Mikko Pohjola, Zhaodong Sun, Alexander Vedernikov, Miriam Nokia, Joona Muotka, Miiamaaria Kujala, Risto Puutio, Anu Tourunen, Xiaobai Li, and Virpi-Liisa Kykyri

Reducing strain and increasing gain of remote work group meetings with physiological indicators.

Final Report of the PhinGAIN project

JYVÄSKYLÄN YLIOPISTON PSYKOLOGIAN LAITOKSEN JULKAISUJA 360

REPORTS FROM THE DEPARTMENT OF PSYCHOLOGY UNIVERSITY OF JYVÄSKYLÄ 360

JYVÄSKYLÄN YLIOPISTON PSYKOLOGIAN LAITOKSEN JULKAISUJA 360
REPORTS FROM THE DEPARTMENT OF PSYCHOLOGY UNIVERSITY OF JYVÄSKYLÄ 360
PUBLISHER: University of Jyväskylä, Department of Psychology
SCIENTIFIC EDITOR OF THE SERIES: Taru Feldt
ISBN: 978-951-39-9756-4

Jyväskylä 2023
Permanent address of the publication: http://urn.fi/URN:ISBN:978-951-39-9756-4

TIIVISTELMÄ

Käsillä oleva raportti on Työsuojelurahaston rahoittaman *Työryhmien etäpalaverien kuormittavuuden vähentäminen ja hyödyllisyyden lisääminen fysiologisten mittareiden avulla* (PhinGAIN) -tutkimushankkeen loppuraportti. Tutkimus toteutettiin tammikuun 2021 ja elokuun 2023 välisenä aikana konsortiohankkeena Jyväskylän ja Oulun yliopistojen tutkimusryhmien yhteistyönä. Hanke jakautui kahteen osaan, joista ensimmäinen oli sisällöllinen ja toinen tekninen.

Hankkeen yleisenä tavoitteena oli tuottaa uutta ja laajasti sovellettavaa tietoa siitä, miten hyödyllistä osallistumista etäpalavereissa voidaan tukea ja palaverien kuormittavuutta osallistujille ja palaverien vetäjille vähentää fysiologisten mittareiden antaman tiedon avulla. Tutkimuksen aineisto koostui neljästä työryhmien työnohjausprosessista, joissa kukin noin seitsemästä yhdeksään henkilön ryhmä tapasi kuusi kertaa työnohjaajan johdolla. Tapaamisia oli aineistossa yhteensä 24 ja ne toteutettiin ja tallennettiin Zoom-verkkokokouspalvelussa. Tutkimukseen osallistui 31 mielenterveys- ja päihdepalveluiden ammattilaista sekä kaksi kokenutta työnohjaajaa. Osallistujat tallensivat tapaamisten aikana omat kasvokuvavideonsa ja mittasivat sykettään pulssioksimetrilaitteella. Tapaamisten jälkeen sekä ohjaajat että ohjattavat arvioivat istunnon vuorovaikutusta sekä tunnistivat keskustelun hyödyllisiä ja kuormittavia kohtia. Lisäksi he arvioivat työtilanteensa yleistä kuormittavuutta tapaamisen aikaan. Aineiston keruu tapahtui Covid-19 pandemian vuoksi kokonaan etäyhteydellä. Tapaamisten jälkeen ohjaajat ja osallistujat lähettivät tallenteet tutkijoille suojattua verkkoyhteyttä käyttäen. Asetelma ja aineiston keruun tehtävät olivat vaativia, joten kaikille osallistujille järjestettiin tutkimusavustajien toteuttama henkilökohtainen tuki ennen ja jälkeen kunkin tapaamisen.

Tutkimuksen ensimmäisen osan tavoitteena oli tunnistaa, mikä on työskentelyn kannalta optimaalinen aktiivisuuden ja vireyden tila, ja millaiset keinot edistävät sen syntymistä ja ylläpitoa ryhmän työnohjaustapaamisissa. Tutkimme aktivaatiota kahdella eri tasolla: (i) vuorovaikutukseen sitoutumista (engagement) ilmaisevien sanallisten ja sanattomien käyttäytymisten sekä (ii) osallistujien fysiologisen vireystilan vaihteluiden tasolla. Havaitsimme, että hyvää ja osallistujien hyödylliseksi kokemaa vuorovaikutusta saadaan aikaan videovälitteisesti. Aktiivinen osallistuminen palaveriin sekä omilla puheenvuoroilla että läsnäoloa ja kiinnostusta viestivien sanattomien vuorovaikutuskeinojen avulla oli yhteydessä palaverin koettuun hyötyyn ja vuorovaikutuksen laatuun. Mitä enemmän osallistuja puhui ja ilmaisi vuorovaikutukseen sitoutumista myös ilmeillään, eleillään ja liikkeillään, sitä paremmaksi tapaamiset arvioitiin.

Havaitsimme myös, että tapaamista edeltävän työtilanteen kuormittavuus vaikutti kokemukseen sen hyödyllisyydestä. Mitä kuormittavampi työtilanne oli, sitä heikommaksi osallistujat arvioivat tapaamisen laadun. Tärkeä havainto oli, että tapaamisten alussa toteutettu lyhyt yhteinen hiljainen jakso silmät suljettuina rauhoitti työnohjaukseen osallistujien fysiologista tilaa. Vain kahden minuutin mittainen jakso riitti siihen, että osallistujien sykkeet laskivat tilastollisesti merkitsevästi. Arviointilomakkeilla monet osallistujat tunnistivat, että rauhoittumishetki oli ollut tärkeä ja hyödyllinen.

Kahdessa otokseltaan pienessä laadullisessa tutkimuksessa tarkastelimme vuorovaikutusta hyödyllisiksi arvioiduissa keskusteluepisodeissa. Havaitsimme, että ohjaajat tukivat niissä ohjattavien aktiivista osallistumista, edistivät keskustelujen koherenssia, sekä haastoivat ja kutsuivat osallistujia ottamaan uusia näkökulmia puheena oleviin asioihin.

Tällaisten hyödyllisten keskustelujaksojen yhteydessä havaittiin, että aktiivisesti keskustelussa mukana olevien osallistujien sykevälivaihtelu lisääntyi, mikä viittaa fysiologiseen rauhoittumiseen.

Tutkimuksen toisessa osassa tavoite oli tekninen: kehittää sykkeen ja hengityksen analysointia videokuvasta erityisesti, kun kuvattava puhuu ja saattaa myös kokea ja ilmaista voimakkaita tunteita. Kehitimme ohjaamattoman oppimisen menetelmän, joka pystyy langattomasti mittaamaan fotopletysmografisignaalia (rPPG) eli sykettä kasvokuvavideoista. Testeissä osoittautui, että menetelmän avulla saavutetaan erittäin hyvä mittaustarkkuus.

Lisäksi hankkeessa kokeiltiin multimodaalisia tekniikoita, joiden avulla voidaan yhdistää fysiologisia ja käyttäytymispiirteitä stressi- ja kuormittavuustasojen mittaukseen. Kolme päähavaintoa olivat: 1) rPPG-piirteet toimivat paremmin stressin arviointiin kuin käyttäytymispiirteet; 2) kahden minuutin aikaikkuna on riittävä rPPG analyysiin keskittymistasojen mittaukseen; 3) kasvonilmeiden lihasten jännittyminen ja pään liikkeet ovat tehokkaita käyttäytymispiirteitä keskittymistasojen mittaukseen videolta havaittuina.

Reaaliaikaiseen rPPG-signaalin mittaukseen kehitettiin matkapuhelinsovellus ja menetelmää testattiin keräämällämme etäpalaveriaineistolla.

Avainsanat: Etäpalaverit, työnohjaus, vireytyminen, syke, sykevälivaihtelu, psykofysiologia, kuormitus, konenäkö, koneoppiminen, sykkeen estimointi, rPP

ABSTRACT

This is the final report of the *Reducing strain and increasing gain of remote work group meetings with physiological indicator*s (PhinGAIN) research project funded by the Finnish Work Environment Fund. The PhinGAIN research was conducted 1/2021- 8/2023 as a consortium project in collaboration with the University of Jyväskylä's and the University of Oulu's research teams. The research tasks were divided into two parts, the first one being thematic and the second one technical.

The main aim for the study was to bring about new and largely applicable knowledge on how useful participation could be supported and participants' strain could be reduced in remote work group discussions using physiological measures during the meetings. The data for this study consists of four online work group supervision processes, six sessions for each group, 24 sessions altogether. Participants were 31 supervisees (seven to nine per group) and two experienced supervisors. Sessions took place and were recorded on the Zoom videoconference service. Additionally, participants recorded their own face videos, and measured their heart rate with pulse oximeter device. Participants evaluated the interactions after each session with a web-based questionnaire and rated the amount of strain at work at the time of the session. Due to the Covid-19 pandemic, the data were collected remotely.

In the first part of the study, our aim was to study what is an optimal state of activation for working group meetings and how it can be facilitated in interaction during the meetings. We studied activation in participants' (i) verbal and nonverbal behaviors displaying engagement, and (ii) changes in the level of physiological arousal. We found that good interaction that the participants consider useful is possible in online mediums. For the success of a meeting, active participation in discussion is beneficial. The participants considered a meeting to be more useful the more actively they participated in the discussion with their own speaking turns, and also with their nonverbal behaviors. Meetings were found to be less useful when people participated in them when they already felt their work situation straining. We also noticed that a shared, quiet calming-down moment with closed eyes at the beginning of the remote meeting was useful for the participants. In only two minutes, participants' heart rates lowered to a statistically significant degree, and this brief calming down had been important for many participants. In qualitative case studies we focused on episodes that the participants had mentioned being especially useful for them. Supervisors had an important role in facilitating active participation and engagement of the supervisees. They also promoted conversational coherence and challenged the supervisees to take new perspectives. During this kind of useful dialogue, the heart rate variability (HRV) of the participants who actively took part in the conversation increased.

In the second part of the study, our aim was to develop a procedure for measuring heart rate from face videos in a conversational setting during which participants may talk and express emotions. We developed an unsupervised learning method, which can remotely measure photoplethysmography (rPPG) signals from facial videos. It was tested on our collected online group meeting video dataset and achieved superior accuracy for heart rate measurements. We further explored multi-modal approaches to fuse physiological and behavioral features for estimating stress and engagement levels. Three main findings are: 1) rPPG features work better than behavior features for stress estimation; 2) a time window of two minutes is sufficient for rPPG analysis for engagement measure; 3) Facial Action Units and head rotation are efficient behavioral features for engagement measure. We implemented

the proposed rPPG method as a mobile application, which works well for real-time heart rate measurement.

Keywords: Work group supervision, activation, heart rate, heart rate variability, psychophysiology, machine vision, machine learning, remote PPG

FOREWORD

This final report is a summary of the findings of the research project Reducing strain and increasing gain of remote work group meetings with physiological indicators (PhinGAIN).

Principal Investigator of the PhinGAIN research has been *Associate Professor Virpi-Liisa Kykyri*, who has also led the research team at the University of Jyväskylä. Researchers in the team were Associate Professor Miriam Nokia, Academy Resercher Miiamaaria Kujala, Dr. Mikko Pohjola, Dr Risto Puutio, and University Teacher Joona Muotka. *Assistant Professor Xiaobai Li* has led the research team at the University of Oulu. Researchers in the team were Dr. Alexander Vedernikov and Doctoral Student Zhaodong Sun.

This research has been funded by the Finnish Work Environment Fund. Consortium partners, the University of Oulu, and the University of Jyväskylä, have also provided funding for the study. We wish to thank all the funders. Especially funding from the Finnish Work Environment Fund has been crucial for the project. We wish to thank research specialist Mikael Saarinen for his excellent advice and support throughout the entire project.

Partners of the research project have been the City of Jyväskylä, Metanoia Institute, and Arinna Ltd. We are grateful for the partners for their collaboration and support during the project. Especially we want to thank Risto Puutio (Metanoia Institute), Tuija Bäck (Arinna Ltd), as well as Eeva-Liisa Liimatainen, Tarja Lappi, and Harri Säkkinen (City of Jyväskylä) for your invaluable help with both practical and technical details and solving of several minor and major problems during the data gathering phase.

We thank Chief Laboratory Technician Petri Kinnunen and Laboratory Engineer Arto Lipponen at the Department of Psychology for your help and support in data gathering. Many thanks to the research assistants, psychology students Aliina Kyyrönen, Erika Saarinen, Eeva Mäkelä, Nea Juntunen, Joanna Palén, Lotta Enestam, Linnea Matikainen and Roosa Heinonen for your invaluable contribution during the data gathering. It would not have been possible to succeed in this very complicated task without your help.

In preparing the data for the analyses, as well as performing time-consuming pre-processing and coding tasks, the following research assistants, and psychology students as part of their dissertation project work have participated in this study: Eeva Mäkelä, Aliina Kyyrönen, Erika Saarinen, Nea Juntunen, Oona Laitila, Aino Jokinen, Venla Ruuhonen, Heikki Törmi, Tiina Lehtonen, Janina Henttonen, Piatta Pynnönen, Noora Haapanen, Noora Hanhikoski. Thank you for your important contributions to the study. We also acknowledge the CSC-IT Center for Science, Finland, for providing computational resources.

And last, but not least, we are most grateful to the participants in this study. We wish to express our deepest gratitude to you. Without your willingness to participate, and without a notable amount of your committed work in completing the individual tasks during the data collection, we would not have had any possibility of doing this research. Thank you for providing us with access to this valuable data.

In Jyväskylä and in Oulu, 29.9.2023,

Authors

CONTENTS

1. INTRODUCTION

In this final report we present the objectives, implementation, and results of the project "Reducing strain and increasing gain of remote work group meetings with physiological indicators (PhinGAIN)". This research was carried out by a consortium of the Department of Psychology at the University of Jyväskylä and the Machine Vision and Signal Processing Unit at the University of Oulu.

This research is divided into two parts. In Part I, the aim was to study what is an optimal state of activation for working group meetings, and how it can be facilitated in interaction during the meetings. The University of Jyväskylä's research team was responsible for the Part I study, as well as data gathering for both studies, and providing qualitative observational analyses to be used in Part II studies.

In Part II, the aim was to develop a procedure for measuring heart rate from face videos in a conversational setting during which participants may talk and express emotions. The University of Oulu's research team was responsible for Part II.

1.1 Theoretical perspective

The PhinGAIN research is anchored in the so-called embodied turn in social sciences (Cromby, 2012), as it combines physiological measurement data with multimodal analysis of social interactions. Moreover, this study is connected with the growing field of mediated online interaction studies (Thompson, 2018; Rettie, 2009).

In the PhinGAIN study we examined interactions during synchronous online work group meetings. More specifically we studied clinical supervision targeted to mental health and substance abuse service professionals. Supervision sessions for groups of seven to ten participants were led by an experienced supervisor and conducted remotely by using videoconference software. In work supervision sessions, facilitated by a professional supervisor, participants address problematic work situations and can also share emotions experienced at work, providing a good context for exploring the co-regulation of emotions and arousal, such as calming down or the stimulation of shared motivation and enthusiasm.

Remote meetings are often felt more stressful than face-to-face meetings by the participants, as technical problems are common (McColl & Michelotti, 2019). Also, non-verbal interaction is only partially transmitted over the network. What is not fully transmitted are the bodily messages, i.e,, bodily elements of communication, that enable bodily resonance or sympathetic vibration, which are needed to identify (sometimes non-consciously) one's own and others' emotions and to regulate them (Fuchs & Koch, 2014). In the absence of bodily messages, it becomes more complicated and requires more effort to monitor and respond to the reactions of others, which can be straining (Ngien & Hogan, 2022). On the other hand, video mediation provides distance, which can also be helpful when dealing with emotionally intense topics (Woo, Bang, Lee & Berghuis, 2020).

Previous research has therefore explored the specific characteristics of distance interaction. However, there is a lack of information on the impact of videoconference mode on the collective regulation of the group's state of arousal, and how this regulation could be supported during remote meetings by means of interaction and real-time biofeedback.

The study builds on the classic observation, known as the Yerkes-Dodson (1908) law, that the relationship between performance and arousal follows a linear curve for simple tasks and an inverted u-shaped curve for complex tasks: an appropriate amount of physiological and psychological arousal improves performance, but if arousal is excessive, performance starts to deteriorate. For example, strong arousal can interfere with the recall of what was realized during the effort.

The main research focus in our study was therefore the variation in participants' activation during the work group supervision sessions. Activation can be observable in behaviors, and also measurable by recording the participants' autonomic nervous system responses. Therefore, activation was examined in two different ways in this study:

- As verbal and nonverbal behavioral displays of engagement in the conversational interaction through e.g., speech, nods, smiles, and movements.
- Changes in participants' physiological state, as reflected in participants' heart rate and heart rate variability.

Rooting on Goffman's (1957, 1963) early notions of the obliging nature of the expectation of participation and responsiveness in the human social interaction and based on the definition by Peräkylä and colleagues (2021), we used the concept of engagement as

referring to how the participants in a conversation show their involvement in the conversation to each other through talk and nonverbal behaviors.

1.2 Physiology and virtual meetings

The rise of online meetings has introduced new complexities to interaction. The nuanced bodily signals that facilitate resonance between individuals are lost in virtual interactions, affecting the recognition of emotions and attitudes, especially in tense situations (Fuchs & Koch, 2014; Pönkänen et al., 2011). Despite the growing acknowledgment of technostress and Zoom fatigue, there is a lack of research exploring the wellness implications of remote meetings (Kilcullen et al., 2021). This gap is particularly noticeable in the context of integrating physiological stress measurements with video-mediated interaction analyses (Riedl, 2022; Hoogeboom et al., 2021).

The Autonomic Nervous System (ANS) plays a critical role in regulating an individual's physiological state, including levels of arousal or alertness. The ANS has two primary branches: the sympathetic nervous system, which controls active states like exercise and work, and the parasympathetic nervous system (PNS), responsible for passive states such as rest and digestion (Jänig, 1989). Heart rate (HR) and heart rate variability (HRV; the fluctuation in the time intervals between adjacent heartbeats) are frequently utilized to assess ANS activity. High HRV is generally associated with PNS activation and states of relaxation and recovery, whereas low HRV suggests stress (Shaffer & Ginsberg, 2017; Plans et al., 2019; Van Amelsvoort et al., 2000).

