatore, India, 15–16 March 2019; pp. 728–731.
50. Kora, P.; Meenakshi, K.; Swaraja, K.; Rajani, A.; Raju, M.S. EEG Based Interpretation of Human Brain Activity during Yoga and Meditation Using Machine Learning: A Systematic Review. *Complement. Ther. Clin. Pract.* **2021**, *43*, 101329. [CrossRef]
51. Fishburn, F.A.; Norr, M.E.; Medvedev, A.V.; Vaidya, C.J. Sensitivity of FNIRS to Cognitive State and Load. *Front. Hum. Neurosci.* **2014**, *8*. [CrossRef]
52. Miller, E.K.; Freedman, D.J.; Wallis, J.D. The Prefrontal Cortex: Categories, Concepts and Cognition. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2002**, *357*, 1123–1136. [CrossRef]
53. Vanderhasselt, M.-A.; De Raedt, R.; Baeken, C. Dorsolateral Prefrontal Cortex and Stroop Performance: Tackling the Lateralization. *Psychon. Bull. Rev.* **2009**, *16*, 609–612. [CrossRef]
54. Funahashi, S. Working Memory in the Prefrontal Cortex. *Brain Sci.* **2017**, *7*, 49. [CrossRef]
55. MacDonald, A.W.; Cohen, J.D.; Stenger, V.A.; Carter, C.S. Dissociating the Role of the Dorsolateral Prefrontal and Anterior Cingulate Cortex in Cognitive Control. *Science* **2000**, *288*, 1835–1838. [CrossRef]
56. Kaller, C.P.; Rahm, B.; Spreer, J.; Weiller, C.; Unterrainer, J.M. Dissociable Contributions of Left and Right Dorsolateral Prefrontal Cortex in Planning. *Cereb. Cortex* **2011**, *21*, 307–317. [CrossRef]
57. Vassena, E.; Gerrits, R.; Demanet, J.; Verguts, T.; Siugzdaite, R. Anticipation of a Mentally Effortful Task Recruits Dorsolateral Prefrontal Cortex: An FNIRS Validation Study. *Neuropsychologia* **2019**, *123*, 106–115. [CrossRef]
58. Mansouri, F.A.; Koechlin, E.; Rosa, M.G.P.; Buckley, M.J. Managing Competing Goals—A Key Role for the Frontopolar Cortex. *Nat. Rev. Neurosci.* **2017**, *18*, 645–657. [CrossRef]
59. Boschin, E.A.; Piekema, C.; Buckley, M.J. Essential Functions of Primate Frontopolar Cortex in Cognition. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E1020–E1027. [CrossRef]
60. Ashtanga Yoga Full Primary Series with Ty Landrum. Available online: https://www.youtube.com/watch?v=K-s4IIxVBc8&t=2607s (accessed on 10 November 2020).
61. Oostenveld, R.; Praamstra, P. The Five Percent Electrode System for High-Resolution EEG and ERP Measurements. *Clin. Neurophysiol.* **2001**, *112*, 713–719. [CrossRef]
62. Aasted, C.M.; Yücel, M.A.; Cooper, R.J.; Dubb, J.; Tsuzuki, D.; Becerra, L.; Petkov, M.P.; Borsook, D.; Dan, I.; Boas, D.A. Anatomical Guidance for Functional Near-Infrared Spectroscopy: AtlasViewer Tutorial. *Neurophotonics* **2015**, 2. [CrossRef]
63. Fonov, V.; Evans, A.C.; Botteron, K.; Almli, C.R.; McKinstry, R.C.; Collins, D.L. Unbiased Average Age-Appropriate Atlases for Pediatric Studies. *NeuroImage* **2011**, *54*, 313–327. [CrossRef] [PubMed]
64. Fonov, V.; Evans, A.; McKinstry, R.; Almli, C.; Collins, D. Unbiased Nonlinear Average Age-Appropriate Brain Templates from Birth to Adulthood. *NeuroImage* **2009**, *47*, S102. [CrossRef]
65. Santosa, H.; Zhai, X.; Fishburn, F.; Huppert, T. The NIRS Brain AnalyzIR Toolbox. *Algorithms* **2018**, *11*, 73. [CrossRef]

66. NIRStar | FNIRS Systems | NIRS Devices | NIRx. Available online: https://nirx.net/nirstar-1 (accessed on 17 December 2019).
67. LabStreamingLayer. Available online: https://labstreaminglayer.readthedocs.io/index.html (accessed on 28 April 2021).
68. *The IMotions Platform*; iMotions: Copenhagen, Denmark, 2020.
69. Friard, O.; Gamba, M. BORIS: A Free, Versatile Open-Source Event-Logging Software for Video/Audio Coding and Live Observations. *Methods Ecol. Evol.* **2016**, *7*, 1325–1330. [CrossRef]
70. Jacques, S.L. Optical Properties of Biological Tissues: A Review. *Phys. Med. Biol.* **2013**, *58*, R37–R61. [CrossRef]
71. Barker, J.W.; Aarabi, A.; Huppert, T.J. Autoregressive Model Based Algorithm for Correcting Motion and Serially Correlated Errors in FNIRS. *Biomed. Opt. Express* **2013**, *4*, 1366. [CrossRef]
72. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
73. Huppert, T.J. Commentary on the Statistical Properties of Noise and Its Implication on General Linear Models in Functional Near-Infrared Spectroscopy. *Neurophotonics* **2016**, *3*, 010401. [CrossRef]
74. Fishburn, F.A.; Ludlum, R.S.; Vaidya, C.J.; Medvedev, A.V. Temporal Derivative Distribution Repair (TDDR): A Motion Correction Method for FNIRS. *NeuroImage* **2019**, *184*, 171–179. [CrossRef] [PubMed]

Appendix C13: Academic contribution 13

Dybvik, H., Kuster Erichsen, C., Steinert, M. (2021) ‘Description of a Wearable Electroencephalography + Functional Near-Infrared Spectroscopy (EEG+fNIRS) for In-situ Experiments on Design Cognition’, in Proceedings of the International Conference on Engineering Design (ICED21), Gothenburg, Sweden, 16-20 August 2021. https://doi.org/10.1017/pds.2021.94

DESCRIPTION OF A WEARABLE ELECTROENCEPHALOGRAPHY + FUNCTIONAL NEAR-INFRARED SPECTROSCOPY (EEG+FNIRS) FOR IN-SITU EXPERIMENTS ON DESIGN COGNITION

Dybvik, Henrikke;
Kuster Erichsen, Christian;
Steinert, Martin

Norwegian University of Science and Technology

ABSTRACT

We developed a wearable experimental sensor setup featuring multimodal EEG+fNIRS neuroimaging data capture applicable for in situ experiments at a lower financial threshold. Consistent application of a good protocol and procedure for sensor application and signal quality control is crucial for researchers to obtain valid data. This paper provides an exhaustive description of the sensor setup, the data synchronization process, procedure for sensor application, and signal quality control. Potential design cognition experiments with the proposed EEG+fNIRS are also described. In summary, the setup is mobile and provides multimodal neuroimaging data of high quality. We encourage the design community to take advantage of the setup and adapt it to new experimental setups in situ.

Keywords: EEG+fNIRS, Mobile Experiments, Human behaviour in design, Design cognition, Research methodologies and methods

Contact:
Dybvik, Henrikke
Norwegian University of Science and Technology
Department of Mechanical and Industrial Engineering
Norway
henrikke.dybvik@ntnu.no

Cite this article: Dybvik, H., Kuster Erichsen, C., Steinert, M. (2021) 'Description of a Wearable Electroencephalography + Functional Near-Infrared Spectroscopy (EEG+FNIRS) for In-Situ Experiments on Design Cognition', in *Proceedings of the International Conference on Engineering Design (ICED21),* Gothenburg, Sweden, 16-20 August 2021. DOI:10.1017/pds.2021.94

1 INTRODUCTION

Interaction design, human-centred design (HCD) and human-computer interaction (HCI) use experiments for many purposes. To evaluate user interfaces and styles of interaction; understand how people use, experience, perceive and process interactive and increasingly complex technology; and to further understand the underlying mechanisms of human cognition in interaction with objects and technical systems (Balters & Steinert, 2017; Blessing & Chakrabarti, 2009; Cairns & Cox, 2008). Information gathered through neuroimaging techniques and physiology sensors can provide knowledge of mental state and cognition which benefit interface design and product development both in the early design process and at the evaluation stage, for new and existing systems (Balters & Steinert, 2017; Cairns & Cox, 2008; Wulvik et al., 2019). On a methodological level, design research is increasing experiments with physiology and neuroimaging measurements (Balters & Steinert, 2017; Goucher-Lambert et al., 2019; Hay et al., 2020; Steinert & Jablokow, 2013). Such experiments often exist with a trade-off between experimental control and ecological validity (Hay et al., 2020), which is problematic since laboratory settings simply doesn't produce results replicable in the real word (Cairns & Cox, 2008; Okamoto et al., 2004). Highly ecologically valid (aka *in situ*) studies demonstrate how humans appropriate technological solutions in their intended context. Such studies accommodate the often unpredictable, real-world environments in which technology is used (Consolvo et al., 2007). Thus, they are suited to study engineering design solutions and activities (Balters & Steinert, 2017; Hay et al., 2020; Mayseless et al., 2019). Electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) are two portable sensor modalities serving as a substitute for neuroscience gold standard fMRI. EEG (electrical brain activity) and fNIRS (cerebral hemodynamic response) are complementing techniques that don't strictly immobilize participants (Herold et al., 2018; Jacko, 2012; Pinti et al., 2018). Neuroimaging modalities may best be used in combination with other modalities, such as systemic measurements (physiology sensors such as electrocardiography and galvanic skin response) and behavior (eye tracking, motion capture, video recordings) (Pinti et al., 2018; Xu et al., 2019), because it enables data and method triangulation. Triangulation increase validity of results, reduce bias and error, and accommodate individualism (i.e. individual behavior, physiology and psychology) (Balters & Steinert, 2017; Blessing & Chakrabarti, 2009; Steinert & Jablokow, 2013). Multimodal EEG+fNIRS is used in cognitive neuroscience, design research, HCI research, allowing investigation of cognitive states (Ahn et al., 2016; Ahn & Jun, 2017; Hassib et al., 2017; Jacko, 2012; Lukanov et al., 2016; Mayseless et al., 2019; Pinti et al., 2018). EEG+fNIRS are now not only portable, but becoming increasingly wearable (Hassib et al., 2017; Pinti et al., 2018; Xu et al., 2019), thus suited for *in situ* studies. However, portable neuroimaging systems are expensive compared to other portable physiology measures (Cisler et al., 2019). Moreover, development of wearable and wireless technology usually come with the cost of low resolution (Piper et al., 2014) and there are limited options for multimodal data capture systems (Ahn & Jun, 2017). Thus, there is a need for a wearable experimental sensor-setup featuring multimodal neuroimaging with high resolution data capture, applicable for *in situ* studies, at a lower financial threshold.

1.1 The study goal

The goal of this work is to develop a low-cost wearable neuroimaging sensor setup, consisting of concurrent EEG and fNIRS measurements. A wearable fNIRS system was integrated with a low-cost EEG sensor, and a data synchronization process established. This paper contributes an exhaustive description of the sensor setup, and detail the procedure for sensor application, and protocol for signal quality control which is needed to ensure collection of high-quality data.

The second section provides theoretical background on EEG and fNIRS. Then, the functional prototype is presented in the third section. Discussion follows in section five, before the conclusion in section six.

2 BACKGROUND OF EEG AND FNIRS

2.1 Technical principle of EEG

Electroencephalography (EEG) is a neuroimaging tool measuring electrical brain activity, specifically the electric field caused by local current flow induced by neural activation (Malmivuo & Plonsey, 1995). Neural activation generates an electrical potential difference measured by a pair of electrodes mounted on the scalp surface. An EEG signal consists of the sum of neural activity in close vicinity to the electrodes (Balters & Steinert, 2017; Herold et al., 2018). The signal can be recorded inside the skull (*intracranial*), or on the scalp surface (*extracranial*) (Im, 2018; Malmivuo & Plonsey, 1995). Intracranial EEG is not relevant for this work. Signal magnitude is in the range of ±100 μV and the frequency spectrum is conventionally considered to be in the range of 0-100 Hz (Im, 2018; Malmivuo & Plonsey, 1995). EEG signals need to be amplified and converted from analogue to digital before processing and analysis (EEG has low signal-to-noise ratio due to the skulls' inherent low conductivity (Balters & Steinert, 2017; Teplan, 2002). Many different electrodes with different characteristics exists (Lee et al., 2019; Teplan, 2002). The international EEG system provides a standardized electrode positioning system, enabling greater replicability of EEG studies (Oostenveld & Praamstra, 2001).

2.2 Technical principle of fNIRS

Functional Near-Infrared Spectroscopy (fNIRS) is a neuroimaging technique that optically measures the hemodynamic response in brain tissue. Concentration changes in oxygenated and deoxygenated hemoglobin (ΔHbO_2 and ΔHbR respectively) are related to neural activation (Pinti et al., 2018; Xu et al., 2019). Light of different wavelengths in the near-infrared (NIR) spectrum is emitted on the scalp, travels through different layers of the head (e.g., scalp skin, skull, cerebrospinal fluid) before reaching neuronal tissue. Inside the tissue NIR light undergoes absorption and scattering, each contributing to light attenuation. Hemoglobin absorbs NIR light differently depending on oxygen saturation state, which changes light attenuation. Non-absorbed light scatter components are measured and ΔHbO_2 and ΔHbR calculated through the modified Beer-Lambert Law (Herold et al., 2018; Pinti et al., 2018). Two or more optodes, i.e., a minimum of one source and one detector must be used. Active brain regions are generally associated with increased ΔHbO_2 and decreased ΔHbR. The international EEG system (Oostenveld & Praamstra, 2001) is often used for optode placement, but there are works where this is not the case. An interoptode distance of 3 cm is common (Herold et al., 2018).

2.3 Advantages and limitations

EEG gained popularity due to being non-invasive, portable, having high temporal resolution, and being relatively inexpensive compared to other neuroimaging techniques (Balters & Steinert, 2017; Pinti et al., 2018; Solovey et al., 2009). However, EEG limitations include low spatial resolution, proneness to noise from motion, artifacts due to sweat and/or muscle activity (Al-Shargie et al., 2016; Balters & Steinert, 2017; Pinti et al., 2018). fNIRS offers better spatial resolution, is more robust to motion artifacts from head and body, but has lower temporal resolution (Al-Shargie et al., 2016; Herold et al., 2018; Pinti et al., 2018). fNIRS furthermore share several advantages with EEG, notably non-invasiveness, portability, being relatively lightweight, compact, and low-cost. Both are silent in operation, safe for longer term measurement and suitable for multimodal imaging (Al-Shargie et al., 2016; Lee et al., 2019; Pinti et al., 2018; Solovey et al., 2009). fNIRS is limited to cortical layers of the brain (Herold et al., 2018; Jacko, 2012; Pinti et al., 2018). Due to the nature of the hemodynamic response, a ~5 second delay from stimuli onset to peak response is inherent in fNIRS data (Pinti et al., 2018), however, EEG measures electrical responses instantly. fNIRS measurements are sensitive to changes in scalp blood flow unrelated to brain activation, and changes in systemic physiology (e.g. increase in heart rate) (Herold et al., 2018; Jacko, 2012; Pinti et al., 2018), and ambient light can contribute to noisy data (Solovey et al., 2009), but this is not the case for EEG. EEG and fNIRS are complementing techniques and provide much of the information obtainable by neuroscience gold standard fMRI, while rendering strict immobilization of participants unnecessary (Herold et al., 2018; Jacko, 2012).

3 SETUP DESCRIPTION

This section describes the wearable EEG+fNIRS sensor system and include hardware, sensor integration, software for data collection, and process for ensuring high data quality.

3.1 Physical setup

The physical setup consists of a core setup (Fig. 1.) with possibilities for additional physiological measurements and devices.

Core setup:

- Laptop (Dell Latitude 7490)
- fNIRS (NIRSport (NIRSport, 2015), with 8+8 optodes)

- EEG (Open BCI (OpenBCI Inc., 2019), with 8 dry electrodes)
- EEG adapter (Erichsen et al., 2020)
- Flexible fabric cap (128Ch Standard Cap for NIRX, EASYCAP GmbH, Germany)

Additional equipment for a conventional computer setup:

- Monitor, keyboard, mouse, web camera.
- HDMI cable, USB cables, USB port hub

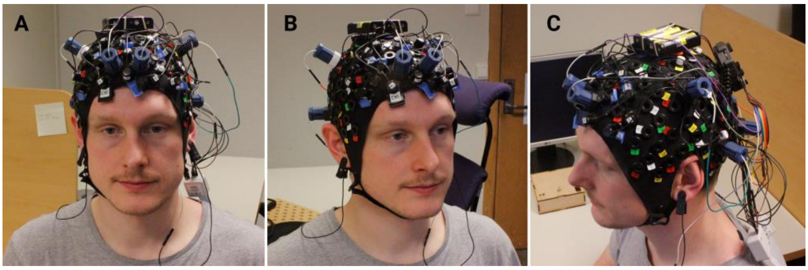


Figure 1. The cap fully integrated with fNIRS and EEG adapters. A) Battery pack placed on the head. B) EEG adapters protrude more than optodes. C) Cables are led toward the back and secured with clips to be of minimal disturbance for the participant. The participant provided written consented to appear in the images published.

3.2 Detailed sensor description of core setup

For collection of fNIRS data the NIRSport system (NIRSport, 2015) with 8 sources and 8 detectors was used with two wavelengths (760 nm and 850 nm) sampled at 7.81 Hz. For collection of EEG data we used the Cyton biosensing board from OpenBCI (OpenBCI Inc., 2019), with 8 spring-loaded, dry electrodes provided as part of the Ultracortex Mark IV EEG headset (Ultracortex "Mark IV" EEG Headset, 2019) sampled at 250 Hz. Both fNIRS optodes and EEG electrodes were mounted on a '128Ch Standard Cap (EASYCAP GmbH, Germany), that follows the five percent system (Oostenveld & Praamstra, 2001). EEG was mounted with the adapter described in the following section (3.3). A dedicated Dell Latitude 7490 laptop was used to collect data. It runs Microsoft Windows 10 Education with a Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz 2.11 GHz processor, 32.0 GB RAM installed, 64-bit operating system, x64-based processor and a 500 GB SSD hard drive.

3.3 Sensor integration in one cap

fNIRS optodes and EEG electrodes were mounted on a standard cap. Initially, this cap only accommodates more expensive EEG electrodes with a geometrical form similar to the fNIRS optodes, and not the selected spring-loaded dry electrodes from OpenBCI. To integrate EEG with fNIRS, we developed a novel adapter that interface between EEG electrodes and the standard cap. The adapter consists of two octagonal parts with a circular, central hole allowing access to the scalp. The adapter is mounted on the cap in the following manner: The bottom part is positioned on the inside of the cap and extends through the precut holes and secured to the top part by means of a threaded connection. The spring-loaded electrode-mounts from OpenBCI are attached to the adapter by a threaded connection, illustrated in Fig. 2. Existing fNIRS optode holders are used to fasten the Cyton board and battery to the participant's head (rendered possible by the small size and low weight (~160 gram) of the battery and board). This imposes fewer restrictions on movement. If the desired optode or electrode configuration conflicts with placement of the battery, the battery can be mounted to the participant's body by elongating the wires from the battery pack. These adapters allow the spring-loaded, dry electrodes to be mounted in any of the 128 standardized positions on the cap that are not occupied by fNIRS optodes. This drastically increases the flexibility and accuracy in electrode placement compared to currently available low-cost EEG equipment e.g., from Emotiv, Neurosky, InteraXon, as well as the standard solutions from OpenBCI. This approach has three advantages. Firstly, the increased flexibility in electrode placement provides more control over which brain regions to study and thus extends possible use cases for OpenBCI hardware. Secondly, accurate electrode positioning is essential in EEG research to ensure replicability and compare results across studies (Oostenveld & Praamstra, 2001). Thus, the approach increases the viability of applying a low-cost

EEG device in scientific research. The complete setup enables low-cost integration of EEG with fNIRS for multimodal brain imaging, lowering the financial barrier to utilize this cutting-edge research method. An exhaustive description of adapter development can be found in Erichsen et al. (2020), and production files at https://github.com/trolllabs/eeg-technology.

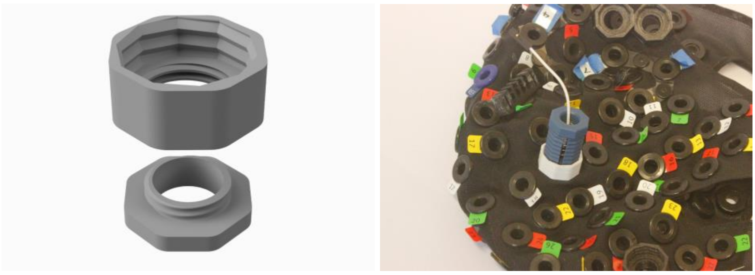


Figure 2. Adapter for EEG-integration. A) Adapter developed to interface the spring-loaded electrodes from OpenBCI with the fNIRS electrode cap. B) Spring loaded electrode mounted on the cap using the adapter.

3.4 Software securing data collection

iMotions 8.1 platform (The IMotions Platform, 2020) present stimuli, collect and synchronize all sensor data. iMotions offer the possibility of registering one or more Lab Streaming Layer (LSL) streams as a sensor. LSL is "*a system for the unified collection of measurement time series in research experiments that handles both the networking, time-synchronization, (near-) real-time access as well as optionally the centralized collection, viewing and disk recording of the data*"[1]. Nirstar 15.2 Acquisition Software for NIRSport (NIRSport, 2015) allows preview, recording and forwarding fNIRS data to third-party software. NIRStar 15.2's LSL option was enabled, and preview mode used to forward a continuous fNIRS data stream through LSL. A python script interfacing with LSL detects the fNIRS data stream and forwards these to iMotions API integration. This enables continuous data collection of all fNIRS channels. For EEG, a separate Python script detects the EEG stream and forwards the data to iMotions using LSL.