HRV is particularly useful in understanding interpersonal interactions and communication dynamics. For instance, higher HRV levels have been linked to positive working relationships (Blanck et al., 2019) and are indicative of more positive and fewer negative emotional states in non-intimate interactions (Mauersberger et al., 2022). Gender-specific responses related to HRV also exist, with males exhibiting higher levels of high frequency HRV being more inclined towards cooperative behavior (Lischke et al., 2018).

Sympathetic activation, marked by increased HR, occurs in preparatory actions and nearly all emotional states (Kreibig, 2010; Pörhölä et al., 1993). Arousal, emotions, and cognitive load appear to be interconnected both at individual and group levels (Butler & Randall, 2013). For instance, in storytelling situations, the listener's alertness increases while

the narrator relaxes, thus sharing the emotional load (Peräkylä et al., 2015). A recent study on team leaders' behaviors in face-to-face meetings found differences in physiological arousal based on their competency ratings during relationship-oriented interactions (Hoogeboom et al., 2021). However, research comparing the physiological burdens of hosting remote versus in-person meetings remains scant.

In our previous study (Lampinen et al., 2018), we showed that in a four-person conversation, the workload of the conversation can be counterbalanced by short calming moments that relax the participants, which we also use in the job supervision sessions in this study.

In summary, as remote work becomes increasingly prevalent, a comprehensive understanding of the physiological and emotional impacts of virtual meetings is vital. In this context, HR and HRV serve as useful indirect measures of participant activation and warrant further investigation.

1.3 Remote measuring of heart rate

The progress of scientific research is highly related to the development of technologies. Traditionally, physiological sensors need to be attached to human bodies in order to make reliable measurement of signals. For example, Photoplethysmography (PPG) (Allen, 2007) is one major approach commonly used for heart activity measurement and HRV analysis. A PPG records blood volume changes in peripheral microvascular, e.g., at an earlobe or a fingertip by using a pulse oximeter. The pulse oximeter is an optical sensor which illuminates skins with LED lights and measures the changes of light absorption by hemoglobin in blood.

One constraint of such contact measurement is that it requires the subject to be physically presented, and professionals (e.g., a physician or a nurse) are usually needed to operate the device. This limits the locations of such measurements only to hospitals or laboratories where the devices and the professionals are available. Besides, another constraint of contact measurement is that the attached sensors will impede subjects' movements and cause uncomfortableness especially for long-duration measurements.

A new approach of heartbeat measurement emerged in 2008, when results of a study (Verkruysse et al., 2008) indicated that it is possible to capture PPG signals remotely

(rPPG) from facial videos under ambient light. The fact is that when the heart pumps blood through the body, it causes subtle facial color changes (due to fluctuation of hemoglobin count within a local region) which are not visible to the eyes but could be captured by cameras. This finding has led the research focus from traditional biomedical field into computer vision and machine learning, that various methods have been proposed for heart rate measurement from facial videos. The research topic has grown fast in the last decade, and more research teams are joining in especially after the breakout of COVID, which pushed our living towards online mode thus contactless, remote technology received unprecedented attention.

Early rPPG methods relied on prior knowledge to design straightforward signal processing steps such as the independent component analysis (ICA) (Poh et al., 2010), but they did not involve learning or training. The performance of such methods is limited, especially on challenging video data with noises. Since 2016, researchers have made more efforts to develop machine learning and deep learning methods that can leverage contact-measured ground-truth signals and achieve more robust performance than traditional methods. One big challenge of the rPPG measurement is that the target signal is very subtle which can be easily impacted and even submerged by noises such as body movements and illumination fluctuations. It is much easier if the videos are recorded in the lab with well-controlled environmental settings, while in this study, we concern practical scenarios of online group meetings in which subjects record their own videos at their workplace and homes with different computers and cameras. This makes it particularly challenging for the rPPG measurement task. We are targeting at developing machine learning methods to counter the impacts and achieve robust and accurate rPPG measurement for the downstream task of emotion analysis.

1.4 Research questions

The overall goal of the PhinGAIN research project was to bring about new and largely applicable knowledge on how useful participation could be supported and participants' strain could be reduced in remote work group discussions using physiological measures during the meetings. In this study, we had two different aims; first, to study what is an optimal state of activation for working group meetings and how it can be facilitated in interaction during the meetings, and second, to develop a procedure for measuring heart rate from face videos in a conversational setting during which participants may talk and express emotions.

Research questions were:

1.How does the level of verbal and non-verbal activation of the participants vary in the sessions, and how does this variation depend on the role of the participant (supervisor, supervisee)? Additionally, how is this variation associated with the participants' ratings of the session interactions?

2. How do the levels of physiological activation among participants differ across the various phases of the session? Additionally, how does this variation correlate with the participant's role, be it supervisor or supervisee?

3.What is straining for the participants in the sessions? What happens in the interaction in segments of the session in which physiological markers of stress are observed in the participants?

4. What types of fluctuations occur in participants' arousal levels during emotionally distinct segments of the session's interactions? What facilitates meaningful dialogue concerning emotionally charged topics?

5. Is there a relationship between variations in both behavioral and physiological activation and shifts in the meaning-making process, such as the emergence of insights and the creation of a shared understanding among participants during the conversation?

6. How to deal with varying video qualities, such as resolution, lighting, and sampling rate? How to reduce the impact of noisy ground truth labels? What are effective ways to utilize data augmentation techniques to enhance the training dataset for the algorithm? What other machine learning approaches can be used to alleviate the constraint of insufficient training data?

7. How can the developed methods be implemented as software or application for real-time physiological measurement of participants during online video meetings? What approaches can be taken to minimize computational load, enabling high-speed and reliable measurement?

2. METHODS

2.1 Data collection

To study remote work group meetings in clinical work supervision contexts, the data of the PhinGAIN research project were gathered in close collaboration with the City of Jyväskylä. The city had started an organizational change in which services targeted for clients with mental health and substance abuse problems were being reorganized. To support this ongoing change, the City of Jyväskylä decided to organize clinical work group supervision for the workers. Earlier, there had been discussions between researchers and the City of Jyväskylä about the possibility of doing research in connection with this transition. The planned study would focus on what are helpful aspects of interaction in supervision conversations, as well as what might be hindering or even stressful in this kind of interaction. Therefore, the plan was also to measure the participants' psychophysiological responses during the supervision sessions.

Due to the Covid-19 pandemic, however, it turned out to be impossible to use the face-to-face design, which had already been piloted for the study. At the same time, the pandemic condition opened up a whole new and timely highly important research aim for our research, namely, how to support helpful conversations in a remote mode between several participants, and how to solve the challenge of conducting psychophysiological measures during these remote meetings within a fully remote mode design. While trying to solve this problem, we quite early on learned that at the University of Oulu, there was very interesting research and development going on about how physiological responses could be estimated from face videos, using machine vision and machine learning procedures (see, Li et al, 2014). We decided to meet each other, and very soon we realized that it would be fruitful to start collaboration between the two research teams.

In a very short time period of only few weeks we were able to modify our face-to-face design into a remote mode design, to be used in the present PhinGAIN study, in which the aim was to investigate what would be an optimal window of activation for this type of an intervention, and how this optimal activation could be achieved through activities in social interaction between the participants and an experienced supervisor during the supervision sessions. There had been recent developments prior to the Covid-19 pandemic to use Zoom

videoconference platform in data gathering (Archibald et al., 2019). In the PhinGAIN design, the supervision sessions for the groups were organized via Zoom videoconference platform, and the coaches/supervisors recorded the sessions. During the sessions, all participants recorded their face videos with a separate software measured their heart rate by using a pulse oximeter device. After the sessions, participants and coaches were given a brief web-based questionnaire to rate the session and to provide an evaluation of their recent work strain level. After each session, participants sent the video recordings and the heart rate files to researchers via Funet FileSender (filesender.funet.fi) service, which is a safe and easy-to-use, browser-based file sharing service. Funet FileSender is based on the open-source FileSender application specially developed for higher-education institutions and research communities.

Since the design was very complex, especially for the participants who were not used to tasks such as installing software or making video recordings of their work, guidelines and support were needed. Therefore, psychology students working as research assistants provided support for participants in completing tasks related to the data gathering. This support was organized via Zoom sessions, in which we used the Zoom Breakout Room functionality to assign each participant with a research assistant into one room, to be able to create privacy and to provide individual support according to the participants' needs for guidance. The participants could share their screen and get individualized technical support from the research assistants to complete the data gathering tasks.

The design is illustrated in Figure **1**. Detailed description of the recordings and measurements is provided on pages 19-22.

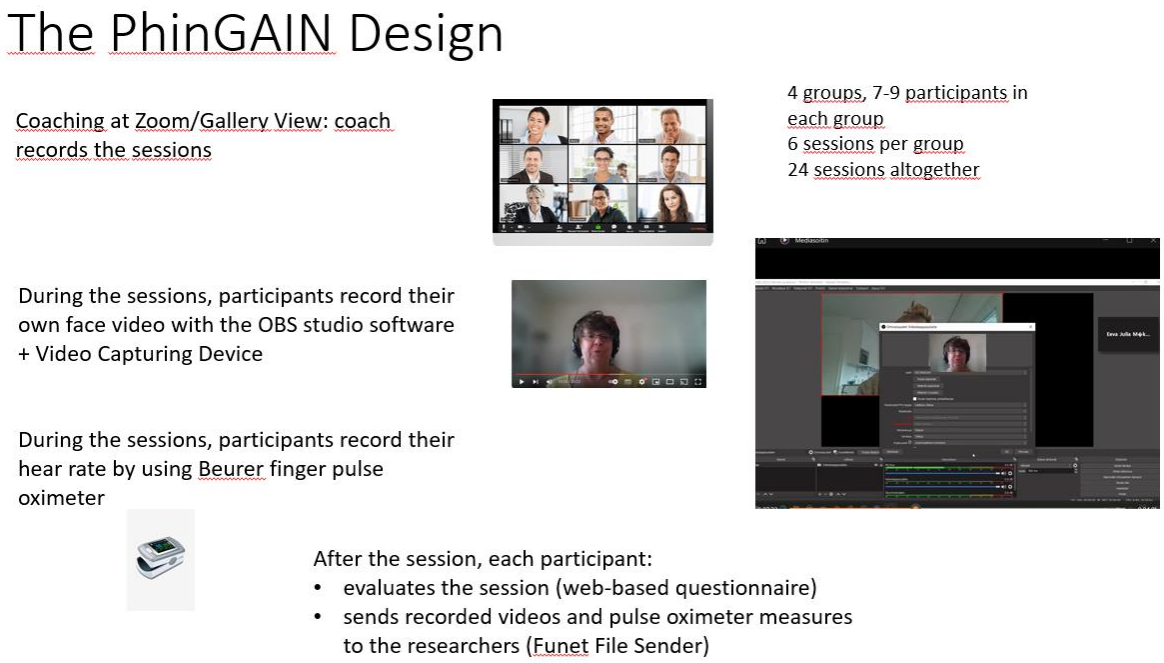

Figure 1. Illustration of the design of the study.

The research plan and the grant application were prepared between June-September 2020. The Ethical Board of the University of Jyväskylä was asked to do the ethical review of the study design and the data gathering procedures. After the ethical review had been successfully completed in the beginning of November 2020, recruiting of the participants and gathering of the data were started. The data gathering phase of the study was completed in May 2021.

2.1.1 Participants

In this study, 21 nurses and 10 psychologists working in public services targeted to clients with mental health problems and substance abuse participated in work group supervision. 27 of the participants identified themselves as female and only 4 identified themselves as male.

31 participants were assigned in four groups for the work group supervision process. The constellation of the supervision groups was decided by the leaders of the organizational unit, and it was based on an objective to include members of various local teams into all four groups. Thus, not all the participants had even met each other prior to the first supervision session. This decision was made since it enabled the sharing of organizational issues and hence learning between different teams in a situation where two previously separate

units of the organization had been merged into one during organizational restructuring. The primary purpose of the supervision was to offer participants space in which to reflect on client-relationships and work-related wellbeing issues during this important but also straining organizational change process.

Each supervision group contained from 7 to 9 participants. Not all participants, however, participated in every session, due to e.g., sick leave, vacations, and other typical issues related to everyday work. Supervision sessions were led and facilitated by two experienced supervisors. One of the supervisors identified herself as female and one as male. Each supervisor was responsible for the supervision sessions in two groups.

The ethical board of the University of Jyväskylä approved the study design, and also the City of Jyväskylä gave permission to gather data for this study. All participants, both supervisors and supervisees, joined the research project voluntarily. After the participants had been given information about the study, they signed an informed consent for gathering and using their data in scientific research. Employees who declined to take part in the research were offered the opportunity to take part in a supervision group outside the research project.

2.1.2 Video-recordings of the sessions

During the research project, six 90-minute sessions were organized for each of the four supervision groups. Our data consists thus of video recordings of 24 work group supervision sessions. Sessions were held on the Zoom videoconferencing platform (Zoom Security Guide, 2016) with each supervisee participating from their own office desk, by using only the integrated webcam on their laptops. No extra cameras were used. Most, but not all participants had audio headsets and used these when available. Sessions were video recorded by the supervisor in the Zoom Gallery View mode and converted to mp4 format.

Each participant, both supervisors and supervisees, recorded their own face video separately using the Open Broadcaster Software Studio (OBS) software (https://obsproject.com/). OBS Studio is free and open-source software for video recording and live streaming. To be able to use the OBS Studio for video recording during a Zoom meeting with only one integrated web camera, a virtual camera was needed. We used the procedure described by Pralat (2020). First, a VirtualCam plugin was installed, and then it was configured with Zoom, the aim being in creating a virtual webcam with the output of OBS and opening

this virtual webcam on Zoom. In OBS studio recordings, we used settings that enabled optimizing quality over the size of the video recording file. This meant that in some cases, video files were large.

Participants were given detailed written instructions and individualized online guidance by research assistants, who used Zoom Breakout Rooms and screen share to provide guidelines for starting and stopping the recordings, as well as sending the files to the researchers via the secure Funet File Sender service. Once the software was installed and the procedure was individually tested and practiced with each participant, the participants were able to complete this recording task relatively easily with the help of research assistants. However, it took some time to start the software and check the quality of the video and audio. Therefore, and since some technical problems occurred, it was important to start preparing the recordings about 20-30 minutes before the session.

2.1.3 Physiological activation

The physiological activation of the participants during the supervision sessions was recorded using a Beurer PO80 pulse oximeter device (https://beurer.com.cn/en/product/pulse-oximeter-po-80/) and the SpO2 Assistant (BEU)V3.1.0.1 software. In the first sessions, we used the pulse oximeter device wireless, to enable participants to move naturally during the sessions. However, as soon as we noticed that not all measured variables were saved when the data were downloaded from the device to the SpO2 Assistant software after the recording, the procedure was changed and in subsequent sessions, the pulse oximeter device was connected to the software via USB cable throughout the session.

2.1.4. Questionnaires

In this study, only basic background information was gathered from the participants. Along with the contact information, participants were only asked about their profession and their gender.

After each supervision session participants were asked to fill in an online questionnaire for the evaluation of the session (Appendix 1). They were asked to answer one question on their personal job strain and four questions from the Session Rating Scale (Duncan

et al., 2003). They were asked to rate the relationship in the supervision session, the topics and goals of the session, the approach and methods of the session, and give an overall rating of the session. Moreover, participants were asked if there were any moments in the session that they felt were a strain or stressful, or moments that they felt were important, useful, or helpful. They were also asked to specify the time at which these important or stressful moments occurred, and what happened in the interaction during these moments.

2.1.5 Technical problems leading to missing data items

Technical problems are likely in this kind of field study with a complex design, and these also occurred during the data gathering, leading in some cases to missing or technically failed recordings and/or signals.

First, because of the large number of the participants, there were some technical problems with the timely installation of the software on the participants' computers on their workstations. The IT services of the City of Jyväskylä provided both permissions and services for forced installation of the Zoom client, the OBS studio, the virtual camera tool, and the SoP2 software. Since some of these also required user actions before the installation was completed, we provided support for that with the research assistants.

Sometimes unexpected problems occurred in installation, which required online help from the IT services. Fatal problems, such as when the software was not installed completely or was not functioning properly, sometimes occurred near to the session time. In these cases, it was not always possible to sort out the problems in time. There was also a large delay in the installation work because several participants' workstations were moved to another building during the research project. This meant that we had to reinstall all the software while the data gathering had already been started.

Due to the problems in installing the software, only the supervisors' face videos were recorded during the first supervision sessions in all four groups. Participants only recorded their heart rate by using the pulse oximeter device, and the supervisor recorded the session interactions on the Zoom. During sessions 2 to 6, most of the participants were able to record their face video and measure their heart rate during the sessions. In some cases, there were problems in recording, leading to a corrupted file and missing data.

In the first two sessions, the participants were not able to use the SoP2 software in recording of the pulse oximeter signals by using a USB-cable. Therefore, they had to use the pulse oximeter device wireless and in the offline mode. Unfortunately, and due to the unexpected problem in downloading the recorded signals from the device, the data were not saved in the required .wave format. Therefore, not all the variables were saved in the recordings. Due to these problems, we did not get good quality heart rate signals from most of the participants in the first two sessions. In sessions 3 to 6, most of the recordings were successful.

Other technical problems occurred, but these were random and local, e.g., unstable internet connection during the recording, or problems in downloading and sending the files to the researchers.

Sometimes a participant did not respond to the session rating questionnaire after the session, which led to missing information in the questionnaire data. Research assistants sent reminders, but if the participant had a longer vacation or sick leave, or her/his working contract had expired, it was not always possible to get a response.

In the final sample of 33 participants and 24 work group supervision sessions, the data consists of

- 24 video recordings in the Zoom Gallery View mode,
- 93 participants' face videos, recorded with the OBS studio software,
- 130 pulse oximeter measurement data files, and
- 140 session rating questionnaires completed by 31 participants and 2 supervisors after the sessions
- 33 basic information questionnaires.

2.2 Observational methods

To be able to study variations in participants' behavioral activity, several observational analyses were needed. In this section, we will present details for annotation of the session phases and speech turns, and coding of displays of engagement and emotions.