3.5 Process for ensuring high signal and data quality

Consistency in experimental procedures ensures that every participant has the same experience, and limits confounding variables and aid experiment robustness (Cairns & Cox, 2008). The current procedure includes greeting participants, acquiring informed consent, a short briefing, placing sensors, ensure connectivity and signal quality, before starting the experiment. Variability in required setup time can occur due to difficulties obtaining good signal quality from both fNIRS and EEG simultaneously. Both optodes and electrodes require direct scalp contact, which is affected by amount and thickness of hair, the ease of which hair can be parted, and general skin conductance. The need to manually disperse hair, apply optically or electrically conductive gels, or spend extensive time organizing wires can significantly impact setup time. The consistent application of a good protocol and procedure for sensor application and signal quality control is crucial for researchers to obtain valid data that can be for inference making. Given this importance the following details steps toward obtaining high signal quality, including an electrode/optode placement procedure and visual inspection prior to starting data collection.

3.5.1 Protocol for ensuring data quality in EEG

Electrode placement procedure. Since fNIRS system current can cause electrical inference and distort EEG-signals (Lee et al., 2019), a wire-configuration that minimize electrical inference is crucial. Several strategies minimize crosstalk; 1) Ensure good electrical contact between electrodes and scalp, since poor electrical contact increase EEGs susceptibility to noise. 2) Maximizing distance between EEG and fNIRS leads. 3) Ensure electric isolation between, and separate ground planes for

[1] https://github.com/sccn/labstreaminglayer/ Accessed 13.04.20

fNIRS and EEG. Obtaining proper spacing can be challenging and the difficulty increase with increasing number of optodes and electrodes. The dry EEG electrode's conductive coating is susceptible to mechanical wear, which limits operational life to 20-30 uses. To limit mechanical wear, electrodes should be mounted after the cap is correctly positioned on the head. Hair in electrode locations should be parted (with a wooden applicator, a Q-tip or similar) to facilitate direct contact with the scalp. Preferably the scalp should be visible through the hole in the cap. The spring-loaded electrodes are then inserted into adapters and fastened by clockwise rotation. Electrodes must be able to move freely in the housing. Excessively fastened electrodes negate the function of the spring loading. Therefore, minimizing hair in areas where electrodes are to be mounted is essential. Applying some conducting gel or water solution to increase the conductivity in hairy regions might be advantageous. However, this can introduce drawbacks associated with wet electrodes such as evaporation of the conducting compound and the need to clean equipment after use. This may also limit the possible duration of the recording.

Visual inspection of signal quality before recording. OpenBCI GUI (open-source software) was used to visually inspect EEG data and signal quality. This was necessary since iMotions is not suited to display real-time EEG-data. OpenBCI GUI allows inspection of data in numerous forms, including a live time-series plot and an FFT-plot indicating the frequency distribution for each channel. Impedance for each channel can also be measured. The FFT-plot should indicate gradually decreasing power towards higher frequencies. fNIRS crosstalk will produce peaks in the FFT-plot at harmonics of the fNIRS sampling rate (7.81 Hz), indicative of excessive noise in the signal, and is to be avoided. Very poor electrical contact between electrodes and scalp is indicated by "Near railed" or "Railed" in the time-series plot.

3.5.2 Protocol for ensuring high data quality in fNIRS

Optode placement procedure. For best signal quality ensure a low-loss contact between skin and optodes. Care should be taken to stabilize optodes on the skin to prevent signal variations and artifacts from motion. The hair should be properly parted, preferably clearly seeing the scalp. A wooden applicator can be used to part hair before placing optodes if needed. If participants have particularly dense hair use of a nonconductive clear gel is recommended, since it improves optical contact and can make parting hair easier. A light source can be helpful in this process. Additional steps to ensure a stable optode can be taken; 1) The use of a stiffening joint increase mechanical rigidity of the montage and facilitate equal optode distance. 2) Fiber optic cables can be spatially organized in a cable tree to avoid undue torque and tipping of optodes. 3) A retaining overcap can be used to exert even pressure on probes and stabilize setup against motion. This overcap also block external light sources from affecting the NIR-light. This was not possible due to the EEG integration, which are larger in size, and the standardized overcap would either distort the signal or destroy the adaptive holder.

Visual inspection of signal quality before recording. NIRStar 15.2's preview mode was used to visually inspect fNIRS data and signal quality. First, a calibration must be performed, and the results analyzed to control signal quality. A Quality Scale tool in NIRStar composes several parameters affecting signal quality: the gain levels, the signal level, the noise level, and a hemodynamic index. A color scale of green, yellow, and red illustrate signal quality collectively. All channels should be green, however yellow may also be accepted. The parameters can be viewed and inspected individually, which should be performed if channels are not acceptable (red or yellow). Gain and intensity of received light have an inverse relationship, if a signal is weak it must be amplified more, which increase noise levels. Noise can be computed as a Coefficient of Variation (CV), indicating variability of the signal with respect to its mean value. Channels with CV>15% will most likely not yield good signal quality. Viewing parameters individually gives a better understanding of why poor signal quality occurs and may help when trying to improve signal quality. High signal quality is generally indicated by visible oscillations of approximately 1 Hz (corresponding to the cardiac response), and overall smooth signal curve. Abrupt spikes do not correspond to the hemodynamic response, they could indicate motion artifacts or noise and should not be visible. Motion artifacts should not be visible if the cap is set up properly, which could be tested by the participant moving their head, in which case the signal should not deteriorate.

4 DISCUSSION AND FUTURE WORK

4.1 Flexibility and robustness

The EEG adapter interface with the spring-loaded electrode mount from OpenBCI, which is compatible with three different dry electrode geometries: flat, spikey, and 5 mm combs. We applied 5mm comb electrodes in hair-covered regions (C3, C4, P7, P8, O1, O2) and spikey electrodes in hairless locations (AFp1 and AFp2). Electrodes can be placed in any position, except for positions occupied by fNIRS. Should different EEG positions be selected or should hair density/conditions vary across participants, electrode geometry can be selected to facilitate scalp contact. We can select different battery packs and different 3D printing filament, which in turn affects both price and robustness. The battery pack can be relocated to a different body part. To reduce electrical inference from fNIRS in EEG signals various shielding measures were tested, but results did not indicate any obvious effects on electrical inference. We found best signal quality when ensuring maximum distance between fNIRS wires and EEG leads. We settled with shielding reference wires and placed fNIRS wires close to the face but will continue testing additional shielding measures to improve setup robustness.

We do not consider data synchronization to be completely robust yet. The iMotions platform do synchronize data, but we have had experiences with occasional unexpected crashes of the platform. Moreover, the platform is demanding for the computer. Our recommendations for those that may be interested in acquiring a similar setup is to investigate open-source software. Our experience with open source software, notably using LSL has been good so far, and we see other studies also using it (Cisler et al., 2019). We will continue developing our software setup by further utilizing LSL, investigating its use in combination with stimuli presentation tools suited to our experiment designs and research questions, and we encourage the community to do the same. Furthermore, not all studies incorporate concurrent physiological measurements disclose exactly how data was synchronized (Liu et al., 2017) and since we are interested in best practice of data synchronization, we highly encourage other studies to disclose their data synchronization process.

4.2 Portability

All individual components are portable. EEG stream data wirelessly over Bluetooth to the laptop. fNIRS requires a cabled connection to the amplifier and can be extended up to 3 m. It can still be applied in experiments involving a moderate amount of movement. The setup can be operated completely on battery power, although only for a limited time. EEG and fNIRS each have dedicated batteries; fNIRS lasts 8 hours; EEG is power by 4 AA alkaline batteries, which lasts at least as long as fNIRS. The longest recording conducted completely on battery power lasted 66 minutes and collected EEG, fNIRS and one web camera recording. This was limited by laptop battery. If the laptop receives socket power, or uses a power bank, this is no longer the limiting factor. The minimum necessary hardware components to conduct an *in situ* study include EEG electrodes and fNIRS optodes, their respective amplifiers and battery, and a laptop. This entire configuration can be put in a normal backpack. The setup has been relocated several times, from offices, to research laboratory and different apartments for testing. Another paper (Dybvik et al., 2021) demonstrate use in a car while driving and a yoga practice, which required transportation to, and set up in a car and a living room. Considering these characteristics in combination with these use cases we pose the setup a viable tool for capturing physiological responses during *in situ* experiments.

4.3 Cost

The costs associated with the EEG integration includes $849 for the hardware supplied by OpenBCI (excluding shipping), and roughly $5 in filament cost for printing adapters, assuming free access to a 3D-printer. The dry electrodes and ear clips must be replaced after 20-30 uses since the conductive coating wear off. Replacement electrodes can be acquired for a cost as low as $0.50/pcs depending on volume. We regard this as low-cost when compared to the price of other wearable EEG systems we're aware of, e.g., $ 26 500 for BIOPAC bioharness and BIOPAC B-Alert System. Since wearable fNIRS systems (with more than 1 channel) amount to € 40 000 or more, cost associated with concurrent EEG+fNIRS quickly aggregates, and we hope the EEG adaptation can be a low-cost option for the community.

4.4 Comparing our setup with existing setups

Current high resolution multimodal EEG+fNIRS systems either require two desktop computers (Ahn et al., 2016; Ahn & Jun, 2017), or two separate instrument systems (Al-Shargie et al., 2016; Shin et al., 2017), each not portable and at a significant cost. Despite fNIRS' and EEGs general increase in portability (Hassib et al., 2017; Xu et al., 2019) and potential for wireless data transmission (Pinti et al., 2018; von Lühmann et al., 2015), wireless fNIRS is not the norm. Moreover, many of the most portable technologies collect data at the lowest resolution possible, which is single channel measurements for both EEG and fNIRS (Xu et al., 2019). This naturally contributes to low weight, less expenses, and easier use *in situ* since preparation time reduces with the number of channels, but this highly limits brain regions and constructs of interest. Researchers have built wireless EEG+fNIRS, developing open source systems with up to 4+2 channel EEG and fNIRS measurements (von Lühmann et al., 2015, 2017). One system support 32 sources and 4 detectors, and up to 16 electrodes, but require use of gel (Safaie et al., 2013). Another portable system support 8 EEG and 32 fNIRS channels, but require a control module to be carried around with the participant (Sawan et al., 2013), similar to our setup. However, we are unable to find participant studies using *any* of the mentioned systems. We speculate system reproduction might be too time-consuming for practical purposes. Our setup resolution is 8 EEG electrodes and 8+8 fNIRS optodes. Regarding data quality, the EEG OpenBCI Cyton board exhibits comparable quality as medical-grade equipment EEG (Frey, 2016), whereas NIRx dedicate their equipment for scientific applications. One limitation of our setup is that we cannot collect concurrent EEG and fNIRS measurements from the same location, but this also applies for other setups.