2.2.1 Annotation of session phases

After the data gathering was completed, we used the ELAN 6.2 software for further analyses. ELAN is an annotation tool for audio and video recordings (ELAN (Version 6.2) [Computer software]. (2021). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://archive.mpi.nl/tla/elan).

To identify and locate different phases of the session, the Zoom recording file was opened in the Elan for annotation. Tiers for (i) a resting period, i.e., an episode during which participants were asked to close their eyes and just quietly relax, (ii) for a working phase, during which the conversations of the session took place, as well as (iii) for technical problems, such as problems in the Zoom, camera, microphone, or lights turning off or extra people coming in the room, were created. The working phase was further divided into four sub-phases: orientation, start inquiry, working on certain topic or case, and reflection.

The annotation results were then exported in the .txt format and saved in Excel for use in further analyses, which most often were focused on the third sub-phase in which the supervision group worked on a specific topic or case.

2.2.2 Annotation of speech turns

ELAN software was used in annotating the speech turns of the sessions. Each participant was given an individual tier for the annotation. Research assistants watched the Zoom videos and carefully annotated both the beginning time and the end time for each speech turn. Results were carefully checked and after that, exported to .txt file, to be opened in Excel. The exported file contained the following information: Participant, beginning time of each speech turn, end time, total length (in milliseconds).

For the analyses, we calculated the following values separately from the whole sessions and from the working phases of the sessions: amount of speech turns, total length of each participant's speech turns, and the average length of each participant's speech turns. From the working phases of the sessions, we also calculated the proportion of an individual participant's speech turns of all speech turns during that phase.

2.2.3 Coding of displays of engagement

Based on the definition by Peräkylä and colleagues (2021), we used the concept of engagement as referring to how the participants in a conversation show their involvement in the conversation to each other. This is done by using both verbal and nonverbal behaviors. Markers of engagement are talking and showing active listening through vocalizations and nonverbal behaviors, such as gaze, nods, and facial expressions such as smile. Disengagement, on the other hand, is observed when no markers of engagement are observed and, in clear cases, a subject is even turning away or focusing on other activities than ongoing conversation.

Psychology students who worked as research assistants performed engagement coding, which was based on observation of the target subject's behavior from the face (OBS) video. In coding, the DARMA software (Girard & Wright, 2018) was used. DARMA is a continuous measurement system that synchronizes media playback and the continuous recording of observational measurement conducted with a computer joystick. The DARMA software is designed for two-dimensional coding, but in the engagement coding, only one axis (y) was used in the coding. For coding, the axis magnitude was set to 10 and the sampling rate was set to 20Hz. As a result, the DARMA software provided a continuous coding time series consisting of two values per second, 10 being the maximum value and -10 being the minimum value.

Before starting the coding, research assistants participated in training, in which they first familiarized themselves with the four broad engagement categories. These categories were:

- full engagement (FE): subject is either speaking or attempts to take a turn or is speaking on top of someone else, or as a listener, subject is actively showing engagement through minimal vocalizations (mm etc.), nods, and/or facial expressions
- moderate engagement (ME): subject gazes at the speaker, looks like he/she is listening to the speaker
- moderate disengagement (MD): subject gazes away, looks like he/she is still listening
- full disengagement (FD): subject gazes away or her/his eyes are closed, looks like the person is not following the conversation, possible side involvements, such as turning away, opening emails etc).

After being able to identify these four categories from the training videos, the research assistants continued their training by using continuous coding with the DARMA software and a joystick. Figure 2 below illustrates how the four categories informed the continuous coding.

Then the research assistants performed the actual coding for each subject for the active part of the video recording of each group meeting. Inter-rater reliability in control sessions was observed to be good (correlation coefficient over .8).

Figure 2. Four engagement coding categories are located in the y axis of the DARMA coding.

Results of the engagement coding were saved in .txt and .xlsx format. The categories of moderate disengagement and full disengagement were combined to form the category of disengagement. Thresholds were: Full engegement=5<, Moderate engagement=0> … 5; Disengagement=<0.

In Excel, we calculated the time (seconds) spent in full engagement, moderate engagement, and disengagement for each supervisee for each session's working phase. We also calculated the proportion (%) that each supervisee displayed each of these three different kinds of engagement during the working phase of the sessions.

2.2.4 Coding of emotions

Annotation of emotional expressions was done by Henttonen and Pynnönen (2022) based on the Russell's (1980) valence model. The visible emotional expressions of the subjects were

located by observing video recordings one subject at a time. When defining emotional expressions, consideration was given to facial expressions, gestures, tone of voice, and the content of speech, as the video footage was primarily limited to the facial area of the subjects.

Emotional expressions were then grouped and color-coded according to their valence into the following categories: negative, positive, and conflicting. Conflicting emotions were defined, following Larsen and McGraw (2014), as the simultaneous occurrence of both negative and positive emotions. In the original handling of the data, a fifth valence category undefined was formed but was later excluded from the final classification. This undefined category mainly included expressions, gestures, and speech related to showing empathy, which initially were challenging to categorize into other classes. Because empathy can involve either a positive or negative connection to another's emotional state, it was decided to allocate expressions of empathy to the existing categories of negative, positive, and contradictory emotional valences.

Annotations were cross-verified for reliability within the research team. The relative proportions of the subjects' emotional expressions were calculated by valence, both individually and at the group level, in order to compare how much emotional expression occurred during the session in relation to a neutral state.

For quantitative analysis, the valences of emotional expression were coded as follows: Negative = -1, Neutral = 0, Positive = 1, Conflicting = 2. Negative valence included emotional expressions that conveyed feelings of stress, sorrow, anger, and negative empathy. Positive emotional expressions were considered to include joy, amusement, and positive empathy. Conflicting emotional expressions included situations where the participant's speech and emotional expression were in conflict with each other or where strong emotional states were being regulated. In neutral situations, the participant often had a basic expression, where the face was relaxed, and no particular emotional states could be read from it. Neutral expression was also determined by the participant's tone of voice, which did not specifically express negative or positive emotions.

The ELAN software was employed for emotion coding, and specific instances were selected to study emotional synchronization, where either the majority or all of the group members manifested emotional expressions with the same valence. Graphs were generated using Python software to visually represent the distribution of emotional expressions and speech turns for each participant throughout the session.

2.3 Discursive approaches to the session interactions

Discursive approaches were used to gain understanding on interactional features in those supervision meetings that were rated by the participants as useful. In the qualitative analyses, we focused on the Working phase during which the supervision group worked on specific topics. Within each Working phase, there were several topical episodes. Altogether, 72 topical episodes were identified from 24 supervision sessions. From these 72 episodes, we located 11 topical episodes that at least one of the supervisees had identified as useful in the questionnaire. From these 11 topical episodes, we selected 3 episodes for detailed qualitative analyses. The selected episodes were each from the different supervision group. The topics in these episodes were:

- One supervisee's patient case
- The clients, who tend to continue conversation when the meeting time limit is over
- Feeling stress and strain at work

These three episodes were selected because the usefulness of the episode was stated most explicitly by the supervisee. Here are some of the supervisees' written evaluations regarding these three episodes:

- "*I got good new ideas from the whole group for my working with the patient", "it was useful to focus on the patient case"*
- *"The supervisor's comment on how I may give different impression for the client that I had intended to [*was useful for me*]", "The supervisor made a good observation ja gave useful feedback for other supervisee, I also started to think about that myself."*
- "*For me, it was useful to hear one of the supervisees verbalizing the shared experience related to feelings of strain and pressure at work...I felt that she found a new verbalization for the experience that I have often had. This verbalization makes it easier for me to understand the feeling of stress*."

We transcribed the conversation in these three episodes using Jefferson's transcription symbols. This enabled us to analyze how the conversation unfolded sequentially, in turn-by-turn interaction. In the transcription of the first of these topical episodes, we also added participants' nonverbal communication, such as emotional displays, and hand and head

movements. Owing to restrictions caused by the video camera angle, we only added nonverbal behaviors when the change in a participant's expression was clearly observable.

Our qualitative data analysis was informed by discourse analytic approaches that direct the analytical focus on what is done with the language and thus approach the discourse as action (Hepburn & Jackson, 2009; Potter et al. 1993). According to these approaches, conversationalists use different discursive resources to construct social reality. This methodological choice allowed us to approach supervisor's interactional practices such as questions, invitations, validations, and answers to supervisee's initiatives as actions aimed to promote the goals of supervision. We were also interested in how the supervisees responded to these practices and what kind of new understanding they displayed in the episodes that they had themselves assessed as being helpful and generating insights or new understanding. By focusing on how the meanings are created, negotiated, and forged in turn-by-turn interaction (Sacks et al. 1974), it was possible to analyze how new meanings and understanding are collaboratively generated in supervision conversations.

In the analysis process, we watched the video recordings from these topical episodes several times to specify the nature of the supervisor's interactional practices such as questions, invitations, validations, and answers to supervisee's initiatives paying special attention to their interactional function.

In the analysis of the first episode, we also applied the concept of "engagement" (Goffman, 1963) to understand supervisees' participation in the episode. In the analysis of engagement, we were informed by the definition by Peräkylä and colleagues (2021): "To be engaged means to show with one's actions and body that one willingly and wholeheartedly takes part in the encounter at hand and focuses one's attention to it and its participants." (p. 2). Thus, we focused in this episode on the participants' verbal (turn-taking and asking / answering) and nonverbal (direction of gaze, body and head movements, facial expressions) actions.

2.4 Physiological variables: obtaining the heart rate measures

From the photoplethysmogram (PPG) signal stored at a frequency of 60 Hz we obtained the heart rate (HR, bpm) and the heart rate variability (HRV, ms) using MATLAB (MathWorks

Inc., US). Briefly, to detect HR, we determined the peaks in the PPG signal (Figure 3), indicative of heart beats, and determined their frequency per minute. To obtain HRV, we calculated the root mean square of successive differences between heartbeats (RMSSD). To assess changes across time, we then calculated the average HR and HRV in a sliding window of 1 min and 5 min, respectively (Figure 4).

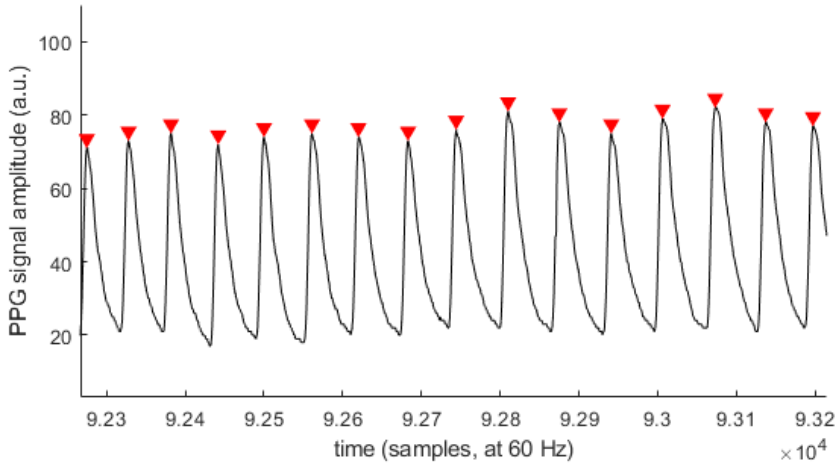

Figure 3. PPG signal (black) with the peaks (red arrowheads) detected.

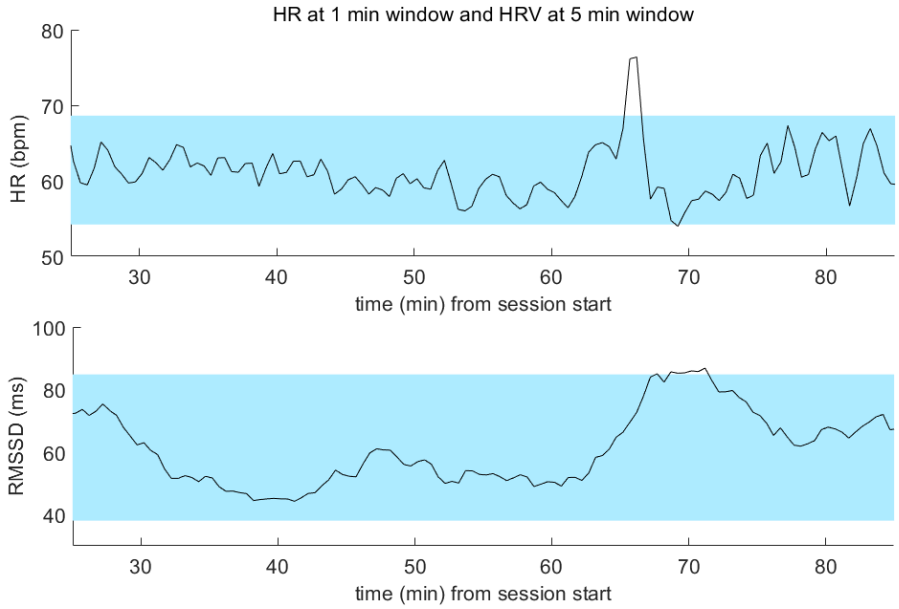

Figure 4. HR and HRV across the session in one representative participant. Blue area indicates the range of +/- 2 standard deviations.

2.5 Statistical analyses

Due to the hierarchical data (participants had several observations), the complex method was used in the models to account for the clustered sample by correcting the standard errors using a sandwich estimator, thus providing more reliable p-values. Group and session effects were tested and, if significant, controlled for in the model. Analyses were carried out with full-information maximum likelihood (FIML) estimation. Estimation accounts for missing values at random (MAR) and includes all the available data.

2.6 Video-based remote PPG (rPPG) measurement

2.6.1 Traditional rPPG method

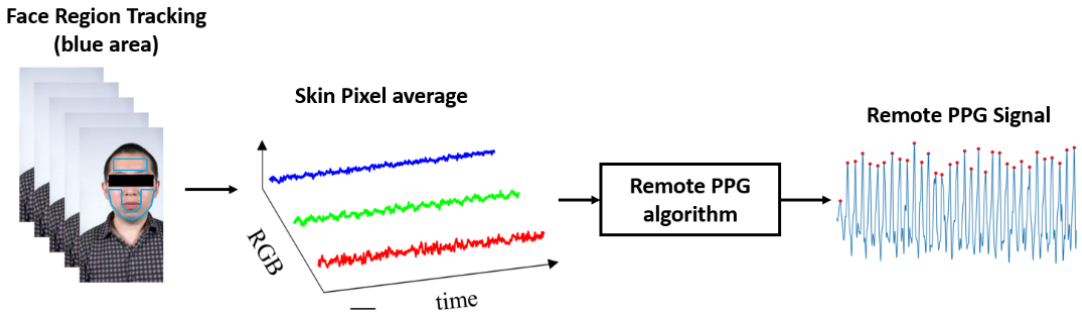

Figure 5. Traditional rPPG extraction steps.

Figure 5 shows the general procedures of traditional method to extract rPPG signals. First, we initiate the preprocessing of face videos to account for motion interference. We utilize OpenFace software (Baltrusaitis et al., 2018) to identify and monitor facial landmarks, ensuring the stability of facial area tracking. Subsequently, we refine these interim outcomes to obtain RGB signals. The traditional methods like POS (Wang et al., 2016) doesn't necessitate training and permits direct pulse signal extraction. We compute the average heart rate over a 30-second interval and generate a heart rate curve to monitor fluctuations in heart rate. Additionally, we employ postprocessing techniques to filter the heart rate curve and eliminate any anomalies. These heart rate curves can be used for the identification of emotional events in the context of remote group meetings.

2.6.2 Deep learning-based method

Deep learning methods are demonstrated to achieve better rPPG signal quality but require ground truth (GT) signals for model training. The GT signals should be synchronized with facial videos and recorded from specific biomedical equipment like pulse oximeters, which cause some burden for data collection. Therefore, we propose Contrast-Phys (Sun & Li, 2022), an unsupervised training method for rPPG measurement as shown in Figure 6. Contrast-Phys is based on contrastive learning and only requires facial videos for model training. We incorporate rPPG prior knowledge (e.g., rPPG spatiotemporal similarity, rPPG cross-video dissimilarity, and heart rate range constraint) into our method to achieve unsupervised training. Our method achieves results similar to supervised methods (Yu et al., 2019; Liu et al., 2020). The method is suitable for TSR dataset since TSR dataset contain noisy, missing, or unsynchronized GT signals, which makes it infeasible for supervised rPPG training. On the other hand, our unsupervised method can easily tackle the issues in GT signals since our method does not require GT signals during training. The proposed unsupervised method achieves similar performance to supervised methods. During inference, unlike traditional methods, the proposed deep learning method can directly use facial videos to output rPPG signals in an end-to-end fashion.

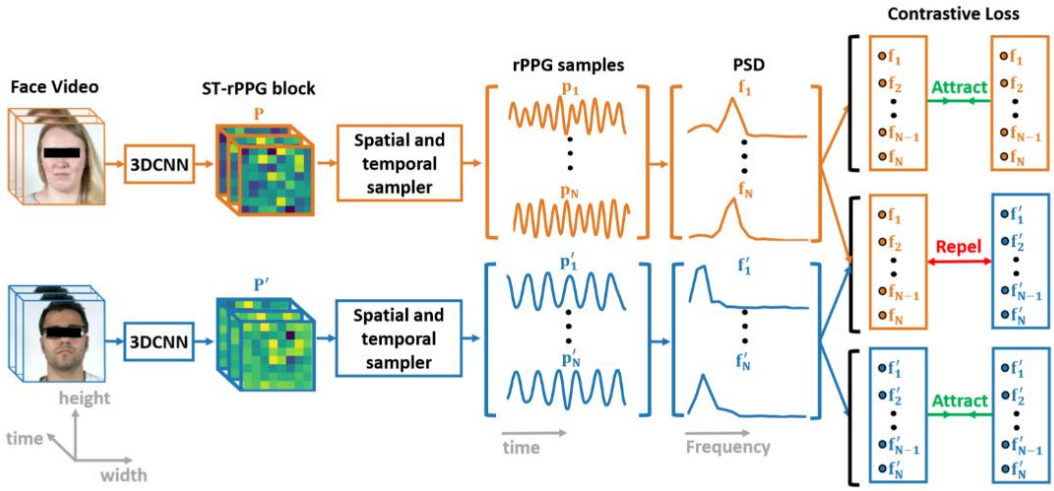

Figure 6. Contrast-Phys: an unsupervised deep learning method for rPPG measurement (Sun & Li, 2022)

2.7 Heart rate variability as a physiological indicator

Heart rate variability (HRV) is calculated from the time intervals between successive heartbeats, known as inter-beat intervals, as shown in Figure 7 below, which can be obtained from rPPG signals.