4.5 Proposed examples of usage

Neuroimaging have been used to investigate designers cognitive style in divergent and convergent design problems (Steinert & Jablokow, 2013), the effectiveness of inspirational stimuli (Goucher-Lambert et al., 2019), and the difference in cognition associated with problem-solving and open-ended design tasks (Vieira, Gero, Delmoral, Gattol, et al., 2020; Vieira, Gero, Delmoral, Li, et al., 2020). We now know that brainstorming and TRIZ evoke different cognitive processes (Shealy et al., 2020), and begin to understand the role of stress during conceptual design (Nguyen & Zeng, 2014). There are numerous opportunities in design cognition, e.g. how different types of memory and knowledge are used in design tasks, the general effects of stimuli on creativity, and how higher-order processing operates in problem-solution coevolution (Hay et al., 2020). We think these topics are interesting and suggest them as potential topics for other researchers to investigate with the proposed EEG+fNIRS setup.

4.6 Further work

Further work includes improving data capture and synchronization, first by further utilizing LSL. We both expect and experience increased noise and motion artifacts *in situ*, and we will investigate means to further minimize noise, by testing and learning from hardware development mentioned (Safaie et al., 2013; Sawan et al., 2013; von Lühmann et al., 2015, 2017). Shielding and wire layout will be improved to minimize crosstalk. *In situ* testing will continue and improvements made to the setup to advance its robustness and to further reduce setup reproduction time.

5 CONCLUSION

This paper presents a wearable experimental sensor setup featuring multimodal EEG+fNIRS neuroimaging data capture applicable for *in situ* experiments. A low-cost EEG was integrated with a wearable fNIRS system, for which we provide an exhaustive description of sensor setup, data synchronization process, procedure for sensor application, and signal quality control. An EEG+fNIRS sensor setup is both applicable in and valuable for e.g., design cognition research. We will continue to develop the setup and conduct experiments. We encourage the community to do the same: take advantage of the setup and adapt it to your *in situ* experiments.

ACKNOWLEDGMENTS

The study was approved by the Norwegian Centre for Research (NSD). This research is supported by strategic funds from the department for Mechanical and Industrial Engineering (MTP) at NTNU.

REFERENCES

Ahn, S., & Jun, S. C. (2017). Multi-Modal Integration of EEG-fNIRS for Brain-Computer Interfaces – Current Limitations and Future Directions. Frontiers in Human Neuroscience, 11. https://doi.org/10.3389/fnhum.2017.00503

Ahn, S., Nguyen, T., Jang, H., Kim, J. G., & Jun, S. C. (2016). Exploring Neuro-Physiological Correlates of Drivers' Mental Fatigue Caused by Sleep Deprivation Using Simultaneous EEG, ECG, and fNIRS Data. Frontiers in Human Neuroscience, 10. https://doi.org/10.3389/fnhum.2016.00219

Al-Shargie, F., Kiguchi, M., Badruddin, N., Dass, S. C., Hani, A. F. M., & Tang, T. B. (2016). Mental stress assessment using simultaneous measurement of EEG and fNIRS. Biomedical Optics Express, 7(10), 3882. https://doi.org/10.1364/BOE.7.003882

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

Blessing, L. T., & Chakrabarti, A. (2009). DRM, a design research methodology. Springer Science & Business Media.

Cairns, P. E., & Cox, A. L. (2008). Research methods for human-computer interaction. Cambridge University Press.

Cisler, D., Greenwood, P. M., Roberts, D. M., McKendrick, R., & Baldwin, C. L. (2019). Comparing the Relative Strengths of EEG and Low-Cost Physiological Devices in Modeling Attention Allocation in Semiautonomous Vehicles. Frontiers in Human Neuroscience, 13. https://doi.org/10.3389/fnhum.2019.00109

Consolvo, S., Harrison, B., Smith, I., Chen, M. Y., Everitt, K., Froehlich, J., & Landay, J. A. (2007). Conducting In Situ Evaluations for and With Ubiquitous Computing Technologies. International Journal of Human–Computer Interaction, 22(1–2), 103–118. https://doi.org/10.1080/10447310709336957

Dybvik, H., Erichsen, C. K., & Steinert, M. (2021). Demonstrating the feasibility of multimodal neuroimaging data capture with a wearable electroencephalography + functional near-infrared spectroscopy (EEG+FNIRS) IN SITU. Proceedings of the Design Society: International Conference on Engineering Design.

Erichsen, C. K., Dybvik, H., & Steinert, M. (2020). Integration of low-cost, dry-comb EEG-electrodes with a standard electrode cap for multimodal signal acquisition during human experiments. DS 101: Proceedings of NordDesign 2020, Lyngby, Denmark, 12th - 14th August 2020, 1–12. https://doi.org/10.35199/NORDDESIGN2020.19

Frey, J. (2016, May 30). Comparison of a consumer grade EEG amplifier with medical grade equipment in BCI applications. International BCI meeting. https://hal.inria.fr/hal-01278245

Goucher-Lambert, K., Moss, J., & Cagan, J. (2019). A neuroimaging investigation of design ideation with and without inspirational stimuli—Understanding the meaning of near and far stimuli. Design Studies, 60, 1–38. https://doi.org/10.1016/j.destud.2018.07.001

Hassib, M., Schneegass, S., Eiglsperger, P., Henze, N., Schmidt, A., & Alt, F. (2017). EngageMeter: A System for Implicit Audience Engagement Sensing Using Electroencephalography. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 5114–5119. https://doi.org/10.1145/3025453.3025669

Hay, L., Cash, P., & McKilligan, S. (2020). The future of design cognition analysis. Design Science, 6. https://doi.org/10.1017/dsj.2020.20

Herold, F., Wiegel, P., Scholkmann, F., & Müller, N. G. (2018). Applications of Functional Near-Infrared Spectroscopy (fNIRS) Neuroimaging in Exercise–Cognition Science: A Systematic, Methodology-Focused Review. Journal of Clinical Medicine, 7(12), 466. https://doi.org/10.3390/jcm7120466

Im, C.-H. (Ed.). (2018). Computational EEG Analysis: Methods and Applications. Springer Singapore. https://doi.org/10.1007/978-981-13-0908-3

Jacko, J. A. (2012). The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications (3rd ed.). CRC Press.

Lee, S., Shin, Y., Kumar, A. R., Kim, M., & Lee, H. (2019). Dry Electrode-Based Fully Isolated EEG/fNIRS Hybrid Brain-Monitoring System. IEEE Transactions on Biomedical Engineering, 66, 1055–1068. https://doi.org/10.1109/tbme.2018.2866550

Liu, Y., Ayaz, H., & Shewokis, P. A. (2017). Multisubject "Learning" for Mental Workload Classification Using Concurrent EEG, fNIRS, and Physiological Measures. Frontiers in Human Neuroscience, 11. https://doi.org/10.3389/fnhum.2017.00389

Lukanov, K., Maior, H. A., & Wilson, M. L. (2016). Using fNIRS in Usability Testing: Understanding the Effect of Web Form Layout on Mental Workload. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 4011–4016. https://doi.org/10.1145/2858036.2858236

Malmivuo, J., & Plonsey, R. (1995). Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields. Oxford University Press. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195058239.001.0001/acprof-9780195058239

Mayseless, N., Hawthorne, G., & Reiss, A. L. (2019). Real-life creative problem solving in teams: FNIRS based hyperscanning study. NeuroImage, 203, 116161. https://doi.org/10.1016/j.neuroimage.2019.116161

Nguyen, T. A., & Zeng, Y. (2014). A physiological study of relationship between designer's mental effort and mental stress during conceptual design. Computer-Aided Design, 54, 3–18. https://doi.org/10.1016/j.cad.2013.10.002

NIRSport. (2015). NIRx Medical Technologies, LLC. https://nirx.net/

Okamoto, M., Dan, H., Shimizu, K., Takeo, K., Amita, T., Oda, I., Konishi, I., Sakamoto, K., Isobe, S., Suzuki, T., Kohyama, K., & Dan, I. (2004). Multimodal assessment of cortical activation during apple peeling by NIRS and fMRI. NeuroImage, 21(4), 1275–1288. https://doi.org/10.1016/j.neuroimage.2003.12.003

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. Clinical Neurophysiology, 112(4), 713–719. https://doi.org/10.1016/S1388-2457(00)00527-7

OpenBCI Inc. (2019). OpenBCI Inc. https://openbci.com/

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2018). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. Annals of the New York Academy of Sciences. http://dx.doi.org/10.1111/nyas.13948

Piper, S. K., Krueger, A., Koch, S. P., Mehnert, J., Habermehl, C., Steinbrink, J., Obrig, H., & Schmitz, C. H. (2014). A Wearable Multi-Channel fNIRS System for Brain Imaging in Freely Moving Subjects. NeuroImage, 85(0 1). https://doi.org/10.1016/j.neuroimage.2013.06.062

Safaie, J., Grebe, R., Moghaddam, H. A., & Wallois, F. (2013). Toward a fully integrated wireless wearable EEG-NIRS bimodal acquisition system. Journal of Neural Engineering, 10(5), 056001. https://doi.org/10.1088/1741-2560/10/5/056001

Sawan, M., Salam, M. T., Le Lan, J., Kassab, A., Gélinas, S., Vannasing, P., Lesage, F., Lassonde, M., & Nguyen, D. K. (2013). Wireless Recording Systems: From Noninvasive EEG-NIRS to Invasive EEG Devices. IEEE Transactions on Biomedical Circuits and Systems, 7(2), 186–195. https://doi.org/10.1109/TBCAS.2013.2255595

Shealy, T., Gero, J., Hu, M., & Milovanovic, J. (2020). Concept generation techniques change patterns of brain activation during engineering design. Design Science, 6. https://doi.org/10.1017/dsj.2020.30

Shin, J., von Lühmann, A., Blankertz, B., Kim, D.-W., Jeong, J., Hwang, H.-J., & Müller, K.-R. (2017). Open Access Dataset for EEG+NIRS Single-Trial Classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25(10), 1735–1745. https://doi.org/10.1109/TNSRE.2016.2628057

Solovey, E. T., Girouard, A., Chauncey, K., Hirshfield, L. M., Sassaroli, A., Zheng, F., Fantini, S., & Jacob, R. J. K. (2009). Using fNIRS brain sensing in realistic HCI settings: Experiments and guidelines. 10.

Steinert, M., & Jablokow, K. (2013). Triangulating front end engineering design activities with physiology data and psychological preferences. 109–118.

Teplan, M. (2002). Fundamentals of EEG Measurement. Measurement Science Review, 2, 11.