Here's a simplified step-by-step guide to calculate HRV from rPPG signals:

- **Preprocessing**: a. *Filtering:* Apply a bandpass filter to the rPPG signal to remove noise and motion artifacts. b. *Peak Detection*: Identify the peaks in the rPPG signal corresponding to the cardiac cycles (heartbeats). This can be done using algorithms such as peak detection (Makowski et al., 2021).
- **Calculate RR Intervals:** a. Determine the time intervals between successive peaks in the rPPG signal. These time intervals are the inter-beat intervals.
- **Time Domain HRV Analysis**: a. Calculate basic HRV metrics in the time domain: Mean RR Interval, Standard Deviation of RR Intervals (SDNN), Root Mean Square of Successive RR Interval Differences (RMSSD).
- **Frequency Domain HRV Analysis**: a. Perform a Fourier transform or another spectral analysis method on the inter-beat intervals to convert them into frequency domain components. b. Calculate frequency domain HRV metrics, including: very low-frequency power (VLF): 0.0033 to 0.04 Hz, low-frequency power (LF): 0.04 to 0.15 Hz, and high-frequency power (HF): 0.15 to 0.4 Hz.
- **Nonlinear HRV Analysis**: a. Optionally, you can perform nonlinear HRV analysis, which includes metrics like Poincaré plots, entropy measures, and fractal analysis.

We can analyze the HRV metrics to gain insights into the autonomic nervous system activity. Higher HRV is generally associated with better autonomic flexibility and overall health. HRV is also related to human emotion and is a useful physiological indicator. It's essential to note that estimating HRV from rPPG signals may not be as accurate as using ECG signals, as rPPG signals are more susceptible to motion artifacts and environmental factors. Still, it can provide useful insights into heart rate variability in situations where ECG data is not available such as online video meetings.

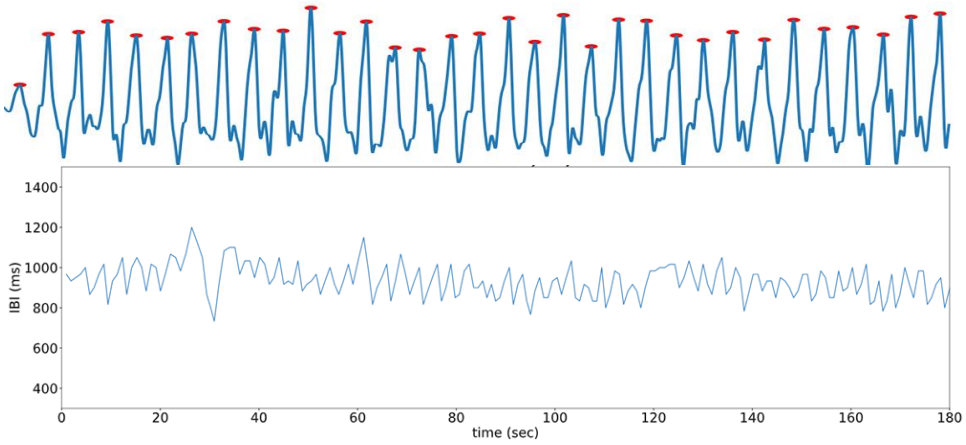

Figure 7. Extraction of inter-beat-interval (IBI) from an rPPG signal. The upper figure shows an example of reconstructed rPPG signal with the peak of each heartbeat marked with a red dot. The lower figure shows the inter-beat-interval signal achieved by subtracting the difference between two adjacent heartbeat peak time points. The IBI curve is used to calculate HRV features.

2.8 Multimodality fusion for emotion recognition

We used rPPG HRV features and behavioral features for emotion recognition such as stress estimation and engagement estimation. rPPG and behavioral (facial expression and motion) features were extracted from the recorded facial videos. Contact PPG (cPPG) features were extracted from pulse oximeter signals and used as a reference. Various combinations of the mentioned features were used for the work stress estimation. The method workflow is shown below in Figure 8.

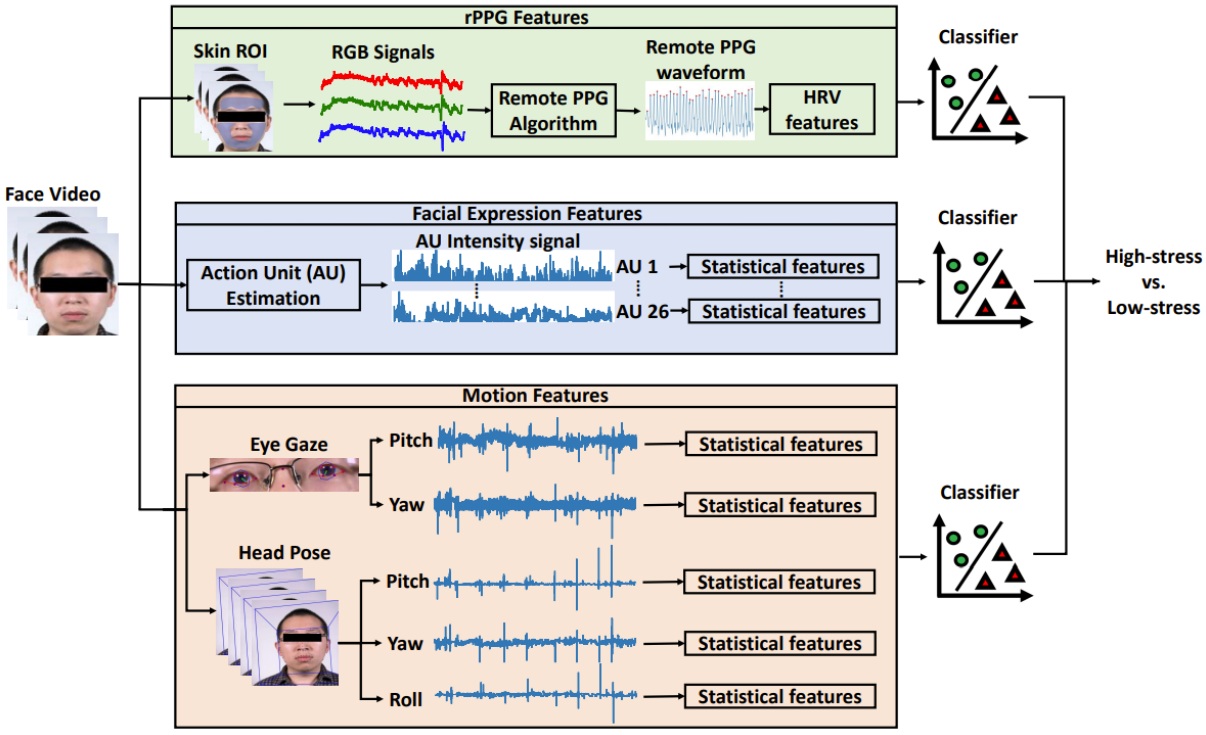

Figure 8. Multimodality fusion for emotion recognition (Sun et al., 2022).

To compute facial expression features, we employed action units (AUs). On each video frame, the OpenFace library was employed to detect and track 16 key types of AUs. These selected AUs encompass the most commonly occurring ones associated with emotions as outlined in the Facial Action Coding System (FACS) (Friesen & Ekman, 1978) Investigator's Guide. The intensity of AUs, ranging from 0 to 5, served as the output. We calculated various statistics, including the mean, median, standard deviation, minimum, maximum, and range, across the temporal dimension of the AUs. Ultimately, this process yielded a 96-dimensional facial expression feature vector for each facial video.

For motion features, we extracted eye gaze and head pose information using the OpenFace library. This encompassed the pitch and yaw of eye gaze, as well as the pitch, yaw, and roll of head pose. To compensate for differences in camera positions, we subtracted the mean value of each feature along the temporal dimension from the original feature signal. Subsequently, we computed statistics such as the median, standard deviation, minimum, maximum, and range for each motion feature across the temporal dimension. This resulted in a 25-dimensional motion feature vector for each facial video.

Finally, we concatenate rPPG HRV features, facial expression features, and motion features for stress classification and engagement estimation.

Stress was classified into two categories (high and low) by thresholding the stress scores reported by participants (Sun et al., 2022). The threshold was chosen to balance the two classes and distinguish the high and low stress levels. Since the stress labels are at the video level and we have limited samples for training and testing, we chose a simple logistic regression model for stress classification.

Engagement was classified into three categories (high, middle, and low) using a variety of classifiers in our study. Since our engagement labels are at the second level, we have many samples for training and testing. We employed a diverse set of classifiers, which encompassed methods such as KNN, Random Forest, AdaBoost, CatBoost, xgboost, lightgbm, Support Vector Machine, Decision Tree, Gaussian Naive Bayes, Quadratic Discriminant Analysis, Neural Net, as well as ensemble combinations of KNN with Random Forest and xgboost with Random Forest. Our analysis involved assessing and comparing the performance of these classifiers to identify the most effective ones.

3. FINDINGS: PART I

In this chapter, findings from Part I of the study will be reported. The University of Jyväskylä's research team was responsible for the Part I analyses and findings.

3.1. Participants' verbal and nonverbal activity and associations with the ratings of the sessions

The first research question in this study was: How does the level of verbal and non-verbal activation of the participants vary in the sessions, and how does this variation depend on the role of the participant (supervisor, supervisee)? Additionally, how is this variation associated with the participants' ratings of the session interactions?

To answer our first research question, we looked at the session ratings (3.1.1), analyzed the associations between the ratings and the level of strain at work (3.1.2), reported our observations on the verbal activity (talk) in sessions (3.1.3), analyzed the associations between verbal activity and session ratings (3.1.4), and reported our observations on displays of engagement in the sessions, as well as analyzed the association between engagement and session ratings in whole sample, and in a smaller sample (3.1.5).

3.1.1 Overall ratings of the sessions

140 individual ratings from 31 participants were given by the participants for 24 sessions in the questionnaire. Cut-offline was 8, scores between 8 to 10 represent good, while scores below 8 stand for poor rating of the session interactions.

Descriptive statistics on Table 1 illustrate that on average, participants gave good ratings of the session in all measured variables. Best ratings were given to quality of the working relationship, in which both the relationship with the supervisor and the supervisees was included in one item of the questionnaire (9.04). Thus, these statistics indicate that on average, participants were quite satisfied with the sessions. However, standard deviations

between 0.97 and 1.56 indicate that there was variation between the participants in the ratings, and not all participants were satisfied with all the sessions.

Table 1. Participants' (supervisees') ratings of the work supervision sessions.

	Quality	**Topics**	**Approach**	**Overall**
Average	9.04	8.37	8.43	8.65
SD	0.97	1.56	1.45	1.33

Note. The scale in the ratings was 1-10. Values between 8 and 10 are good and values below 8 poor. Quality = Relationship quality in the session; Topics = Topics and goals of the meeting; Approach = Approach and methods of the meeting: Overall = overall rating of the session.

3.1.2 Associations between strain in the participants' work situation and session ratings

In the questionnaire, the participants were asked to evaluate how straining their situation at work was at the time of the supervision session. In the questionnaire, the evaluation scale was between one and ten, one indicating not straining at all and ten indicating extremely straining.

It was observed that supervisees in our data (N=114) felt their work situation quite straining, the mean being 7,2. The variance of the ratings was 2,740 the standard deviation (SD) being 1,66.

We wanted to look at if the supervisees' strain at work at the time of the session was linked with their ratings of the session interactions. The findings indicated that the more strain there was at the supervisees' work, the weaker they rated the working relationship (alliance) and the approaches and methods used by the supervisor in the session. The findings are illustrated in Table 2.

Table 2. The relationship (full-information maximum likelihood (FIML) estimation) between experienced strain in the individual work situation and the questions sampling evaluation of the session. Experienced strain correlated with the evaluated relationship quality and approach and methods of the meeting. Quality = Relationship quality in the session; Topics = Topics and goals of the meeting; Approach = Approach and methods of the meeting: Overall = overall rating of the session.

	Estimate	S.E. Est.	S.E.	P-value
Quality	-0.168	0.080	-2.107	0.035*
Topics	-0.177	0.123	-1.434	0.152
Approach	-0.231	0.111	-2.087	0.037*
Overall	-0.160	0.132	-1.211	0.226

* p <.05; ** p < .01; *** p <.001.

3.1.3 Participants' verbal activation during the sessions

On average, supervisees talked 15.8 % of the session time. There was, however, a large variation between individuals in their talking activity (SD = 10.5). While some of them talked over 20 % of the session time, some spent most of the session listening and could have less than one percent, or even not speech at all during the working phase of the session. Supervisors, on the other hand, showed activity in talking and typically, the supervisor was talking 25-33 % of the session time.

The speech turns tended to be allocated by the supervisor and typically there was not much spontaneous dialogue between the supervisees. This is illustrated in Figure 9, which is an example from one topical episode.

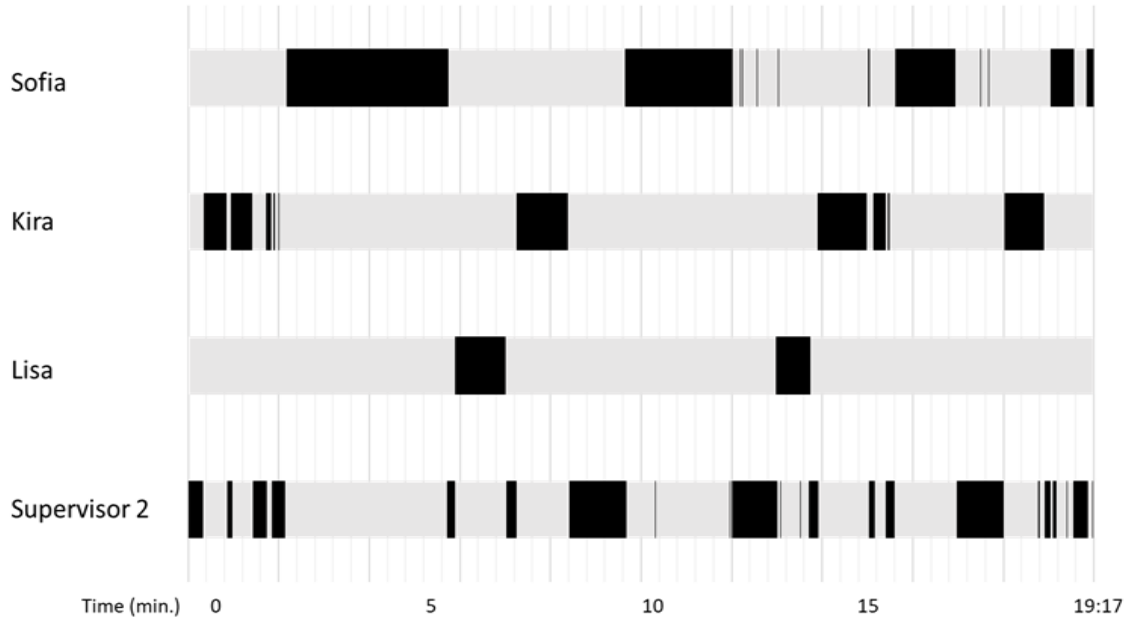

Figure 9. Illustration of the organization of the speech turns in one topical episode.

3.1.4 Associations between verbal activity (speech turns) and session evaluation

We wanted to study whether there were associations between participants' verbal activity and their evaluation of the session. Complex models were used in the analyses, which indicated that the more the participants talked during the working phase of the session, the better they rated the aims and topics of the session. Participants were also overall more satisfied with the session, when their proportion of the speech turns of the session was larger. Findings of the analysis are shown in Table 3.

Table 3. Standardized coefficients (full-information maximum likelihood (FIML) estimation with group and session effects controlled) of the duration and frequency of the participant speech turns during the working phase of the sessions, and percentage of the participant speech duration (relative to all the talk within session) with the questionnaire responses of the participant. Strain = Felt strain at work; Quality = Relationship quality in the session; Topics = Topics and goals of the meeting; Approach = Approach and methods of the meeting: Overall = overall rating of the session.

	Strain	**Quality**	**Topics**	**Approach**	**Overall**
Total duration of the speech turns	0.036	0.204*	0.272**	0.181	0.183*
Average duration of the speech turns	0.090	0.104	0.081	0.084	0.022
Frequency of the speech turns	-0.039	0.161	0.240**	0.114	0.166*
Proportion of a participant's speech turns (%) of all speech turns	0.037	0.179	0.301***	0.155	0.212**

* p <.05; ** p < .01; *** p <.001.

3.1.5 Participants' displays of engagement during the sessions, and associations between engagement and session evaluation

During the working phases of all the 24 sessions of our dataset, the proportion of full disengagement was 8%, moderate engagement 71%, and full engagement 21%.

We wanted to study what possible associations there are between the participants' verbal and nonverbal displays of engagement and the evaluations of the session. Complex models were used in the analyses.

The findings indicate that the more the participant is displaying full engagement through talk and nonverbal behaviors, the better ratings she/he gives for the session in all the measured variables.

On the other hand, moderate engagement was negatively correlated with the session ratings, this meaning that the more time the supervisee was only looking at other participants and only appearing to listen, but did not participate through talk or nonverbal behaviors, the poorer the ratings of the session were.

It was noticed that disengagement was not associated with any of the session rating variables. When interpreting this finding, it is important to keep in mind that the proportion of disengagement was only 8% of the sessions. All the findings are shown in Table 4.

Table 4. Standardized coefficients (full-information maximum likelihood (FIML) estimation with group and session effects controlled) of the participant level of engagement during sessions in percentage and in duration of the total session time. Strain = Felt strain at work; Quality = Relationship quality in the session; Topics = Topics and goals of the meeting; Approach = Approach and methods of the meeting: Overall = overall rating of the session.