The iMotions Platform (8.1). (2020). [Computer software]. iMotions. https://imotions.com/

Ultracortex "Mark IV" EEG Headset. (2019). OpenBCI Inc. https://shop.openbci.com/products/ultracortex-mark-iv

Vieira, S., Gero, J. S., Delmoral, J., Gattol, V., Fernandes, C., Parente, M., & Fernandes, A. A. (2020). The neurophysiological activations of mechanical engineers and industrial designers while designing and problem-solving. Design Science, 6. https://doi.org/10.1017/dsj.2020.26

Vieira, S., Gero, J. S., Delmoral, J., Li, S., Cascini, G., & Fernandes, A. (2020). Brain activity in constrained and open design spaces: An EEG study. Proceedings of the Sixth International Conference on Design Creativity (ICDC 2020), 068–075.

von Lühmann, A., Herff, C., Heger, D., & Schultz, T. (2015). Toward a Wireless Open Source Instrument: Functional Near-infrared Spectroscopy in Mobile Neuroergonomics and BCI Applications. Frontiers in Human Neuroscience, 9. https://doi.org/10.3389/fnhum.2015.00617

von Lühmann, A., Wabnitz, H., Sander, T., & Müller, K.-R. (2017). M3BA: A Mobile, Modular, Multimodal Biosignal Acquisition Architecture for Miniaturized EEG-NIRS-Based Hybrid BCI and Monitoring. IEEE Transactions on Biomedical Engineering, 64(6), 1199–1210. https://doi.org/10.1109/TBME.2016.2594127

Wulvik, A. S., Dybvik, H., & Steinert, M. (2019). Investigating the relationship between mental state (workload and affect) and physiology in a control room setting (ship bridge simulator). Cognition, Technology & Work. https://doi.org/10.1007/s10111-019-00553-8

Xu, J., Slagle, J. M., Banerjee, A., Bracken, B., & Weinger, M. B. (2019). Use of a Portable Functional Near-Infrared Spectroscopy (fNIRS) System to Examine Team Experience During Crisis Event Management in Clinical Simulations. Frontiers in Human Neuroscience, 13. https://doi.org/10.3389/fnhum.2019.00085

Appendix C14: Academic contribution 14

Dybvik, H., Erichsen, C. K., Steinert, M. (2021) ‘Demonstrating the Feasibility of Multimodal Neuroimaging Data Capture with a Wearable Electoencephalography + Functional Near-Infrared Spectroscopy (EEG+fNIRS) in-situ’, in Proceedings of the International Conference on Engineering Design (ICED21), Gothenburg, Sweden, 16-20 August 2021. https://doi.org/10.1017/pds.2021.90

C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16

DEMONSTRATING THE FEASIBILITY OF MULTIMODAL NEUROIMAGING DATA CAPTURE WITH A WEARABLE ELECTOENCEPHALOGRAPHY + FUNCTIONAL NEAR-INFRARED SPECTROSCOPY (EEG+FNIRS) IN SITU

Dybvik, Henrikke;
Erichsen, Christian Kuster;
Steinert, Martin

Norwegian University of Science and Technology

ABSTRACT

We developed a wearable experimental sensor setup featuring multimodal EEG+fNIRS neuroimaging applicable for in situ experiments of human behavior in interaction with technology. A low-cost electroencephalography (EEG) was integrated with a wearable functional Near-Infrared Spectroscopy (fNIRS) system, which we present in two parts. Paper A provide an exhaustive description of setup infrastructure, data synchronization process, a procedure for usage, including sensor application, and ensuring high signal quality. This paper (Paper B) demonstrate the setup's usability in three distinct use cases: a conventional human-computer interaction experiment, an in situ driving experiment where participants drive a car in the city and on the highway, and an ashtanga vinyasa yoga practice in situ. Data on cognitive load from highly ecologically valid experimental setups are presented, and we discuss lessons learned. These include acceptable and unacceptable artefacts, data quality, and constructs possible to investigate with the setup.

Keywords: EEG+fNIRS, in situ experiments, Human behaviour in design, User centred design, Research methodologies and methods

Contact:
Dybvik, Henrikke
Norwegian University of Science and Technology
Department of Mechanical and Industrial Engineering
Norway
henrikke.dybvik@ntnu.no

Cite this article: Dybvik, H., Erichsen, C. K., Steinert, M. (2021) ‘Demonstrating the Feasibility of Multimodal Neuroimaging Data Capture with a Wearable Electoencephalography + Functional Near-Infrared Spectroscopy (EEG+FNIRS) in Situ’, in *Proceedings of the International Conference on Engineering Design (ICED21),* Gothenburg, Sweden, 16-20 August 2021. DOI:10.1017/pds.2021.90

1 INTRODUCTION

Engineering design research is trending towards increasing experiments that integrate physiology and neuroimaging measurements (Balters & Steinert, 2017; Goucher-Lambert et al., 2019; Hay et al., 2020; Steinert & Jablokow, 2013). Such experiments often exist with a trade-off between experimental control and ecological validity (Hay et al., 2020), which is problematic since laboratory settings simply cannot provide results replicable in the real word (Cairns & Cox, 2008; Okamoto et al., 2004). Highly ecologically valid (aka *in situ*) studies demonstrate how humans appropriate technological solutions in their intended context, accommodating the often unpredictable, real-world environments in which technology is used (Consolvo et al., 2007). Thus, they are suited for design research (Balters & Steinert, 2017; Hay et al., 2020; Mayseless et al., 2019). Electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) are two portable and complementing neuroimaging modalities serving as a substitute for neuroscience gold standard fMRI, by measuring electrical brain activity and cerebral hemodynamic response respectively (Herold et al., 2018; Jacko, 2012; Pinti et al., 2018). However, portable neuroimaging systems are expensive, often come at the cost of low resolution and limited options (Ahn & Jun, 2017; Cisler et al., 2019; Piper et al., 2014). Moreover, most neuroimaging studies are often confined to the comfort of a laboratory or an educational setting, and not *in situ* (Gero & Milovanovic, 2020; Goucher-Lambert et al., 2019; Hay et al., 2020; Steinert & Jablokow, 2013). Thus, there is a need for a portable experimental sensor-setup featuring multimodal neuroimaging data capture at a lower financial threshold, and there is a need to demonstrate its feasibility *in situ*.

1.1 The study goal

This work contributes by demonstrating the feasibility of concurrent EEG and fNIRS measurements *in situ*. We increase ecological validity within these demonstrations by testing real-world experimental use cases of increasing complexity. This paper demonstrates three use cases: a conventional human-computer interaction experiment, a driving experiment *in situ* involving city and highway driving, and a moving yoga practice *in situ*. Together with concreate design experiment examples this exemplification expands the range of what is possible within a design research experiment.

The paper briefly describes sensor placement, pre-processing and analysis of EEG and fNIRS, before demonstrating the cases in section 3. Here we provide a description of experimental procedure, results and lessons learned for each case. Section 4 provide examples of potential design experiments. Concluding remarks follow.

2 SENSOR PLACEMENT, DATA PROCESSING AND ANALYSIS

Sensor placement. We used a wearable EEG+fNIRS (Dybvik et al., Under Review). FNIRS optodes were placed over the prefrontal cortex (montage by NIRx Medical Technologies, LLC) using 8 sources and 7 detectors. The sources (denoted Sx) were placed as follows: S1: F3, S2: AF7, S3: AF3, S4: Fz, S5: Fpz, S6: AF4, S7: F4, and S8: AF8. The detectors (denoted Dx) were placed as follows D1: F5, D2: F1, D3: Fp1, D4: AFz, D5: F2, D6: Fp2, and D7: F6. This configuration gives 40 channels. EEG electrodes were placed as follows: AFp1, AFp2, C3, C4, P7, P8, O1 and O2, and one reference electrode placed on each earlobe. Optodes and electrodes were positioned according to the five percent system (Oostenveld & Praamstra, 2001).

Data processing and analysis. For EEG, an initial power spectral density analysis within the frequency bands delta (1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-100 Hz) was performed, a common approach for cognitive load investigations (Borghini et al., 2014). Pre-processing included bandpass filtering (1 - 50 Hz), and a notch filter (48 - 52 Hz) to attenuate line noise, using a Hamming windowed sinc FIR filter. Bad channels and artifacts were removed by visual inspection. Power frequency density was calculated at 1 Hz using Welch's power spectral density estimate. Results were converted to μV^2/Hz, calculating total band power and relative band power. For fNIRS, pre-processing included conversion from raw data to optical density, then conversion to hemoglobin concentration using the modified Beer-Lambert Law. Then a general linear model regression, using an autoregressive pre-whitening method with iteratively reweighted least-squares (AR-IRLS) (Barker et al., 2013; Santosa et al., 2018), was performed for first-level statistics. This procedure is more robust to effects of physiology and motion compared to prior recommendations (e.g. bandpass filtering, motion artifact correction, cardiac response removal etc.), and is therefore now recommended (Santosa et al., 2018). EEGLAB (Delorme & Makeig, 2004) and NIRS toolbox (Santosa et al., 2018) in MATLAB was used.

3 DEMONSTRATION CASES

This section presents concrete test cases selected to demonstrate flexibility in our multimodal EEG+fNIRS sensor setup, through a variety of experimental scenarios. These include a conventional human-computer interaction setup, *in situ* city and highway driving, and *in situ* ashtanga vinyasa yoga practice. We describe each respective case, show resulting data, and detail lessons learned. Procedure for sensor application is described in-depth elsewhere (Dybvik et al., Under Review). Collected data was processed according to the pipeline described in section 2.

3.1 Case 1: Conventional laptop setup

3.1.1 Description of conventional human-computer interaction setup

A conventional computer setup was used as a pilot experiment to test signal quality and analyze data. Participants were tasked with playing a computer game at different difficulties and responding to random auditory alarms using an Arduino device. Computer game difficulty was set to *easy* in the first condition, and *hard* in the second condition. The participant was first exposed to a baseline condition, sitting and relaxing in front of a static screen with instructions to relax for two minutes. An assimilation period followed, before exposure to each condition for four minutes, with a two-minute resting period between conditions. This experiment was set up in a standard office cleared for everything except desks, chairs, and a wall separating participant and experimenter, see Fig. 1a.

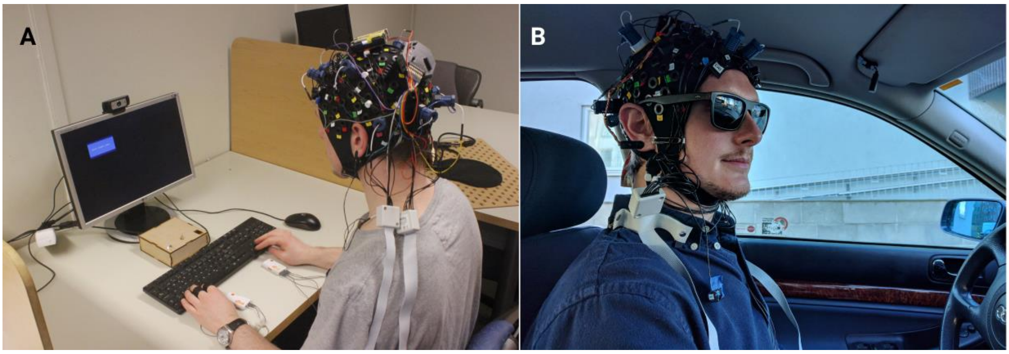


Figure 1. Full sensor setup in A) a conventional laptop configuration, and B) in situ driving.