	Strain	**Quality**	**Topics**	**Approach**	**Overall**
Disengagement %	-0.147	-0.008	-0.055	-0.039	-0.090
Disengagement	-0.148	0.009	-0.027	-0.018	-0.084
Moderate engagement %	0.041	-0.261	-0.317**	-0.214*	-0.229*
Moderate engagement	-0.057	-0.246*	-0.237***	-0.160**	-0.228**
Full engagement %	0.070	0.311*	0.386**	0.265*	0.319*
Full engagement	-0.017	0.269*	0.361**	0.268*	0.293*

* p <.05; ** p < .01; *** p <.001. All tests are two-tailed.

Results from studying interactional engagement and alliance in a smaller sample (n = 15 supervisees; two sessions from two groups) showed that full engagement during remote supervision sessions occurred in 25%, moderate engagement in 72%, low engagement in 2%, and no engagement in 1% of the time spent in remote supervision sessions (Haapanen & Hanhikoski, 2022). The sessions were selected to include both the best and worst alliance-evaluated sessions from two supervisors, aiming to create a diverse dataset in terms of alliance evaluations and to enable comparison between sessions. Due to the different interactive roles of supervisors and supervisees, the study was limited to only the supervisees.

Interestingly, almost half of full engagement was nonverbal rather than verbal communication (48%). Supervisees expressed their full engagement in the interaction not only through speech, but also through gestures and facial expressions. This is illustrated in

Figure 10. There was a statistically significant positive correlation between full engagement and alliance (r = .61), meaning that more engagement was connected with better alliance ratings.

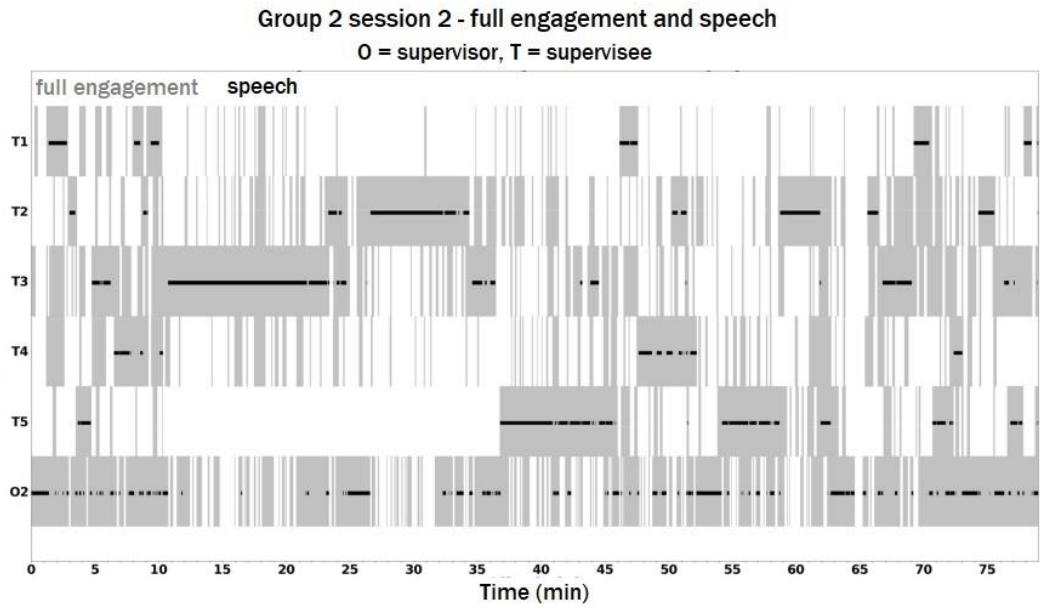

Figure 10. Full engagement and speech in Group 2's session 2. Adapted and reprinted with permission from Haapanen & Hanhikoski (2022).

3.2 Arousal state across the session

To answer our second research question, we wanted to study what can be learned about the variation of physiological stress and relaxation across the sessions. In this section, we will report the findings on the effects of relaxation periods (3.2.1), variation in the participants' arousal level in different segments of the sessions (3.2.2), and associations between engagement and heart rate variables in a smaller sample (3.2.3).

3.2.1 The effects of relaxation period

We first studied whether the brief relaxation period in the beginning of the session has an effect on the arousal state of the participants. To enable comparison between participants and to

pinpoint possible statistically significant deviations from baseline, we used (z-scored) HR values in the analysis, averaging over person and session. Data was available from altogether 23 supervisees. Our analysis indicated that in supervisees, HR was lower during the relaxation period compared to before (paired samples t-test: t (22) = 2.51, p = 0.020) and after (t (22) = 3.61, p = 0.002) the relaxation period. There was no difference between the pre and post relaxation periods, t (22) = 0.64, p = 0.530 (see Figure 11). This suggests that the decrease in HR indicative of lowered arousal state and activation of the parasympathetic nervous system was momentary and limited to the relaxation period.

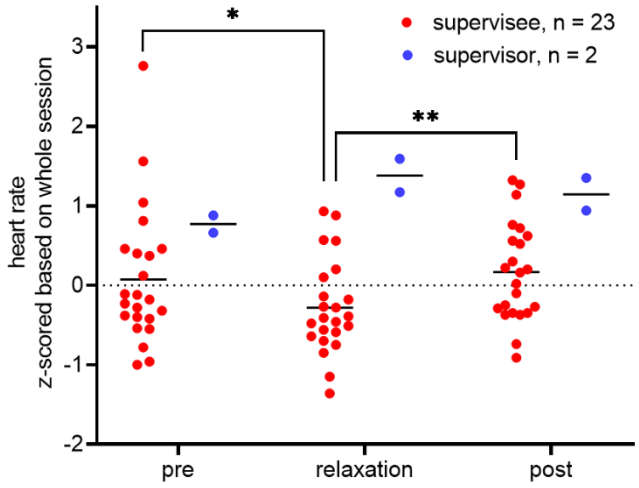

Figure 11. During the relaxation, HR slowed in supervisees. HR is z-scored to allow visualization.

3.2.2 Variation in the participants' arousal level in different segments of the session

We then studied whether the HR and HRV in participants were different at the beginning vs. the end of the session. For this we used the average HR (bpm) and HRV (RMSSD in ms) during the first vs. last 10 minutes of the session, from which data was available. That is, averages were calculated for each supervisee with valid data (n = 24), over all sessions, from which data was available. Paired samples t-test indicated no statistically significant difference in HR (t (23) = 1.77, p = 0.090) but a statistically significant increase in HRV (t (23) = 3.56, p = 0.002; see Figure 12). To summarize, arousal state seemed to decrease towards the end of the session

as indicated by the increased HRV suggesting activation of the parasympathetic nervous system.

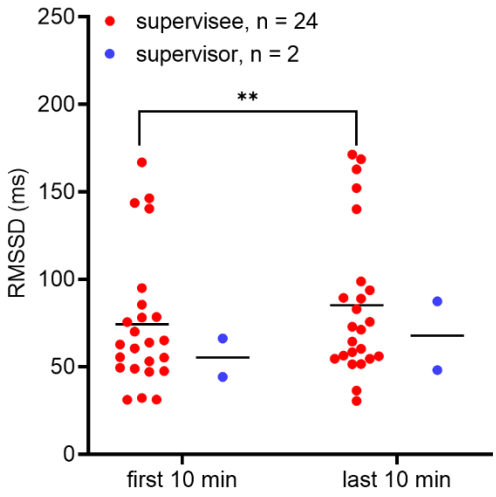

Figure 12. HRV was higher at the end of the session compared to the beginning in supervisees.

3.2.3 Associations between engagement, heart rate variables, and session ratings

In a sample of 15 supervisees from two sessions from two groups, we investigated whether supervisees' displays of engagement were associated with heart rate variables and the session ratings. Supervisees' full engagement was associated with higher heart rate (HR), greater heart rate variability (HRV), and better experienced working alliance (Haapanen & Hanhikoski, 2022). During the peak of full engagement, supervisees' heart rates were statistically significantly higher than after the peak of engagement. HRV was statistically significantly lower before than during the engagement peak, suggesting higher stress preceding full engagement.

3.3 Strain in the sessions

In this segment, we will answer our third research question: What is straining for the participants in the sessions? What happens in the interaction in segments of the session in which physiological markers of stress are observed in the participants?

In what follows, we provide a description of the technical and other problems related to the remote context (3.3.1), focus on participants' (supervisors' and supervisees') experiences of what was straining in the sessions (3.3.2), and provide a description of what happens in the interaction between the participants in episodes where at least one of the participants had low heart rate variability (3.2.3).

3.3.1 Technical problems and strain in the sessions

Based on earlier studies it was assumed that technical problems could be straining for the participants. According to our data, nearly 75% (18/24) of remote sessions encountered at least one distraction related to the online context. These distractions commonly ranged from one to three incidents per meeting and were most frequently attributed to issues with audio or video "freezing." Beyond technological hindrances, nearly a third of meetings were disrupted by distractions in participants' physical environments, such as unexpected room entries, or abrupt turning off of the lights in the room. These frequent disruptions not only interrupt the flow of conversation but also make remote meetings cognitively more demanding compared to face-to-face interactions, as participants must continually refocus on the discussion at hand.

Although in some sessions there were technical problems which lasted nearly throughout the session, such as video freezing, most often technical problems or other distractions only lasted for a few seconds rather than minutes. As a consequence, the segments were mainly too short to enable reliable calculations of e.g., changes in participants' HRV. Therefore, and even though it would have been in line with our research task to investigate variation in the participants' arousal level in relation with technical problems, we were not able to do it.

Instead, to learn more about what is straining in the remote meetings, we decided to focus on 1) to participants' open-ended responses in questionnaire, concerning what is straining in the sessions, and 2) to use qualitative analysis on the session interactions by focusing on the segments of the session in which low HRV values were observed in at least one participant.

3.3.2 Participants' experiences of what was straining in the sessions

To identify other possible sources of the strain, we decided to look at more closely the participants' responses to the open-ended question about what was straining in the sessions and where in the session this straining event took place.

We used qualitative content analysis to analyze the participants' responses. The participants reported that they felt strain in the sessions in those moments when there were problems related to technology or distractions related to working environment from where they participated in the meeting, when they felt insecurity about using technology, when they found it difficult to interpret others' responses, when they felt that is was difficult to orient and to concentrate into the meeting, when they felt that it was difficult to participate (e.g., getting a turn, having a possibility to say one's opinion on the topic at hand). Also, some participants reported strain related to being in a novel situation, talking to others and getting attention, talking about emotionally laden topics or stressful situations, and because they felt that the conversation is not useful or does not promote the creation of insights.

3.3.3 Interaction during segments in which some participants' HRV decreased

Our analysis of physiological data revealed 31 instances where participants' heart rate variability (HRV) decreased significantly, i.e., fell below a pre-set threshold of mean minus two standard deviations, — 9 instances from supervisors and 22 from supervisees. A closer examination of these moments, using video data and observable behavior, shed light on the underlying triggers.

For supervisors, HRV reductions were linked to various activities such as orchestrating transitions between different phases of the meeting, discussing their own emotionally charged experiences, and moments of confusion regarding the meeting's agenda. On the other hand, supervisees showed lowered HRV when transitioning from daily work tasks to the supervision meeting, introducing themselves to others, waiting for their turn to speak, discussing emotional experiences, withdrawing from the conversation, listening to challenging work scenarios, or while multitasking. These findings highlight the intricate physiological responses that can occur during supervised interactions, shaped by roles, behaviors, and the topics discussed

3.4 Emotional displays and cardiac activation

In this section, we will answer our fourth research question: What types of fluctuations occur in participants' arousal levels during emotionally distinct segments of the session's interactions? What facilitates meaningful dialogue concerning emotionally charged topics?

Emotional displays and cardiac activation in 28 participants (2 supervisors and 26 supervisees) were studied by selecting one session each from four groups (Henttonen & Pynnönen, 2022). In this data set, emotional expression was distributed over time in relation to neutral as follows: negative expression averaged 6.5 (SD = 9), positive 11.5% (SD = 8) and conflicting 7.7% (SD = 8). Most of the emotional expressions were neutral, 74.4% (SD = 16). There were differences in the number and distribution of emotional expressions both between groups and between participants.

Participants showed large differences in visible emotional expression depending on which group they belonged to (Figure 13). Supervisor B showed significantly more emotional expressions compared to supervisor A. The supervisor's active emotional displays seemed to play a crucial role in creating the emotional environment of supervision sessions. The supervisor explained about 66% of all the emotional expressions, 25% of negative emotional expressions, 28% for positive emotional expressions and 35% for conflicting emotions.

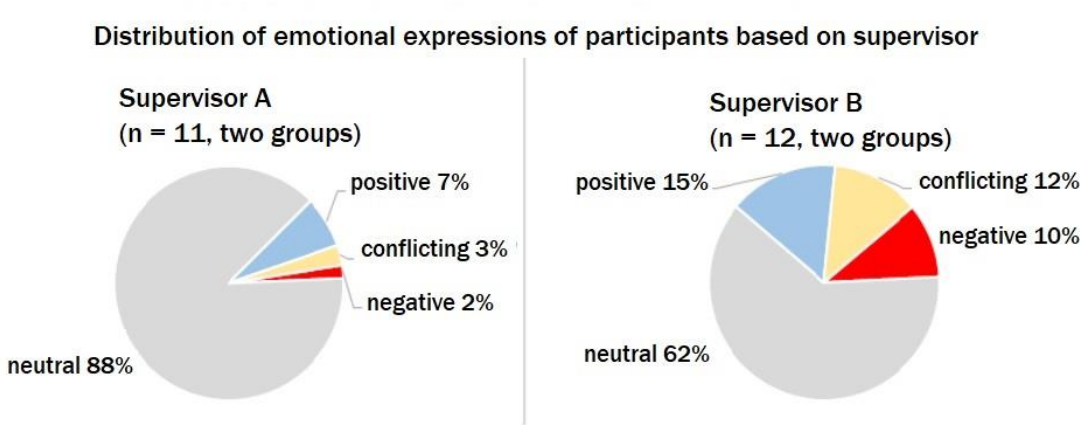

Figure 13. Differences in the visible emotional displays among participants, presented as percentage distributions. Adapted and reprinted with permission from Henttonen & Pynnönen (2022).

By qualitatively examining the graphs, several moments of emotional synchronization could be identified, of which 16 were statistically analyzed (Group 1: seven moments, Group 2: one moment, Group 3: three moments, and Group 4: four moments). Most of the moments of emotional synchronization were positive (14), one was negative, and one was conflicting.

In Group 3's session, three moments of emotional synchronization were observed (Figure 14, rectangles). The first moment of synchronization lasted 3.5 minutes, and during that time, all subjects expressed positive emotions, either smiling or laughing. In the comparison periods (3.5 minutes before and after the moment), represented in the figure with dashed lines, random emotional expressions occur, but mostly neutral emotional expression is observed. In Group 4's session, there were four moments of emotional synchronization. The longest and only negative moment of emotional synchronization unification was observed in the figure between 18–28 minutes. Most of the participants had negative emotional expressions either for an extended period (supervisor B; participants 23, 26) or multiple times (participants 20, 21, 24) during that moment.

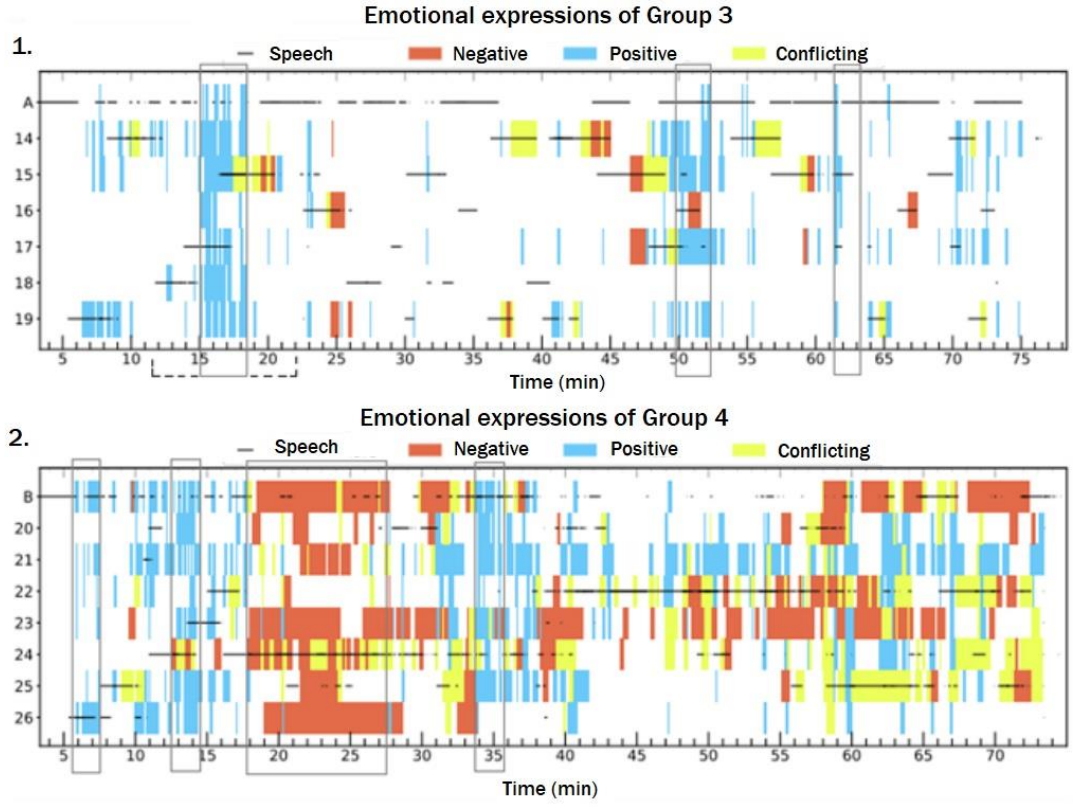

Figure 14. Moments of emotional synchronization. On the y-axis, labels 'A' and 'B' represent individual supervisors, while numbers represent identifiers of distinct participants. Adapted and reprinted with permission from Henttonen & Pynnönen (2022).