3.1.2 Results from preliminary data analysis

Result of preliminary analysis EEG. The power spectral density (PSD) plots indicated decent signal quality during *baseline*, and *easy* condition (Fig. 2a. and 2b). The PSD plot for condition *hard* indicated a moderate amount of inference from the fNIRS, expressed by regular peaks at harmonics of the fNIRS sampling rate (7.8 Hz) (Fig. 2c). Thus, this data was not included in further analysis. According to literature, increasing cognitive workload correlates with decreased power in the alpha band at central, and posterior locations (Antonenko et al., 2010; Klimesch, 1999; Sterman et al., 1993), and an increase in the theta band at frontal locations (Borghini et al., 2014). The results obtained during this pilot experiment indicate the same trends, showing increased theta power at location AFp2, and alpha-suppression in central, and posterior locations C3, C4, P7, P8, O1, and O2 (see Fig. 3). This indicates that playing the computer game induce higher cognitive load than a baseline condition, as expected.

Result of preliminary analysis fNIRS. Three representative data slices for each condition were selected for illustration in Fig. 4, which plots concentration changes in oxygenated hemoglobin, ΔHbO_2, from all channels. Differing mental demands can be seen in the graph as changes along the y-axis, which allows for visual observation of changes in cognitive load. The scale of the y-axis depends on participant, context and fNIRS calibration, and is not absolute. Fig. 4 clearly shows the cardiac response as periodic peaks at ~ 1 Hz throughout the data, and we observe an overall trending line over time and condition. Apart from a couple of spikes from slight motion artifacts (at 13 and 15s) this selection is an example of data of good quality. The general trend for this participant is an increased cognitive load during condition *hard* and a lower demand during condition *easy*, although not as low as the baseline. We clearly see a difference in neural activation in the three conditions. The participant is working harder mentally during the game with high difficulty.

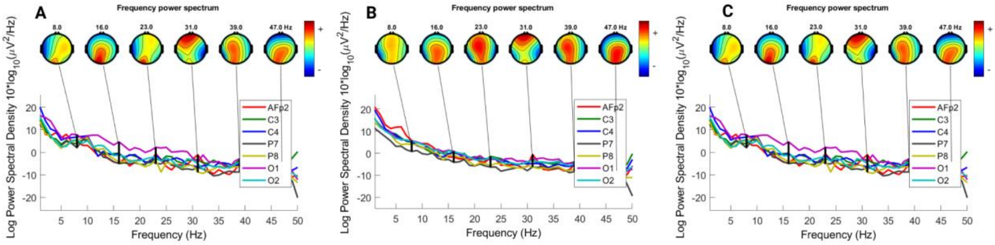


Figure 2. Power spectral density plots for the baseline (A), easy (B), and hard condition (C).

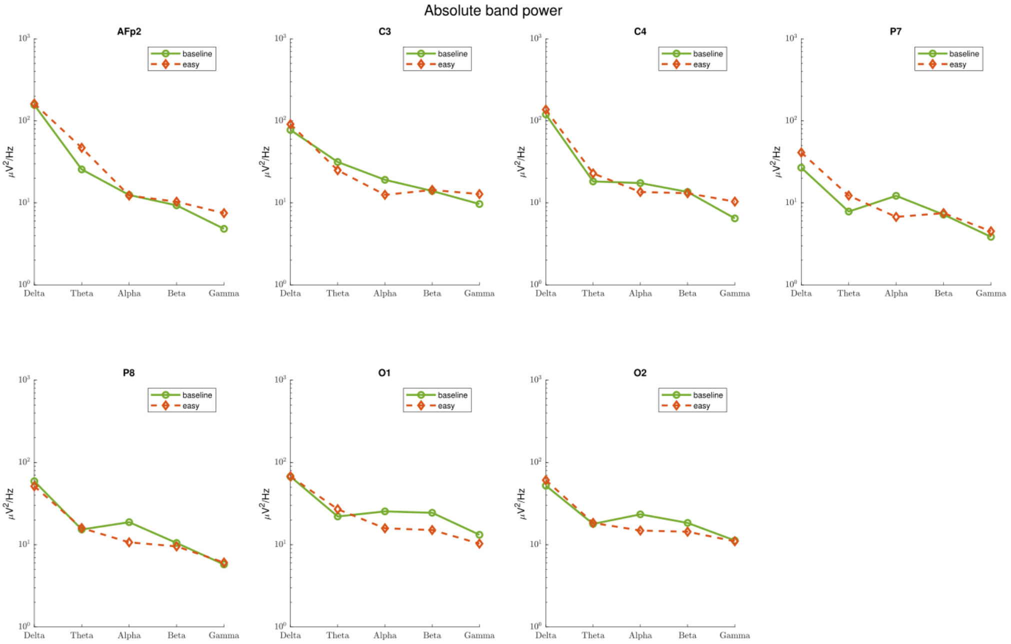


Figure 3. Comparison of spectral band power during baseline and easy condition. One channel was discarded in pre-processing.

3.1.3 Lessons learned - Identifying bad data

Equally important as identifying high quality data is the ability to determine what poor data looks.

Identifying bad EEG data. Crosstalk with fNIRS will exhibit peaks in the FFT-plot at harmonics of the fNIRS sampling rate, which can be seen in Fig. 2c, but also in OpenBCI GUI during visual inspection prior to recording. Furthermore, high signal amplitudes (>50 µV) might indicate presence of excessive noise in the signal. Very poor electrical contact between electrodes and scalp is indicated by "Near railed" or "Railed" in the time-series plot.

Identifying bad fNIRS data. Bad channels are indicated as red in NIRStar 15.2's Quality Scale tool. If such data is recorded, large, inconsistent signal fluctuations (spikes) in some channels dominate the data as seen in Fig 4b, and they will be discarded in an analysis. Close investigation of Fig. 4b reveal good channels fluctuating around zero, with a visible cardiac response.

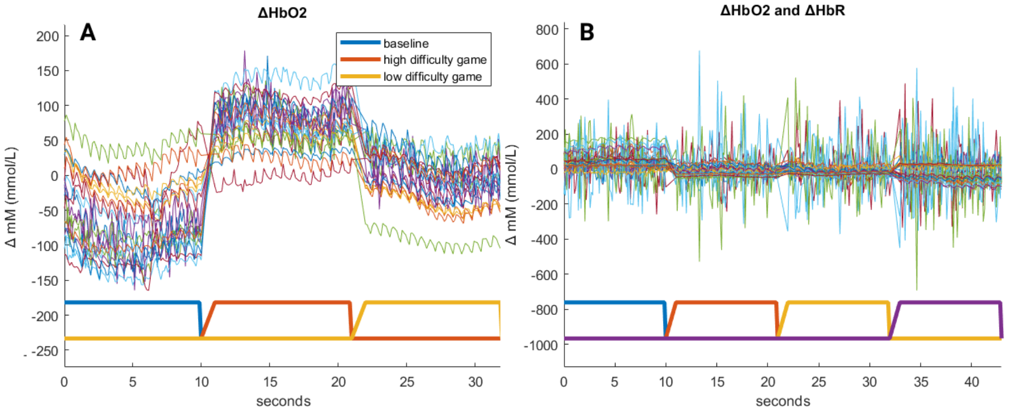


Figure 4. A) ΔHbO$_2$ over time show increased neural activating during high difficulty. B) fNIRS data with insufficient quality.

3.2 Case 2: In situ city and highway driving

3.2.1 Description of an in situ driving experiment

A pilot experiment was conducted to investigate brain activity while driving a car to test the setup's feasibility in *in situ* environments, and test the ability to differentiate between various levels of cognitive load. This pilot experiment was performed on two separate days, with two participants (one male, one female). Fig. 1b depicts one participant wearing the sensors. Additionally, a video was recorded simultaneously using a web camera. A two-minute baseline was recorded while the participant relaxed inside the parked car. Afterwards, participants were instructed to drive around the city center following vocal directions from the experimenter in the front passenger seat. After 20 minutes, participants were instructed to take an exit to the highway. A 20-minute recording of highway driving was made after crossing city border. We continued data collection afterwards to test the setup's battery life, to benchmark for in situ experiments without access to external power sources. We managed a total of 66 minutes of recorded data, with the laptop battery being the limiting factor. Prior to the experiment we hypothesized higher cognitive load associated with city driving than highway driving due to a more complex environment, and increased handling of the car.

3.2.2 Results from preliminary data analysis

Result of preliminary analysis EEG. For visual artifact removal, a five-minute data slice starting at the tenth minute of *city* and *highway*, was extracted. The full two minutes of baseline was processed. During visual inspection, 54% of *city*, 44% of *highway*, and 25% of *baseline* data were rejected. Thus, analysis is based on 106 s. of baseline data, 137 s. of city data, and 169 s. of highway data. Increased cognitive load has been associated with increased relative theta-power at frontal regions (Borghini et al., 2014) and decreased alpha-power at parietal and occipital regions (Borghini et al., 2014). The results (Fig. 5) show a relative increase in theta-power in the frontal regions (AFp1, AFp2), in accordance with hypothesized cognitive load: *baseline* < *highway* < *city*. Additionally, we observe an inverse trend within the alpha band at central and posterior locations (C3, C4, P8, O1, O2), i.e. decreasing alpha-power with increasing cognitive load. There is a substantial difference in relative alpha power between baseline and the two driving conditions. The difference between the two driving conditions is smaller, but the general trend of lower alpha power with hypothesized increased cognitive load holds true for all central and posterior locations, except for C4. Here, relative alpha-power is marginally higher during city driving compared to highway driving. Still, overall results indicate highest cognitive load imposed by *city*, followed by *highway* driving, and *baseline* condition.

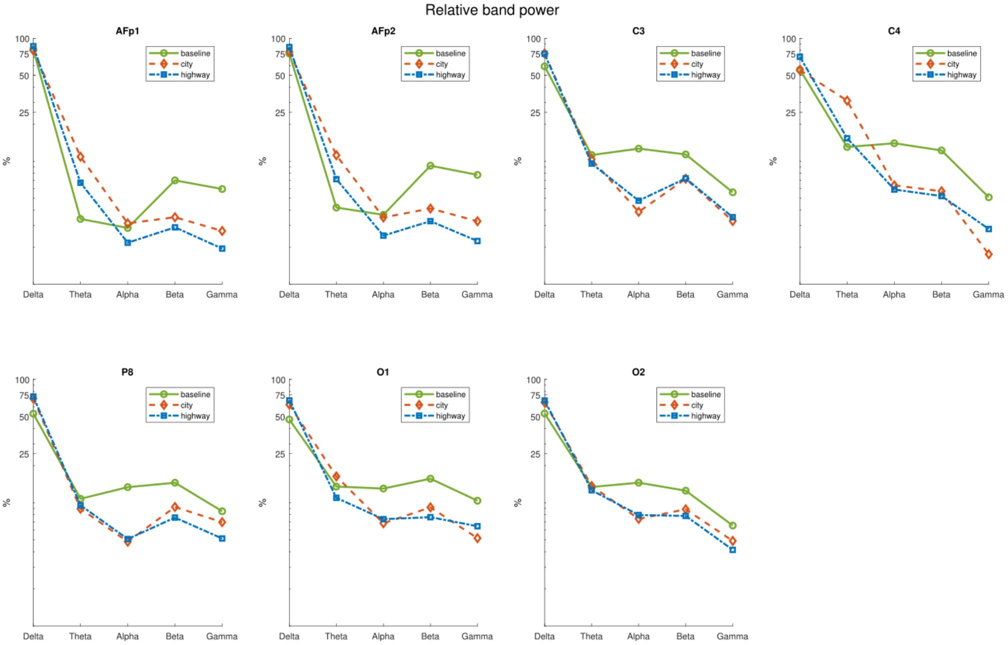


Figure 5. Relative spectral band power. Electrode P7 discarded due to signal quality.