Results from the study of emotional displays and cardiac activation showed that HR and HRV were significantly higher during all emotional displays compared to the neutral state. Especially during mixed emotional displays, there was higher HR and greater HRV. Emotional expressions in general or when divided by valence were not statistically significantly associated with the session's alliance ratings.

3.5 Meaning-making, creation of a shared understanding, and activation

In this section, we will report findings from two qualitative case studies (Pohjola et al., in press.; Pohjola et al., 2023), the aim being in answering the research question 5: Is there a relationship between variations in both behavioral and physiological activation and shifts in the meaning-making process, such as the emergence of insights and the creation of a shared understanding among participants during the conversation?

The online context of the meetings may bring about challenges related to participant engagement (Bagheri & Mohamadi Zenouzagh, 2021), to collective idea generation (Nakatani et al., 2021), and to collaborative learning (Dumford & Miller, 2018). Our qualitative analyses were aimed to gain understanding on features of interaction that may promote shared reflection and generation of new understanding in online meetings and help to tackle these challenges. To answer this question, we focused on interaction in three conversational episodes the participants had rated as useful. Our qualitative analyses of the supervision interaction suggest that there were several interactional factors that promoted shared reflection and creation of new meanings in these episodes.

First, the supervisors in these episodes actively invited the supervisees to participate in the conversation and thus positioned them as active agents in the meeting (Pohjola et al., 2023). The supervisors suggested topics for the discussion only after asking the supervisees what they would like to share and work on during the session. By building the agenda on the supervisees' own interests the supervisors could ensure that the topic was meaningful for the supervisees. By creating a social order in which the supervisees were expected to take active role in the supervision meeting and to join the discussion by talking about their experiences and giving their points of view, the supervisors may have made it more inviting for the supervisees to join the discussion and less inviting to withdraw from it and start doing other tasks (Pohjola et al., in press).

Second, the supervisors' choice of words seemed to promote a social order in which thinking aloud and pondering together were the expected mode of action. This probably encouraged the participants to reflect on the case without any predetermined goal and to take multiple perspectives on the topic at hand. In the selected episodes, the supervisors also challenged the supervisees to take new perspectives. This kind of interaction can create conditions for productive difference and promote possibilities for learning (Pohjola et al., 2023).

Third, although the supervisees' speech turns tended to be allocated by the supervisors, the supervisors repeatedly invited the supervisees to reflect on each other's views and ideas (Pohjola et al., in press; Pohjola et al., 2023). This promoted conversational coherence as the supervisees became connected to each other in many ways despite not being in the same physical space (Pohjola et al., in press). For example, they could connect by sharing the same conversational topic. In addition to the inter-participant connections created by the shared topic, we observed how, by sharing the same rhythm of communication, the participants were also connected via their nonverbal modalities. Prosodically, the participants' speech rate in the selected episodes was generally slow. There was no rapid turn-taking or overlapping speech, and the interaction proceeded calmly and unhurriedly. This promoted a conversational atmosphere in which reflection, asking questions, making suggestions, and talking about one's own experiences were validated actions. During this kind of dialogue, the HRV of those participants that actively took part in the conversation by sharing their ideas and experiences related to topic at hand, increased (Pohjola et al., in press).

4. FINDINGS: PART II

In this chapter, findings from Part II of the study will be reported. The University of Oulu's research team was responsible for the Part II analyses and findings.

4.1 Measurement accuracy of the heart rate

The first task we have been working on is to develop machine learning methods which can achieve accurate and robust heart rate measurements from facial videos recorded during online group meetings. We started from exploring the traditional rPPG methods proposed in previous related studies, and then we proposed a new unsupervised method targeting to counter the challenges of the collected data, e.g., noisy data and labels, to achieve better heart rate measurement performance.

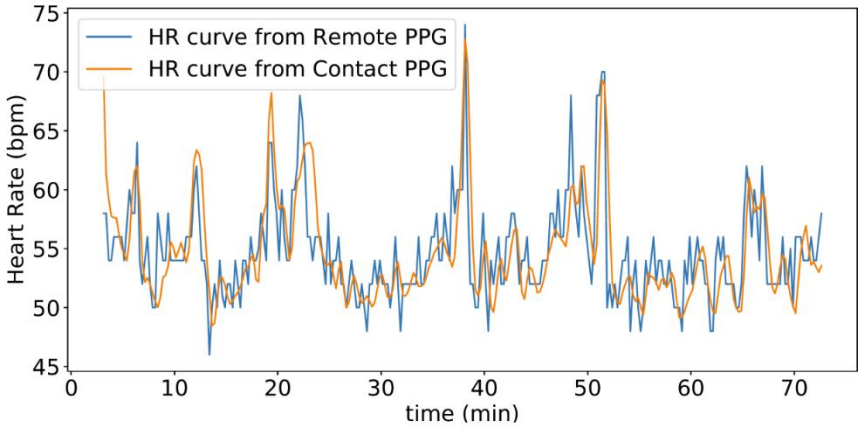

Figure 15. HR curves reconstructed from cPPG and rPPG signals.

In Figure 15, an illustration is presented, displaying an individual instance of HR curves calculated from both rPPG and cPPG signals. Within the traditional methodology, the HR for each 30-second rPPG clip was computed, subsequently contrasting it with the corresponding cPPG-derived HR. The metric for assessment was the Mean Absolute Error (MAE), root mean square error (RMSE) and Signal-to-Noise Ratio. On the other hand, within the unsupervised methodology, the comparison between the HR derived from the reconstructed rPPG signals and the ground truth cPPG was conducted. The comparison of evaluation metrics between the traditional method and the unsupervised method is presented in Table 5. It demonstrates that

the newly developed unsupervised deep learning model outperforms the traditional model in predicting heart rate. Moreover, this analysis yielded encouraging outcomes, especially considering the unregulated nature of the video recordings. In the interest of conducting a comparative analysis, mean absolute error (MAE) values of 3.55, 9.26, and 5.99 beats per minute (bpm) were attained (Sabour et al., 2023) for three distinct settings when the data recorded in controlled laboratory conditions was subjected to state-of-the-art (SOTA) performance evaluation. A significant advantage of the unsupervised deep learning model is that it does not require ground truth data for training. This allows for the application of the unsupervised deep learning model to estimate heart from any recorded facial video without additional training.

Table 5. Heart Rate prediction performance.

	Traditional methodology	Unsupervised methodology
Mean Absolute Error [bpm]	5.48	5.26
Root Mean Squared Error [bpm]	8.42	7.96
Signal-to-Noise Ratio [-]	-3.99	-3.95

4.2 Results of stress analysis

The second task we have been working on is to build a machine-learning framework that can estimate a subject's stress level by automatic analysis of the input video. We considered using different sets of clues, i.e., physiological features from rPPG signals or cPPG signals, head motions, and facial expressions, and we also considered the fusion of two or three modalities in order to achieve more reliable estimation.

Table 6. Stress classification results for a single set of features (first panel), fusion of rPPG with facial expression and motion features (second panel), and fusion of cPPG with facial expression and motion features (third panel).

Features	Acc (%)	AUC
Motion	70.00	0.70
Facial Expression	73.75	0.71
cPPG	**81.25**	**0.80**
rPPG	78.75	0.78

Features	Acc (%)	AUC
rPPG	78.75	0.78
rPPG+Motion	78.75	0.78
rPPG+ Facial Expression	81.25	0.81
rPPG+Motion +Facial Expression	**82.50**	**0.81**

Features	Acc (%)	AUC
cPPG	81.25	0.80
cPPG+Motion	81.25	0.80
cPPG+ Facial Expression	83.75	0.82
cPPG+Motion +Facial Expression	**85.00**	**0.83**

Throughout this experiment, a total of 36 high-stress and 44 low-stress samples were obtained. The threshold value of 7, which represents the median stress level reported by the participants, was employed. In the first column of Table 6 panel one, one can find the presentation of results from a stress classification task that relied on a single set of features. Notably, the highest accuracy (81.25%) and AUC (0.80) were attained in predicting stress levels through the utilization of cPPG information. Comparatively, the task's performance, which was based on rPPG features, yielded slightly lower results, with an accuracy of 78.75% and an AUC of 0.78. Additionally, when employing facial expression features, an accuracy rate of 73.75% and an AUC of 0.71 were observed. Ultimately, the least favorable outcomes were yielded by motion features, resulting in an accuracy of 70.00% and an AUC of 0.70. The presentation of feature fusion outcomes can be found in Table 6 panel two, specifically in the second and third columns. When the fusion of rPPG, motion, and facial expression features was undertaken, a marginal enhancement in both accuracy and AUC was observed. A similar occurrence was noted when combining cPPG with motion and facial expression features, which are shown in Table 6 panel three. We can answer the following three research questions based on the results achieved.

Is the utilization of rPPG signals, as measured from facial videos, a feasible approach for stress estimation? The superiority of stress estimation, rooted in rPPG features, is evidenced by its accuracy rate, standing at 78.75%, surpassing that achieved through motion features, which registers at 70.00%, and facial expression features, which exhibit a comparatively lower accuracy rate of 73.75%. Stress estimation relying on motion and facial expression features is characterized by reduced accuracy, primarily attributed to two underlying factors. In the first instance, neutral facial expressions and minimal motions are commonly observed among participants during online video meetings, rendering facial expressions and motions subtle, brief, and infrequent (Navarathna et al., 2017). In the second scenario, the regulation of facial expressions and motions during states of stress is a self-controlled process, whereas physiological signals remain beyond individual volition, as expounded in references (Yu et al., 2021; Kortelainen et al., 2012).

How does the accuracy of stress estimation based on rPPG and cPPG features differ? The accuracy of stress estimation features derived from rPPG (78.75%) is found to be in proximity to those yielded by cPPG features (81.25%). Nevertheless, stress estimation through rPPG signals is deemed more convenient, primarily due to the ubiquity of facial videos in online video meetings. The utilization of cPPG signals for stress assessment necessitates pulse oximeters, a practical implementation of which is generally considered challenging.

Is stress estimation enhanced by the fusion of behavioral features (facial expression and motion) with features derived from physiological signals (rPPG or cPPG)? The integration of behavioral features with physiological signal features (as illustrated in Table 1's center and right columns) can lead to a slight improvement in stress estimation accuracy. Despite the relatively lower effectiveness of standalone behavioral features in stress estimation, the practice of fusion remains beneficial, as it permits the potential complementation of different features for the purpose of stress assessment.

Following the three main results findings, we further explored one more question: ***are some HRV features more important for stress estimation than the others?*** To answer to this question, the following 24 HRV features split into three groups were considered:

- Time domain features: mean NNI (mean inter-beat interval (IBI)), SDNN (standard deviation of IBI), SDSD (standard deviation of successive), pNN50 (percentage of samples with more than 50 ms difference from the consecutive beat), pNN20 (percentage of samples with more than 20 ms difference from the consecutive beat), NN50 (the number of samples with more than 50 ms difference from the consecutive beat), NN20 (the number of samples with more than 20 ms difference from the consecutive beat), RMSSD (The root mean square of successive differences of IBI), median NNI (median of IBI), range NNI (range of IB), CVSD (the coefficient of variation of successive differences), CVNNI (the coefficient of variation), mean (HR mean heart rate), max HR (maximum heart rate), min HR (minimum heart rate), std HR (the standard deviation of heart rate)
- Frequency domain features: LFnu (the power in the low frequency (0.04Hz-0.15Hz) with normalized unit), LF (freq the low frequency peak), HFnu (the power in the high frequency (0.15Hz-0.4Hz) with normalized unit), HF freq (the high frequency peak), LF/HF (the ratio of LFnu and HFnu)
- Geometrical domain features: SD1 and SD2 (Poincare´ plot standard deviations), SD1/SD2 (SD1 and SD2 ratio).

The robustness in the stress estimation task is demonstrated by the physiological features, and we ranked the features according to their contributions when training the classifier for the task of stress estimation. The feature importance is displayed in Figure 16. The highest ranks were achieved by several frequency features, including LF Freq, HF Freq, LFnu, and HFnu. The importance of each feature is quantified using the permutation feature importance method.

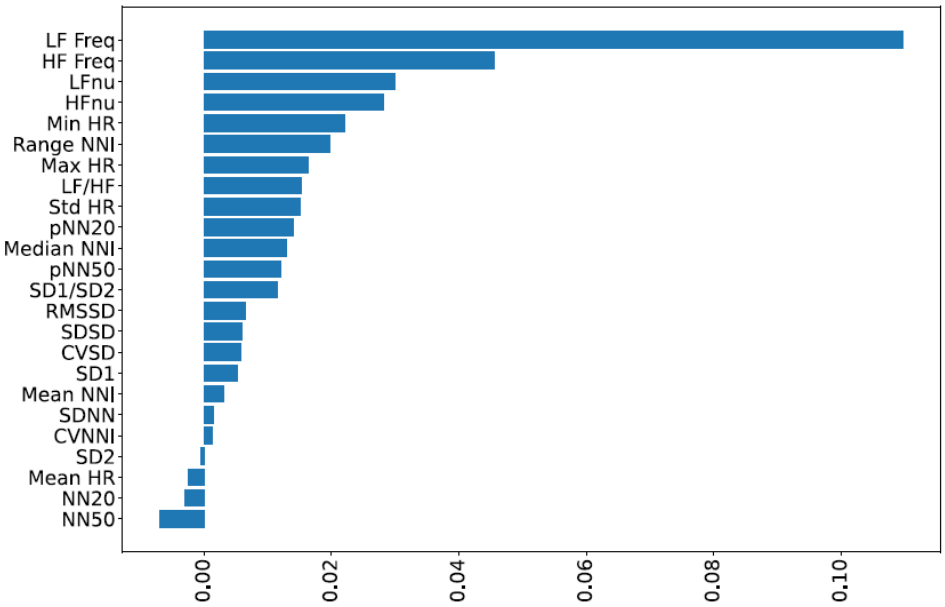

Figure 16. Importance ranking of HRV Features calculated from rPPG signals for stress estimation.

Main observations of this study can be summarized as follows:

Observation 1: rPPG features surpass motion and facial expression features in stress estimation, showcasing higher accuracy rates due to their involuntary nature.

Observation 2: Stress estimation can be enhanced by the fusion of behavioral and physiological features.

Observation 3: Among all the HRV features, the frequency domain features are more important in stress estimation.

4.3 Results of the engagement analysis

The third task we have been working on is to develop a machine learning framework that can measure the subject's engagement levels during the online meeting via analyzing his/her facial video. The task is solved in two phases: firstly, we explored a suitable setup and approach for utilizing reconstructed rPPG signals for engagement measurement; secondly, we further improved the measurement performance by testing and selecting the most suitable set of behavioral feature set to add to the physiological feature sets extracted from rPPG signals.

An unsupervised approach of the pre-trained rPPG model was undertaken to acquire rPPG signals from video conferences, facilitating the derivation of heart rate variability (HRV) features. Further, the impact of varying HRV observation window sizes was analyzed. Through this analysis, the potential for utilizing short observation windows to infer engagement levels based on HRV features was illustrated, displaying a marked improvement when observation windows were expanded to durations ranging from two to four minutes. Following this, an evaluation was conducted on the proficiency of behavioral cues when integrated with physiological data, a process that notably improved the precision of engagement prediction.

Is engagement estimation feasible when using HRV features calculated using short observation windows from the reconstructed rPPG signal? In the gathered dataset, the suggested method's efficacy has been demonstrated to be as much as 49.40% accurate when assessed through the KNN model, utilizing a brief 10-second window for the computation of HRV attributes. Dependable HRV features are not rendered through a short observation window; instead, an extended video duration is essential for a reliable analysis of HRV and engagement. The full potential of the proposed technique is facilitated through the utilization of extended HRV observation windows in the video recordings analyzed in this research.

Is it possible to enhance the accuracy of engagement estimation by using longer observation windows for HRV feature calculations? If so, by how much can it increase? Employing extended observation windows indicated that for the task of identifying engagement, the most potent machine learning classifiers were the KNN and the combination of KNN and Random Forest ensemble (KNN+RF). Table 7 demonstrates the evaluation of the models' efficacy over diverse HRV observation window values. A marked increase in both accuracy and ROC AUC scores was documented when the HRV observation window duration expanded from 60 to 240 seconds.

Table 7. Comparing KNN and KNN-random forest ensemble models for engagement prediction based on various HRV observation window durations: evaluation metrics.

HRV window size [sec]		60	90	120	150	180	210	240
KNN	Accuracy [-]	.816	.882	.910	.924	.931	.937	.937
	ROC AUC [-]	.870	.923	.946	.953	.955	.961	**.960**
KNN + RF	Accuracy [-]	.816	.880	.911	.926	.938	.938	.940
	ROC AUC [-]	.906	.947	.967	.975	.983	.983	**.983**

A notable influence on the efficacy of models predicting engagement is noticed when altering the size of the HRV observation window; enhanced results are achieved with increased window dimensions. Employment of longer observation windows provides more sophisticated and detailed HRV patterns within the rPPG signals, thereby allowing for the generation of predictions that are both more reliable and exact. Additionally, it is identified that over extended periods of observation, the disruptions and transient variations in the signals, which might hinder or reduce the functionality of the classifiers, are more adeptly mitigated. As highlighted in Tab. 7, utilizing a 120-second HRV observation window sufficiently allows the depicted method to display its potential.

Is engagement estimation enhanced by the fusion of behavioral features (facial expression and motion) with features derived from rPPG signals? The inclusion of behavioral features emerges as a vital step for refining assessment metrics, considering that the spectrum of human engagement is influenced not only by physiological nuances but also by understated behavioral cues. Therefore, adopting a comprehensive strategy that adeptly combines both the subtle behavioral indications and the distinct physiological components into a singular model, could potentially serve to notably enhance the reliability and depth of predictions, offering a more nuanced and accurate understanding of human engagement dynamics.

In Table 8, the comprehensive analysis studies the influence of various behavioral feature combinations on the accuracy of the combined KNN and Random Forest predictive model is extensively exhibited. These selected combinations encompass the following aspects: (1) Thorough monitoring of the gaze patterns, providing insights into the individual's focus of attention, (2) Determination of the gaze direction angle, which offers nuanced information on the exact angle and orientation of the gaze, (3) Pinpointing 2D and 3D landmarks at specific,

strategically chosen locations surrounding each eye, facilitating detailed analysis of eye movements, (4) Assessing the 3D space head translations and rotations, an aspect that helps in understanding the spatial orientation and movements of the head in a three-dimensional space, and (5) Analyzing Facial Action Units, which encompasses the detailed study of various facial movements and expressions indicative of underlying emotions or reactions. It is empirically demonstrated that the careful selection of behavioral features has a considerable and marked impact on the efficacy of the predictive model. Specifically, set (3) is identified as less effective during extended HRV observation windows, as opposed to the other sets. Meanwhile, it has been observed that sets (4) and (5) significantly enhance the accuracy of predictions, potentially leading to a more robust and reliable model. A noteworthy improvement in metrics is witnessed as the model's performance experiences an increase from 0.816 to 0.930 for a 60-second observation window when behavioral features are integrated. Enhanced performance attributed to the inclusion of BFs is consistently detected within the entire dataset. However, the magnitude of this enhancement is more prominently observed with smaller HRV observation window sizes.

Table 8. Comparing the performances of the KNN-Random Forest ensemble model for engagement prediction, based on various sets of behavioral features, while utilizing different lengths of HRV observation windows.

HRV window size [sec]	60	90	120	150	180	210	240
HRV	.816	.880	.911	.926	.936	.938	.940
HRV + BF (1)	.864	.903	.929	.934	.937	.939	.943
HRV + BF (2)	.855	.902	.929	.936	.944	.944	.949
HRV + BF (3)	.861	.870	.881	.884	.892	.896	.899
HRV + BF (4)	.896	.936	.952	.953	.954	.955	.956
HRV + BF (5)	.919	.931	.944	.944	.947	.951	.956
HRV + BF (1+2)	.856	.891	.919	.927	.932	.933	.934
HRV + BF (4+5)	.928	.942	.952	.954	.955	.958	**.960**
HRV + BF (all)	.930	.929	.934	.940	.938	.941	.940

Main observations of this study can be summarized as follows:

Observation 1: utilizing longer HRV observation windows in video recordings can significantly enhance the accuracy of engagement estimations. Extended observation windows,

ranging from 120 to 240 seconds, have shown to improve both the accuracy and ROC AUC scores in identifying engagement using KNN and KNN+RF classifiers.

Observation 2: The fusion of behavioral features, which analyze head movements and facial action units, significantly amplifies the accuracy and reliability of engagement analysis in the predictive model, especially within smaller HRV observation window sizes.

5. SUMMARY AND DISCUSSION

In this chapter, we will summarize and discuss our findings. Findings from Part I will be discussed in section 5.1 and findings from Part II studies in section 5.2. We will address strengths, limitations, and our suggestions for further studies in section 5.3, and suggestions for practitioners in section 5.4.

5.1 Part I

At the conversational level, we looked at the variation of activation across different roles; a) supervisor vs. supervisee, b) talker vs. listener. We noticed that in the work group supervision sessions of our data, speech turns tended to be organized via supervisor. Similar findings from other types of online conversations (Nakatani et al. 2021) point to this same phenomenon; something in videoconferences seems to raise the threshold for participation. Yet, active participation would be important: in our study, participation through talk and displaying your engagement also through nonverbal behaviors was significantly correlated with the ratings of the session.

Furthermore, we noticed that the level of strain before the session is important. When the supervisee evaluated his/her work at the time of the meeting more straining, he/she also rated the meeting less favorably. The working relationship between the participants, and the methods used by the supervisor were rated lower when the supervisee's work situation was more straining.

The participants experienced strain in the meetings can be categorized into three categories. Some of the strain that was experienced by the participants in the meetings were clearly related to online context of meetings such as technological problems; these fall into situations creating over arousal as depicted in Figure 17. Other issues that the participants mentioned as straining were possibly, but not necessarily, related to online context such as difficulties in orienting and concentrating into the meeting, which may be associated with under arousal. Then there were also those issues that were felt as straining, but that were not related to online context of the meeting but would have most likely happened also in face-to-face

meetings. These were, for example, becoming nervous about being in a novel situation or about talking to others and getting attention.

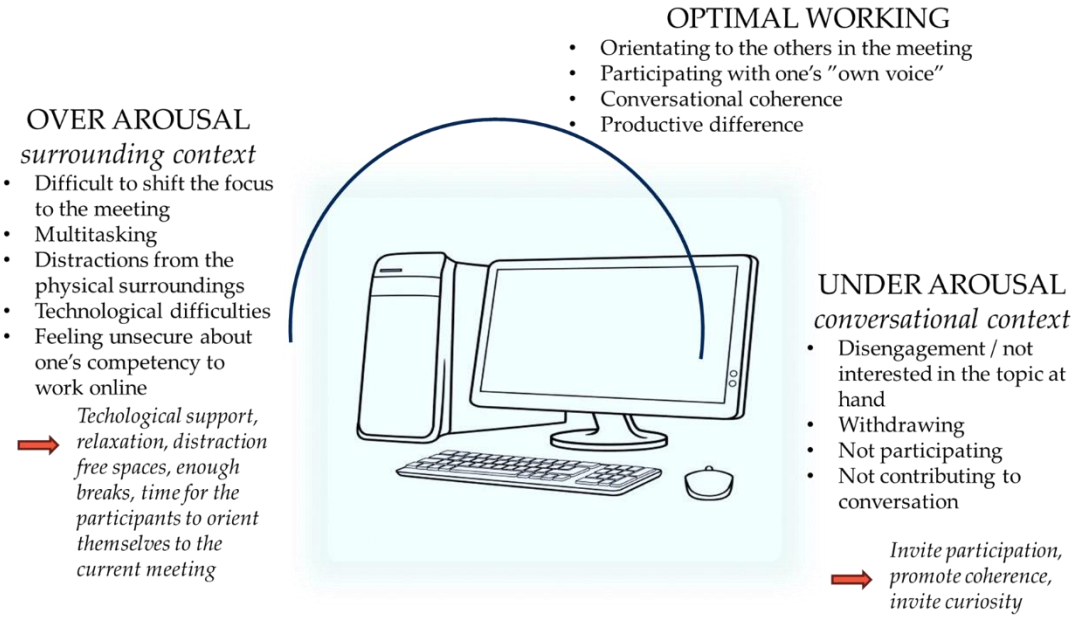

Figure 17. Working during over arousal, optimal working and under arousal conditions and suggested solutions to overcome challenges by over arousal and under arousal.

An important finding was that short resting periods calmed down the supervisees. Only a resting period of two minutes was enough to physiologically calm down supervisees. This was evidenced as their HR decreased. Based on the questionnaires, this resting period was also important and useful for many supervisees. A brief resting period or corresponding exercise at the beginning of the meeting can thus be recommended – especially if there is a lot of strain at work at the time of the meeting.

One of the main findings of this study was that for the success of a meeting, active verbal and nonverbal participation in discussion is beneficial. The more the supervisees talked during the meeting, the better they rated the meeting, especially what comes to the aims and topics of the meeting. The same was true with showing your engagement in the meeting also nonverbally. In the engagement coding, full engagement included both verbal and nonverbal behaviors which supervisees used to display their interest and participation in the meeting interactions. At the level of the entire sample, those who spent more time showing full

engagement rated the session more favorable in all variables, i.e., aims and topics, working relationships, methods, and overall rating. On the other hand, the more supervisees showed only moderate engagement, this is they just looked at the other participants and looked like they were listening, but did not show their engagement actively, the worse they rated the session in all variables.

In a study by Haapanen and Hanhikoski (2022), where four supervisory sessions were examined from two groups based on low and high session ratings, it was observed that supervisees exhibited physiological relaxation while speaking. In this case study, supervisees' full engagement was observed to be connected with higher HR, higher HRV, and better alliance experience. This finding is interesting as it seems to suggest simultaneous activation in the sympathetic and parasympathetic parts of the ANS. This, in turn, would be in line with the idea of an optimal window of physiological activity in which relaxed and active state are simultaneously present. However, it is important to keep in mind that this finding was based on a very limited sample. Further studies with larger samples are therefore needed.

Moreover, case level observations of moments of generation of new knowledge and even insight were made in our research. We noticed that when several supervisees are involved in the interaction by showing actively their engagement, the conversation is more coherent, there are more productive differences, which are considered important for the generation of new knowledge in the interaction, and the supervisees also use more ”ownership talk” (Kykyri et al., 2010), i.e. talk with their own voice on their own behalf in a matter relevant for them.

On the other hand, when there is more disengagement in the group, there is less coherence in the conversation, participants use less ”ownership talk” and the conversation is less generative, i.e., less new meanings are generated. Since these findings are based on a limited sample of sessions, findings should be interpreted with caution.

In our sample, supervisees were on average very satisfied with the sessions, even though technical and other problems related to remote mode were at least as common as earlier studies have reported from all kinds of remote meetings. When interpreting our findings, it is important to keep in mind that the research context was work group supervision. The remote meetings were thus organized and facilitated by trained supervisors, which most probably affected the supervisee's experiences of the meetings. Still, it is possible to conclude that

following some of the good procedures reported in this study, quality of other types of online meetings can be increased.

5.2 PART II

Three major findings can be summarized from the Oulu team's results. Firstly, we have demonstrated that the proposed unsupervised rPPG approach achieves superior performance on heart rate measurement. It has been found that a notable difference in accuracy exists between the traditional methodology and the unsupervised deep learning model in predicting heart rates from the collected online group meeting videos. This indicates that the unsupervised model, which does not require ground truth labels for training, is more efficient and avoids the impact of noisy ground truth signals than the previous models. The enhanced performance of the unsupervised model is attributed to its advanced algorithms based on contrastive learning. These are particularly beneficial given the unregulated nature of video recordings used in these analyses. In conclusion, the unsupervised deep learning model emerges as a promising tool for real-world applications by offering enhanced accuracy and eliminating the need for ground truth data, a significant limitation of traditional models.

Secondly, the study evaluated the effectiveness of cPPG, rPPG, motion, and facial expression features in stress estimation. It was found that the rPPG approach is able to achieve almost comparable accuracy of stress estimation compared with the cPPG approach, while the rPPG features, extracted from facial videos, provides a much more convenient and comfortable way especially for long-duration monitoring. The motion and facial expression features, while less effective on their own, enhance stress estimation when combined with physiological signals, underscoring the value of a multimodality fusion approach. Frequency domain HRV features were particularly significant in stress estimation. As a conclusion, integrating behavioral and physiological features is beneficial for a more comprehensive and nuanced understanding of stress levels, highlighting the promising role of rPPG features extracted from easily accessible facial videos.

Thirdly, it was found that the accuracy of engagement estimation is significantly impacted by the length of HRV observation windows. A suitable window length is within 120 to 240 seconds, which produced the best results. A longer window length (>240 s) will not

further improve the accuracy, while if the window length drops shorter than 120s the performance declines. The incorporation of behavioral features, including gaze patterns, head movements, and facial expressions, further enhances the model's prediction accuracy. This means that the combination of physiological data and behavioral cues offers a more comprehensive and nuanced understanding of human engagement dynamics. As a conclusion, an integrated approach that integrates both HRV features extracted from a suitable observation window and a diverse set of behavioral features serves as an effective strategy for accurate and reliable engagement estimation in video conferencing contexts.

5.3 Strengths, limitations, and needs for the future research

The field study design in our research project has been both an asset and a limitation. Earlier research on remote meetings has often focused on experimental settings with limited designs. E.g., "team" may be just two or three participants, who often are students who are asked to complete somewhat artificial tasks. Studying remote meetings at the real-life workplace settings increases the natural validity of the research. Our dataset is unique and valuable, since it consists of 24 sessions in groups of multiple trained professionals, who discuss topics highly relevant for their work with an experienced coach. Moreover, in the field of computer vision and machine learning studies, a good sample of 90-minute video recordings can be considered highly valuable.

On the other hand, there was complexity in the field study setting, which did not enable controlling the design, leading to missing or poor-quality data and problems with the analyses. E.g., highly emotional segments of the sessions were most often too short to enable analyses of HRV, which usually requires at least two-minute segments for the reliable calculations.

In the future, controlled experiments would also have their place. For example, to study what is better, camera open or closed, or what is the role of time lags, and how straining the technical problems are for the participants, controlled experiments would be useful. We would also recommend including skin conductance (SC) as a measure of sympathetic nervous system activity in the design. In studies which are aimed at illuminating the variation in physiological activity across different styles or patterns of participation, skin conductance would allow capturing rapid changes in the participants' arousal. When integrating

physiological variables into what happens at the conversation, this would be important since in two minutes period, which is needed for HRV calculation, there happens a lot in the conversation, whereas skin conductance responses can be located in episodes of only 10-20 seconds.

Another limitation of our study was that in our design, face-to-face meetings were not gathered, so we did not have a possibility to make any comparisons between the two contexts. Some groups in our sample might have been interested in meeting also face-to-face, but this was not possible, due to the Covid-19 pandemic.

In our study, the groups were not fixed, so it would not have been meaningful to compare. In future studies, however, it is advisable to compare remote meetings and face-to-face meetings of the same teams. Then, it would be possible to investigate whether team members' experiences of collaboration and workplace communality in teams who meet only in remote mode differ from that of the members of the teams that mainly meet face-to-face. Another, and also very interesting setting would be included in the design also hybrid meetings, in which some workers meet in a face-to-face setting while some of the participants join the meeting online.

From the perspective of computer vision research, the task in this project is challenging for developing robust machine learning methods. Since we consider practical online meeting scenes, the video data and the ground truth labels were recorded by the subjects themselves thus inevitably contain lots of noises which impact the model performance. Our proposed solution is an unsupervised learning approach pretrained on other datasets which doesn't require any label from the current dataset. It helps with the noisy label problem and outperformed previous supervised methods. We expect the approach can be further refined to achieve even higher accuracy, and we are working on extending the current method into a semi-supervised approach which can leverage helpful information from the noisy labels.

Furthermore, we also plan to explore the possibility of reconstructing other physiological signals from facial videos, such as blood pressure (BP) and ox saturation (SaO2). In the current study, we focused on measuring the heart rate, heart rate variability, and breathing rate, which can be calculated from the reconstructed ted PPG signals. BP and SaO2 are related to the blood volume pulse but require more sophisticated measurement, e.g., multiple tracks of simultaneous measurement. We plan to explore more advanced rPPG methods in the following

in order to measure BP and SaO2 levels which are also important indicators for people's health and psychological status.

Natural and also highly interesting next step in the research would be a study, which integrates biofeedback of the participants' heart rate into a design that compares remote meetings with face-to-face meetings. This would be possible in the near future, based on the rPPG method developed in Part II of this study. It will be possible to integrate basic research aims with software that already can be helpful for the practitioners.

5.4 Observations and reflections on the role of the work supervisor

5.4.1 General observations

Like all participants, the work supervisor also receives fewer interactive cues in a remote conversation than in a face-to-face encounter, where spontaneous interaction can be observed and sensed.

In remote work setting, the content of the conversations is easily emphasized at the expense of qualitative factors. The project made visible the importance of qualitative factors of interaction in remote working situations.

One key observation related to the quality of the interaction; the tendency of discussions to be organized via the supervisor, rather than flowing freely between the group members. The supervisees address their answers easily to the supervisor and not to each other or to all participants. This reduces the spaces and opportunities for the group's multilateral discussion and the meanings that are produced together.

5.4.2 Learning points for supervision practice

When preparing for work supervision sessions, it is useful for the supervisor to recognize his own state of alertness in order to be sensitive to observing the participants' arousal and state of alertness.

For the supervisee, participating in a work supervision session in the middle of a busy working day can make it difficult to focus on one thing. It is important for the supervisor to remind the supervisees to close other applications and focus on a shared context and task. Instructing the supervisees to exclude other things that attract their attention, such as closing the software that is not needed in work supervision, is one of the supervisor's tasks.

Findings of the project showed that a brief shared silent moment promoted reaching an optimal state of activity in the beginning of the supervisory sessions. In addition to the beginning, the instructor should monitor how the participants' activation state develops during the session and take regular breaks especially in longer sessions.

It was known earlier, and observed also in this project, that remote group meetings are prone to technical disturbances. In supervisor's role, it is useful to treat them as relevant. The supervisor's calmness in facing technical problems and other troubled situations is transmitted to the group members, reducing the stress load and helping to focus on a shared task despite the technological difficulties.

5.4.3 Conclusions and suggestions for the practitioners

Firstly, awareness of the interactional bias produced by the supervisor-centeredness helps to change the practice to increase the initiative and active participation of the supervisees and the interaction between the supervisees in the group. The instructor can do this by asking, for example:

- what did you (addressed to the group) hear in one participant's turn?
- what personal experiences did the description/talk evoke in you? This question can promote authentic speech through personal sharing (told here and now and shared with these people only).
- have you discussed this with anyone involved now?

The instructor can also produce low-threshold questions at the beginning of the sessions to get all the supervisees involved. For example, "where are you at the moment and what does this place mean to you as a place to do your work?"

Secondly, the supervisor can modify the structures of participation to strengthen the participants' agency and contribution to the supervision work during the session. For

example, the instructor can ask only three people from the group to discuss the topic that came up, while the others are listening. The supervisor could even ask the other supervisees to close their videos for that time. After 5 minutes, the supervisor can continue by asking the next three people to continue the conversation, based on what has already been started by the first three supervisees. Others close their videos and just listen.

Another method to facilitate dialogue between the supervisees is to start the session by using a "circulating the microphone around the group" exercise. This refers to a practice where the current speaker, after her/his turn is completed, selects, and invites the next speaker to take the floor. In this way, the immediate experiences of the current work can be shared at the beginning and the needs of the participants can be explored for the supervision session.

By implementing these recommendations, supervisors can significantly enrich the quality of remote work supervision, thereby fostering a more engaging and effective environment.

6. CONCLUSIONS

Remote work group meetings are widely used in organizations, but these can be straining for the participants and leaders. It is highly important to know what can be done to make remote meetings more useful. In the PhinGAIN research project, we have provided both empirical evidence and descriptions of how strain could be reduced and gain of the meetings increased by focusing on both verbal and nonverbal behaviors and the participants' physiological responses. Moreover, we have been able to develop new technical solutions to be used in future applications during remote meetings.