Result of preliminary analysis fNIRS. Three representative data slices were selected for illustrating concentration changes in oxygenated (ΔHbO_2) and deoxygenated (ΔHbR) hemoglobin from three periods during the drive: resting state (*baseline*), *city* driving and *highway*. In Fig. 6a we see ΔHbO_2, which increase with increased brain activation. Fig. 6b depicts ΔHbR, which decrease with brain activation. The optodes' placement over the prefrontal cortex is indicative of cognitive load. As such, we observe an increase in the driver's cognitive demand during *city* driving compared to *baseline* and *highway*. Interestingly *highway* driving seems to impose less cognitive demand than a resting state (*baseline*), contradictory to EEG results. In part, this discrepancy may result from different data slices being analyzed, but also from difference in technical principle of EEG and fNIRS. Mental fatigue may be a reason for lower cognitive load during *highway* compared to *baseline* (as measured by fNIRS), in which case it supports literature that highlights drivers' mental fatigue as major risk and cause factor for road accidents (Ahn et al., 2016; Ingre et al., 2006; Lal & Craig, 2001).

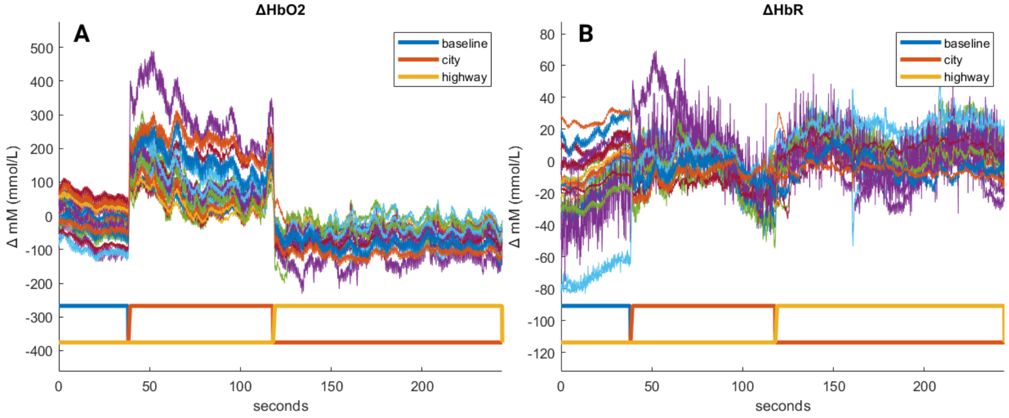


Figure 6. Changes in ΔHbO_2 (A) and ΔHbR (B). ΔHbR channels are centered around zero, whereas ΔHbO_2 channels fluctuate more.

3.2.3 Lessons learned - The setup enable in situ studies of important research topics

Changing experimental setup from a stationary classical experiment room to a mobile *in situ* car driving experiment did not produce major challenges or any corrupt data. In fact, the 50Hz interference in EEG from building power lines disappeared. The level of noise did increase (notably motion artifacts from muscle movement, and ocular artifacts). This could be mitigated by marking movements captured by video and removing this data. Furthermore, there exists several methods for artifact removal in EEG-signals that preserve more raw data (Jung et al., 2000; Kim & Im, 2018; Zeng et al., 2013). Of course, cabling and physical setup in the car do generate some extra, though negligible, work. As mentioned, our data analysis supports research on drivers' mental state, most notably cognitive load, fatigue, and drowsiness (Ahn et al., 2016; Borghini et al., 2014; Ingre et al., 2006; Lal & Craig, 2001). Our setup is able to record potentially crucial changes in mental state, such as cognitive workload, and can thus be used to study such important research topics in situ.

3.3 Case 3: Ashtanga Vinyasa Yoga practice in situ

3.3.1 Description of an Ashtanga Vinyasa Yoga practice

A pilot experiment of an Ashtanga Vinyasa Yoga practice (ashtanga for short) was conducted to further test robustness and flexibility of the setup in studies with movement. Ashtanga is a moving yoga practice consisting of a standardized sequence of physical postures, connected by flowing movements and synchronized breathing patters, performed the same way every time (Mikkonen et al., 2008). Two participants practiced the half primary series by following instructions from a free online class (Ashtanga Yoga Full Primary Series with Ty Landrum, 2020) in their own living room. They were instructed to perform the practice as normal as possible. The sensors were attached to one participant, and fNIRS wires were secured to the participant's body by means of a tight-fitting top. The participant wearing the sensor cap did not perform postures that would affect electrode and optode scalp connection (e.g. headstand), and had two years of experience practicing other yoga types. Several sessions have been recorded. The illustrated session (Fig. 7) collected data from EEG, fNIRS, electrocardiography and two webcams simultaneously. However, only fNIRS data has been analyzed thus far and will discussed in the following section.

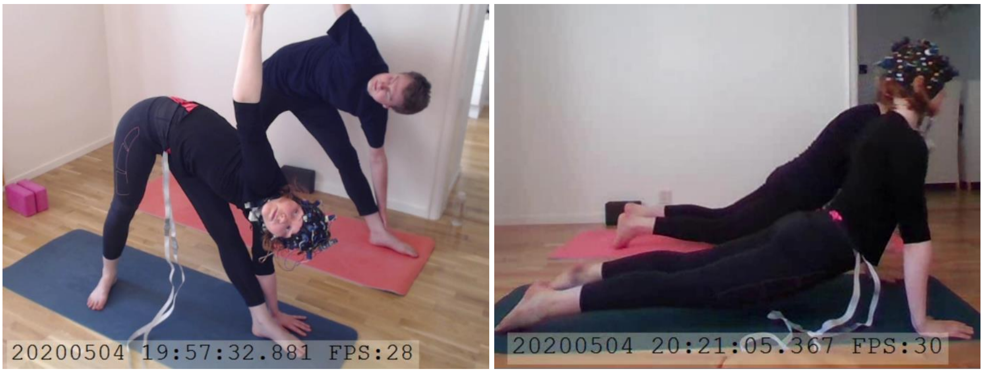


Figure 7. Screenshots from webcam recording during the ashtanga practice.

3.3.2 Results from preliminary data analysis

Result of preliminary analysis fNIRS. Fig. 8 depicts hemoglobin data, and Fig. 9 data after additional filtering. In Fig 8 ΔHbO_2 and ΔHbR are dominated by periodic large spikes across all channels. These are motion artifacts caused by the transition between postures. By investigating data from approx. 500 - 1000 s, we see several "equal" periods where high quality data is collected. Combined with domain knowledge and video recording we infer that these "good" periods are collected during the posture downwards-facing dog. In between each good period there are similar motion artifacts, stemming from the transition between postures. Towards the end of the session the number of seated positions increase, and from 3500 s and out the participant performed less postures which resulted in less transitions and thus less artifacts as is visible in Fig. 8. By applying a TDDR filter (corrects motion artifacts by down-weighting outlier fluctuations) (Fishburn et al., 2019) we can view trend lines in the data over time. As we can see most channels follow the same trend with and same artifacts, being more dispersed at the beginning and end of the practice.

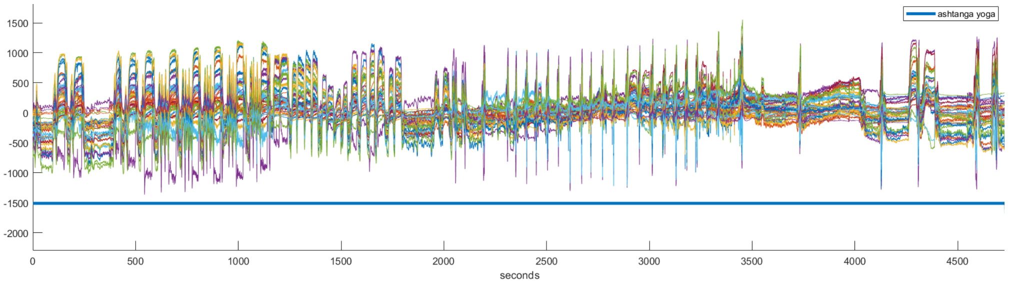


Figure 8. ΔHbO$_2$ and ΔHbR.

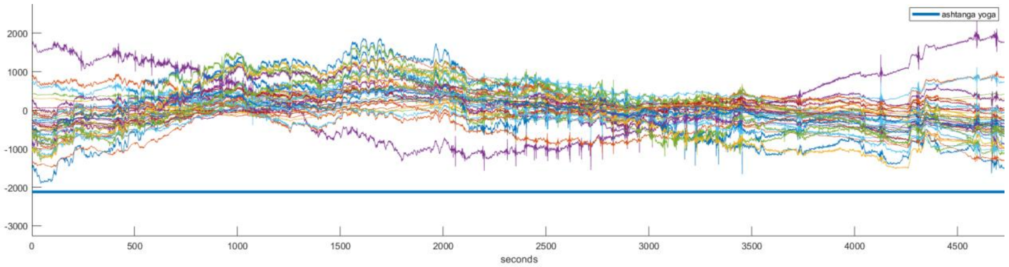


Figure 9. ΔHbO$_2$ and ΔHbR data with TDDR filter (Fishburn et al., 2019). Continuous fNIRS data stream throughout moving sequence of ~ 1 h and 18 min.

3.3.3 Lessons learned - motion artifacts are present; However, we can still acquire acceptable data

In this yoga practice we experience movement artifacts in fNIRS data that are systemic and occur across all channels. Since the newer statistical models (AR-IRLS mentioned in section 2) account for this the systemic artifacts will not be particularly problematic in further statistical inference testing. This differs from signal artifacts discussed in section 3.1.3 and Fig. 4b since the latter affects only one channel. Although not needed for statistical testing, the TDDR motion artifact filter enable visual investigation of overall trends in the data - to the extent that we can separate channels from each other. One channel (S1-D2, purple in Fig. 9) behaves differently from the others, making it subject for additional scrutiny. It's located toward the montage's back left. The EEG battery pack was head mounted, and its weight might alleviate pressure from optode tips, or cause loss of contact with the scalp. This contributes to the large spikes in the data. Furthermore, battery position was slightly to the left on the top of the head, making S1-D2 one of the closest channels. This may contribute to the deviating behavior of the channel. The battery pack should have been placed on the body instead of the head. To summarize, despite motion artifacts, the setup can be used to acquire reasonable data by using motion artifact correction algorithms.