Our results show that participants' active engagement in the meeting interactions both verbally and nonverbally is of essential importance for the success of the meetings. All who participate in remote meetings can take responsibility for this issue on their own part. However, the role of a leader or a facilitator is crucial in creating space for and inviting active displays of engagement from the participants.

Moreover, our results reveal that by using relatively small interventions, strain in remote meetings could be reduced and an optimal state of activation be achieved. Only a two-minute resting period with eyes closed was shown to be enough to calm down the participants, as was evidenced in their lowered heart rate. Well-organized and thoroughly facilitated remote meetings were rated useful by the participants, and our evidence from the entire sample shows that participants' physiological state relaxed towards the end of the session. Taking into account the increasing concern regarding the wellbeing of workers in turbulent and constantly changing working environments, as well as the large number of remote meetings at workplaces, the new knowledge gained in our project is highly important. Therefore, we recommend organizations implement our suggestions for practitioners, which are presented in section 5.4, and also in Appendix 3.

From the technical perspective, to remotely measure the heart rate from facial videos recorded during online meetings is a big challenge for current computational models, as the self-collected videos and ground truth label signals contain lots of noises. We proposed a contrastive learning-based unsupervised method which can achieve superior performance on heart rate measurement without using any ground truth label signals thus avoiding the impact

of noises. We also developed two machine learning frameworks which can utilize both physiological features calculated from the reconstructed rPPG signals and behavioral features extracted from facial videos for analyzing emotional status. Our results showed that physiological features extracted from rPPG signals can work almost as well as those from cPPG signals for the task of stress level estimation, indicating the feasibility of using the remote approach of monitoring physiological responses via a camera. We also demonstrated that for the purpose of estimating subjects' engagement levels during the meeting, an observation window of two to four minutes is the most suitable for extracting rPPG signals for HRV analysis, and additional behavioral features sets of head rotations and facial action unit trajectories will further boost the performance of engagement estimation. Last but not the least, we have implemented one of our proposed rPPG methods as a mobile application, which can run on any Android phone to make real-time heart rate measurement using the video captured by the inbuilt front camera.

REFERENCES

Abbass, A., & Elliott, J. (2021). Emotion-focused and video-technology considerations in the COVID-19 crisis. *Counselling Psychology Quarterly, 34*(3–4), 624–636. https://doi.org/10.1080/09515070.2020.1784096

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement, 28*(3), R1–39. https://doi.org/10.1088/0967-3334/28/3/r01

Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using Zoom Videoconferencing for Qualitative Data Collection: Perceptions and Experiences of Researchers and Participants. *International Journal of Qualitative Methods, 18.* https://doi.org/10.1177/1609406919874596

Bagheri, M. & Mohamadi Zenouzagh, Z. (2021). Comparative study of the effect of face-to-face and computer mediated conversation modalities on student engagement: speaking skill in focus. *Asian-Pacific Journal of Second and Foreign Language Education, 6*(1), 1–23.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018). Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018),* 59–66. IEEE. https://doi.org/10.1109/FG.2018.00019

Blanck, P., Stoffel, M., Bents, H., Ditzen, B., & Mander, J. (2019). Heart rate variability in individual psychotherapy: Associations with alliance and outcome. *The Journal of Nervous and Mental Disease, 207*(6), 451–458. https://doi.org/10.1097/NMD.0000000000000994

Cromby, J. (2012). Feeling the way: Qualitative clinical research and the affective turn. *Qualitative Research in Psychology*, *9*(1), 88–98. http://dx.doi.org/10.1080/14780887.2012.630831

Dumford, A.D., & Miller, A.L. (2018). Online learning in higher education: exploring advantages and disadvantages for engagement. *Journal of computing in higher education, 30*, 452–465. https://doi.org/10.1007/s12528-018-9179-z

Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Reynolds, L. R., Brown, J., & Johnson, L. D. (2003). The Session Rating Scale: Preliminary psychometric properties of a "working" alliance measure. *Journal of brief Therapy*, *3*(1), 3–12. https://api.semanticscholar.org/CorpusID:26848800

Friesen, E., & Ekman, P. (1978). Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2), 5.

Fuchs, T., & Koch, S. C. (2014). Embodied affectivity: on moving and being moved. *Frontiers in psychology*, *5*, 508. https://doi.org/10.3389/fpsyg.2014.00508

Girard, J. M., & Wright, A. G. C. (2018). DARMA: Software for dual axis rating and media annotation. *Behav Res* 50, 902–909. https://doi.org/10.3758/s13428-017-0915-5

Goffman, E. (1957). Alienation From Interaction. *Human Relations*, *10*(1), 47–60.

Goffman, E. (1963). *Behavior in Public Places: Notes on the Social Organization of Gatherings.* New York, NY: Free Press.

Haapanen, N., & Hanhikoski, N. (2022). *Vuorovaikutukseen sitoutuminen, vireys ja allianssi videovälitteisissä työnohjausistunnoissa* (Master's thesis, University of Jyväskylä). http://urn.fi/URN:NBN:fi:jyu-202206223543

Henttonen, J., & Pynnönen, P. (2022). *Tunneilmaisut, yhteistyösuhde ja autonominen hermosto etävälitteisessä työnohjauksessa* (Master's Thesis, University of Jyväskylä). http://urn.fi/URN:NBN:fi:jyu-202209064482

Hepburn, A., & Jackson, C. 2009. Rethinking subjectivity: A discursive psychological approach to cognition and emotion. In: Dennis Fox, Isaac Prilleltensky & Stephanie Austin (eds.), *Critical psychology: An introduction,* 188–219. London: Sage.

Hoogeboom, M. A., Saeed, A., Noordzij, M. L., & Wilderom, C. P. (2021). Physiological arousal variability accompanying relations-oriented behaviors of effective leaders: Triangulating skin conductance, video-based behavior coding and perceived effectiveness. *The Leadership Quarterly*, *32*(6), 101493. https://doi.org/10.1016/j.leaqua.2020.101493

Jänig, W. (1989) Autonomic Nervous System. In: Robert F. Schmidt, Gerhrad Thews (eds.), *Human Physiology*, 333–370. Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-73831-9

Kilcullen, M., Feitosa, J., & Salas, E. (2021). Insights from the virtual team science: Rapid deployment during COVID-19. *Human Factors: The Journal of the Human Factors and Ergonomics Society 64*(8). https://doi.org/10.1177/0018720821991678

Kortelainen, J., Tiinanen, S., Huang, X., Li, X., Laukka, S., Pietikäinen, M., & Seppänen, T. (2012). Multimodal Emotion Recognition by Combining Physiological Signals and Facial Expressions: A Preliminary Study. In *Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5238–5241). IEEE. https://doi.org/10.1109/embc.2012.6347175

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological psychology*, *84*(3), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

Kykyri, V. L., Puutio, R., & Wahlström, J. (2010). Inviting participation in organizational change through ownership talk. *The Journal of Applied Behavioral Science*, *46*(1), 92–118. http://dx.doi.org/10.1177/0021886309357441

Lampinen, E., Karolaakso, T., Karvonen, A., Kaartinen, J., Kykyri, V. L., Seikkula, J., & Penttonen, M. (2018). Electrodermal activity, respiratory sinus arrhythmia, and heart rate variability in a relationship enrichment program. *Mindfulness*, *9*, 1076–1087.

Larsen, J. T., & McGraw, A. P. (2014). The case for mixed emotions. *Social and Personality Psychology Compass*, *8*(6), 263–274. https://psycnet.apa.org/doi/10.1111/spc3.12108

Lischke, A., Mau-Moeller, A., Jacksteit, R., Pahnke, R., Hamm, A. O., & Weippert, M. (2018). Heart rate variability is associated with social value orientation in males but not females. *Scientific reports*, *8*(1), 7336. https://doi.org/10.1038/s41598-018-25739-4

Li, X., Chen, J., Zhao, G., & Pietikainen, M. (2014). Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4264-4271.

Liu, X., Fromm, J., Patel, S., & McDuff, D. (2020). Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33, 19400–19411. https://doi.org/10.48550/arXiv.2006.03790

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., ... & Chen, S. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior research methods 53*, 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Mauersberger, H., Tune, J. L., Kastendieck, T., Czarna, A. Z., & Hess, U. (2022). Higher heart rate variability predicts better affective interaction quality in non-intimate social interactions. *Psychophysiology*, *59*(11), e14084. https://doi.org/10.1111/psyp.14084

McColl, R., & Michelotti, M. (2019). Sorry, could you repeat the question? Exploring video-interview recruitment practice in HRM. *Human Resource Management Journal*, *29*, 637–656. https://doi.org/10.1111/1748-8583.12249

Nakatani, M., Ishii, Y., Nakane, A., Takayama, C., & Akasaka, F. (2021). Improving Satisfaction in Group Dialogue: A Comparative Study of Face-to-Face and Online Meetings. In: Masaaki Kurosu (eds.), *Human-Computer Interaction. Design and User Experience Case Studies.* HCII 2021. Lecture Notes in Computer Science, vol 12764. Cham: Springer. https://doi.org/10.1007/978-3-030-78468-3_41

Navarathna, R., Carr, P., Lucey, P., & Matthews, I. (2017). Estimating Audience Engagement to Predict Movie Ratings. *IEEE Transactions on Affective Computing*, *10*(1), 48–59. https://doi.org/10.1109/TAFFC.2017.2723011

Ngien, A., & Hogan, B. (2022). The relationship between Zoom use with the camera on and Zoom fatigue: considering self-monitoring and social interaction anxiety. *Information, Communication & Society*, *26*(10), 2052–2070. https://doi.org/10.1080/1369118X.2022.2065214

Peräkylä, A., Henttonen, P., Voutilainen, L., Kahri, M., Stevanovic, M., Sams, M., & Ravaja, N. (2015). Sharing the emotional load: Recipient affiliation calms down the storyteller. *Social Psychology Quarterly*, *78*(4), 301–323. https://doi.org/10.1177/0190272515611054

Peräkylä, A., Voutilainen, L., Lehtinen, M., & Wuolio, M. (2021). From Engagement to Disengagement in a Psychiatric Assessment Process. *Symbolic Interaction, 42*(2), 257–296. https://doi.org/10.1002/symb.574

Plans, D., Morelli, D., Sütterlin, S., Ollis, L., Derbyshire, G., & Cropley, M. (2019). Use of a biofeedback breathing app to augment poststress physiological recovery: randomized pilot study. *JMIR formative research*, *3*(1), e12227. https://doi.org/10.2196/12227

Poh, M. Z., McDuff, D. J., & Picard, R. W. (2010). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1), 7–11. https://doi.org/10.1109/TBME.2010.2086456

Pohjola, M., Puutio, R. & Kykyri, V.-L. (2023). *Shared reflection and creation of new understanding in online group supervision*. Manuscript in preparation.

Pohjola, M, Puutio, R., Nokia, M., Muotka, J. & Kykyri, V.-L. (In press). Interaction in systemic online group supervision: multimodal analyses of dialogue quality. In: M. Borcsa & V. Pomini (Eds.): *the Handbook of Online Systemic Therapy, Supervision and Training - EFTA Book Series volume 7*. Springer International.

Potter, J., Edwards, D., & Wetherell, M. (1993). A model of discourse in action. *American Behavioral Scientist, 36(3),* 383–401. https://doi.org/10.1177/0002764293036003008

Pralat, R. (2020). https://raphaelpralat.medium.com/how-to-live-stream-on-zoom-with-obs-72b23aa9721a

Pönkänen, L., Peltola, M., & Hietanen, J. (2011). The observer observed: Frontal EEG asymmetry and autonomic responses differentiate between another person's direct and averted gaze when the face is seen live. *International Journal of Psychophysiology*, *82*(2), 180–187. https://doi.org/10.1016/j.ijpsycho.2011.08.006

Pörhölä, M., Isotalus, P., & Ovaskainen, T. (1993). Communication-elicited arousal in public speaking, small group and dyadic contexts: comparisons of change in heart rate. *Communication Research Reports*, *10*(1), 29–37. https://doi.org/10.1080/08824099309359915

Riedl, R. (2022). On the stress potential of videoconferencing: definition and root causes of Zoom fatigue. *Electronic Markets*, *32*(1), 153–177. https://doi.org/10.1007/s12525-021-00501-3

Rettie, R. (2009). Mobile Phone Communication: Extending Goffman to Mediated Interaction. *Sociology*, *43*(3), 421–438. https://doi.org/10.1177/0038038509103197

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161–1178. https://psycnet.apa.org/doi/10.1037/h0077714

Sabour, R. M., Benezeth, Y., De Oliveira, P., Chappé, J., & Yang, F. (2023). UBFC-Phys: A Multimodal Database For Psychophysiological Studies of Social Stress. *IEEE Transactions on Affective Computing*, *14*(1), 622–636. https://doi.org/10.1109/TAFFC.2021.3056960

Sacks, H., Schegloff, E.A., & Jefferson, G. (1974). A simplest systematics for the organization for turn-taking for conversation. *Language*, *50*(4), 696–735. https://doi.org/10.2307/412243

Shaffer F., & Ginsberg, J.P. (2017). An Overview of Heart Rate Variability Metrics and Norms. *Front Public Health*, *5*, 258. https://doi.org/10.3389/fpubh.2017.00258

Sun, Z., & Li, X. (2022). Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. *European Conference on Computer Vision*, 492–510. Cham: Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2208.04378

Sun, Z., Vedernikov, A., Kykyri, V. L., Pohjola, M., Nokia, M., & Li, X. (2022). Estimating Stress in Online Meetings by Remote Physiological Signal and Behavioral Features. *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, 216–220.

Thompson, J. B. (2018). Mediated Interaction in the Digital Age. *Theory, Culture & Society*, *37*(1), 3–28. https://doi.org/10.1177/0263276418808592

van Amelsvoort, L.G.P.M., Schouten, E.G., Maan, A.C., Swenne, C.A., & Kok, F.J. (2000). Occupational determinants of heart rate variability. *Int Arch Occup Environ Health 73*, 255–262. https://doi.org/10.1007/s004200050425

Verkruysse, W., Svaasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Opt. Express, 16*(26), 6–17. https://doi.org/10.1364%2Foe.16.021434

Wang, W., Den Brinker, A. C., Stuijk, S., & De Haan, G. (2016). Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering, 64*(7), 1479–1491. https://doi.org/10.1109/TBME.2016.2609282

Woo, H., Bang, N. M., Lee, J., & Berghuis, K. (2020). A Meta-Analysis of the Counseling Literature on Technology-Assisted Distance Supervision. *International Journal for the Advancement of Counselling*, *42*(4), 424–438. https://psycnet.apa.org/doi/10.1007/s10447-020-09410-0

Yerkes, R. M. and Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *J. Comp. Neurol. Psychol., 18,* 459–482. https://doi.org/10.1002/cne.920180503

Yu, Z., Li, X., & Zhao, G. (2021). Facial-Video-Based Physiological Signal Measurement: Recent Advances and Affective Applications. *IEEE Signal Processing Magazine, 38*(6), 50–58. https://doi.org/10.1109/MSP.2021.3106285

Yu, Z., Li, X., & Zhao, G. (2019). Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *30th British Machine Visison Conference: BMVC 2019*, 1–12. https://doi.org/10.48550/arXiv.1905.02419

Zoom Security Guide (2016). Zoom-Security-White-Paper.pdf (d24cgw3uvb9a9h.cloudfront.net)

APPENDICES

APPENDIX 1: Questionnaire

1. Name of the participant
2. Date of the session
3. Have you recently felt your work situation strained? 1= not at all; 10= extremely straining
4. Please evaluate the working relationship in this session. Select the number that best describes Your experience of the session. 1=I did not feel that I was heard, understood, and respected.; 10= I felt I was heard, understood, and respected.
5. Please evaluate the aims and topics of today's conversation. Select the number that best describes Your experience of the session. 1= We did not discuss and process the topics that I would have wished; 10=I'm satisfied with the topics we discussed and processed in the session.
6. Please evaluate working tools and methods used in today's conversation. Select the number that best describes Your experience of the session. 1= The supervisor's working style and methods were not suitable for me; 10=I felt that the supervisor's working style and methods suited me very well.
7. What is your general evaluation of today's session? Select the number that best describes Your experience of the session. 1=There was something essential missing from today's conversation; 10= Generally, conversation today was good.

APPENDIX 2: RECOMMENDATIONS FOR ONLINE CONVERSATIONS

TO FACILITATE THE TRANSMISSION OF NONVERBAL CUES:

- Keep the camera turned on
- Locate the camera so that the participants are able to see your upper body movements - not just your face
- Adjust the light so that you are facing the source of the light (window or a lamp)
- Encourage using the Reaction buttons available in videoconferencing tools

TO AVOID DISTRACTIONS RELATED TO TECHNOLOGY AND ONLINE CONTEXT

- Prefer the use of microphone-headphone sets over computers built-in microphone and speakers
- Familiarize yourself with videoconferencing tools
- If possible, find a quiet and peaceful location and mark the location booked

TO PROMOTE OPTIMAL LEVEL OF PARTICIPANT ACTIVATION AND TO AVOID ”ZOOM FATIQUE”

- Avoid Zoom or Teams marathons, organize a break if the meeting will last more than one hour.
- Remember to take care of yourself and have breaks between remote meetings
- Consider carefully for whom the meeting is necessary and what is their role in the meeting
- Enhance the participants time to calm down and to direct their focus on the meeting, by e.g., opening small talk or using quiet relaxation in the beginning of the meeting
- Talk about the purpose and aims of the meeting - this may help to promote the transition from other tasks to the meeting. Also let participants know what working methods will be used.
- Be explicit with the meeting agenda and what is expected from the participants
- Use Breakout Rooms with topics that needs to be discussed – smaller groups help to create active participation
- Promote everybody’s participation by showing interest and curiosity to participants own ideas, opinions, and experiences about topic at hand
- When participation in the group is in low level, invite group members to participate and speak less yourself. For example, invite three participants to talk together about the topic while others focus on listening.

- Keep an eye on the signs of disengagement. Change the way how you ”orchestrate” the meeting if participants are getting disengaged. One way to intervene is to invite the participants to comment and reflect on each other’s turns .
- Keep an eye on even small signs of somebody wanting to take a turn