4 EXAMPLES OF POTENTIAL EXPERIMENTS IN DESIGN RESEARCH

The increasingly complex experiment demonstrations showcase feasibility of acquiring neuroimaging data from participants who are 1) seated at a desk conducting tasks within desk range, 2) driving a vehicle (conducting a multifaceted task outside a laboratory), and 3) moving dynamically in space by exercising intentional motor control. We interchange these tasks for typical tasks in design research to gain understanding of potential experiments. Studies on sketching conventionally have participants seated at a desk while they use sketches as means to complete a design task. Therefore, it is possible to use the EEG+fNIRS setup to measure the cognitive processes that take place while sketching. This may even be coupled with the think aloud protocol for triangulation purposes, and to see if participant introspection concurs with neuroimaging data. The setup may also be used in product evaluation in which users wears EEG+fNIRS while interacting with the product to investigate whether the product is fulfilling its intended purpose, e.g. is a proposed interface as intuitive as intended? This can be particularly helpful when comparing several design alternatives. There are many opportunities to investigate the cognition of designers while engaged in the design process, as highlighted by Hay et al. (2020). Given that design is a collaborative activity often conducted by groups of individuals, it is interesting to investigate how they work together to reach mutual understanding and a design problem solution. The neurological synchrony of designers in ecologically valid situations may indeed be investigated with the proposed EEG+fNIRS setup, building upon prior research (Mayseless et al., 2019).

5 CONCLUDING REMARKS

This paper demonstrates a wearable experimental sensor setup collecting multimodal EEG+fNIRS neuroimaging data capture *in situ*. We demonstrate to which variety of experimental scenarios the setup can be appropriated to. Three distinct use cases were tested: a conventional human-computer interaction experiment, a driving experiment *in situ* involving city and highway driving, and a moving yoga practice *in situ*. Resulting data on cognitive load from highly ecologically valid experimental setups are presented. We also discuss lessons learned including signal quality, motion artifacts, and illustrate acceptable and unacceptable data artifacts. We propose to use this setup to measure construct such as cognitive load and suggest a variety of potential experiments pertinent to design research. Taken together, this demonstrate that an EEG+fNIRS sensor setup is both feasible and valuable for design research, and expands the range of what is possible within a design research experiment. We encourage the community to use the setup and adapt it to your *in situ* experiments.

ACKNOWLEDGMENTS

This research is supported by strategic funds from the department for Mechanical and Industrial Engineering (MTP) at NTNU.

REFERENCES

Ahn, S., & Jun, S. C. (2017). Multi-Modal Integration of EEG-fNIRS for Brain-Computer Interfaces – Current Limitations and Future Directions. Frontiers in Human Neuroscience, 11. https://doi.org/10.3389/fnhum.2017.00503

Ahn, S., Nguyen, T., Jang, H., Kim, J. G., & Jun, S. C. (2016). Exploring Neuro-Physiological Correlates of Drivers' Mental Fatigue Caused by Sleep Deprivation Using Simultaneous EEG, ECG, and fNIRS Data. Frontiers in Human Neuroscience, 10. https://doi.org/10.3389/fnhum.2016.00219

Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using Electroencephalography to Measure Cognitive Load. Educational Psychology Review, 22(4), 425–438. https://doi.org/10.1007/s10648-010-9130-y

Ashtanga Yoga Full Primary Series with Ty Landrum. (2020). https://www.youtube.com/watch?v=K-s4IIxVBc8&t=2607s

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

Barker, J. W., Aarabi, A., & Huppert, T. J. (2013). Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS. Biomedical Optics Express, 4(8), 1366. https://doi.org/10.1364/BOE.4.001366

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neuroscience & Biobehavioral Reviews, 44, 58–75. https://doi.org/10.1016/j.neubiorev.2012.10.003

Cairns, P. E., & Cox, A. L. (2008). Research methods for human-computer interaction. Cambridge University Press.

Cisler, D., Greenwood, P. M., Roberts, D. M., McKendrick, R., & Baldwin, C. L. (2019). Comparing the Relative Strengths of EEG and Low-Cost Physiological Devices in Modeling Attention Allocation in Semiautonomous Vehicles. Frontiers in Human Neuroscience, 13. https://doi.org/10.3389/fnhum.2019.00109

Consolvo, S., Harrison, B., Smith, I., Chen, M. Y., Everitt, K., Froehlich, J., & Landay, J. A. (2007). Conducting In Situ Evaluations for and With Ubiquitous Computing Technologies. International Journal of Human–Computer Interaction, 22(1–2), 103–118. https://doi.org/10.1080/10447310709336957

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods, 134(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dybvik, H., Erichsen, C. K., & Steinert, M. (Under Review). Description of a Wearable Electroencephalography and Functional Near-Infrared Spectroscopy (EEG+fNIRS) for in Situ Experiments on Design Cognition. Proceedings of the Design Society: International Conference on Engineering Design.

Fishburn, F. A., Ludlum, R. S., Vaidya, C. J., & Medvedev, A. V. (2019). Temporal Derivative Distribution Repair (TDDR): A motion correction method for fNIRS. NeuroImage, 184, 171–179. https://doi.org/10.1016/j.neuroimage.2018.09.025

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

Goucher-Lambert, K., Moss, J., & Cagan, J. (2019). A neuroimaging investigation of design ideation with and without inspirational stimuli—Understanding the meaning of near and far stimuli. Design Studies, 60, 1–38. https://doi.org/10.1016/j.destud.2018.07.001

Hay, L., Cash, P., & McKilligan, S. (2020). The future of design cognition analysis. Design Science, 6. https://doi.org/10.1017/dsj.2020.20

Herold, F., Wiegel, P., Scholkmann, F., & Müller, N. G. (2018). Applications of Functional Near-Infrared Spectroscopy (fNIRS) Neuroimaging in Exercise–Cognition Science: A Systematic, Methodology-Focused Review. Journal of Clinical Medicine, 7(12), 466. https://doi.org/10.3390/jcm7120466

Ingre, M., Åkerstedt, T., Peters, B., Anund, A., & Kecklund, G. (2006). Subjective sleepiness, simulated driving performance and blink duration: Examining individual differences. Journal of Sleep Research, 15(1), 47–53. https://doi.org/10.1111/j.1365-2869.2006.00504.x

Jacko, J. A. (2012). The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications (3rd ed.). CRC Press.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. Psychophysiology, 37(2), 163–178. https://doi.org/10.1111/1469-8986.3720163

Kim, D.-W., & Im, C.-H. (2018). EEG Spectral Analysis. In C.-H. Im (Ed.), Computational EEG Analysis: Methods and Applications (pp. 35–53). Springer. https://doi.org/10.1007/978-981-13-0908-3_3

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. Brain Research Reviews, 29(2), 169–195. https://doi.org/10.1016/S0165-0173(98)00056-3

Lal, S. K. L., & Craig, A. (2001). A critical review of the psychophysiology of driver fatigue. Biological Psychology, 55(3), 173–194. https://doi.org/10.1016/S0301-0511(00)00085-5

Mayseless, N., Hawthorne, G., & Reiss, A. L. (2019). Real-life creative problem solving in teams: FNIRS based hyperscanning study. NeuroImage, 203, 116161. https://doi.org/10.1016/j.neuroimage.2019.116161

Mikkonen, J., Pedersen, P., & McCarthy, P. W. (2008). A Survey of Musculoskeletal Injury among Ashtanga Vinyasa Yoga Practitioners. International Journal of Yoga Therapy, 18(1), 59–64. https://doi.org/10.17761/ijyt.18.1.l0748p25k2558v77

Okamoto, M., Dan, H., Shimizu, K., Takeo, K., Amita, T., Oda, I., Konishi, I., Sakamoto, K., Isobe, S., Suzuki, T., Kohyama, K., & Dan, I. (2004). Multimodal assessment of cortical activation during apple peeling by NIRS and fMRI. NeuroImage, 21(4), 1275–1288. https://doi.org/10.1016/j.neuroimage.2003.12.003

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. Clinical Neurophysiology, 112(4), 713–719. https://doi.org/10.1016/S1388-2457(00)00527-7

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2018). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. Annals of the New York Academy of Sciences. http://dx.doi.org/10.1111/nyas.13948

Piper, S. K., Krueger, A., Koch, S. P., Mehnert, J., Habermehl, C., Steinbrink, J., Obrig, H., & Schmitz, C. H. (2014). A Wearable Multi-Channel fNIRS System for Brain Imaging in Freely Moving Subjects. NeuroImage, 85(0 1). https://doi.org/10.1016/j.neuroimage.2013.06.062

Santosa, H., Zhai, X., Fishburn, F., & Huppert, T. (2018). The NIRS Brain AnalyzIR Toolbox. Algorithms, 11(5), 73. https://doi.org/10.3390/a11050073

Steinert, M., & Jablokow, K. (2013). Triangulating front end engineering design activities with physiology data and psychological preferences. 109–118.

Sterman, M., Mann, C., & Kaiser, D. (1993). Quantitative EEG patterns of differential in-flight workload.

Zeng, H., Song, A., Yan, R., & Qin, H. (2013). EOG Artifact Correction from EEG Recording Using Stationary Subspace Analysis and Empirical Mode Decomposition. Sensors (Basel, Switzerland), 13(11), 14839–14859. https://doi.org/10.3390/s131114839

Appendix C15: Academic contribution 15

Dybvik, H., Erichsen, C. K., Steinert, M. (Manuscript) Tetris' effect on cognitive load, performance, and systemic neurophysiology. To be submitted to Frontiers in Human Neuroscience.

Appendix C16: Academic contribution 16

Aalto, P., Dybvik, H., & Steinert, M. (Under review). A stroll through a cathedral: FNIRS and space sequences in architecture. *Frontiers in Neuroergonomics*.

C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16

frontiers

A stroll through a cathedral: fNIRS and space sequences in architecture

Pasi Aalto[1*], Henrikke Dybvik[2], Martin Steinert[2]

[1]Department of Architecture and Technology, Faculty of Architecture and Design, Norwegian University of Science and Technology, Norway

[2]Department of Mechanical and Industrial Engineering, Faculty of Engineering, Norwegian University of Science and Technology, Norway

*** Correspondence:**
Pasi Aalto
pasi.aalto@ntnu.no

Keywords: Architecture$_1$, fNIRS$_2$, in-situ$_3$, method$_4$, neuroimaging$_5$, church$_6$.

Abstract

Combining in situ neural imaging with detailed spatial descriptions and data can provide new insights into how humans experience architecture. In this study, we show a proof-of-concept experiment where mobile optical brain activity monitoring using fNIRS (functional near-infrared spectroscopy) was used to examine the brain activity of a single participant when walking through a Nidaros Cathedral in Norway. The 1.5-hour walk included distinctly different spaces, including the nave, the catacombs, the built-in-wall stairwells, secluded places of worship as well as the rooftop. Control was achieved by closing the cathedral for the experiment, no other people were present. The proof of concept shows that it is feasible to examine architectural space sequences with mobile fNIRS. In this paper we present different means of signal quality assessment, use a generalized linear model to investigate whether different spaces show different responses in the prefrontal cortex, and explore the use of multivariate analysis methods. The study illustrates a viable novel approach to better understand the built environment, both in terms of the response of humans as they move through a spatial sequence, but also to encourage better experiment designs tools that combine neuroimaging and architecture. This combination, when triangulated with other factors like social interactions and spatial data shows great potential as future tool in architectural research.

ISBN 978-82-326-7146-5 (printed ver.)
ISBN 978-82-326-7145-8 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology