Thesis for The Degree of Doctor of Philosophy

From Stress to Strength: Well-Being and Resilience in Software Engineering

Cristina Martinez Montes

Division of Software Engineering
Department of Computer Science & Engineering
Chalmers University of Technology and Gothenburg University
Gothenburg, Sweden, 2026

From Stress to Strength: Well-Being and Resilience in Software Engineering

Cristina Martinez Montes

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5825
ISSN 978-91-8103-368-7
https://doi.org/10.63959/chalmers.dt/5825

Department of Computer Science & Engineering
Division of Software Engineering
Chalmers University of Technology and Gothenburg University
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

This thesis has been prepared using LaTeX.
Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2026.

“”When we are no longer able to change a situation, we are challenged to change ourselves.”
- Viktor Frankl (Neurologist/Psychiatrist)

From Stress to Strength: Well-Being and Resilience in Software Engineering

Cristina Martinez Montes

Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Software engineers face unique circumstances that shape a specific work context distinct from many other professions. They experience frequent stress due to tight deadlines, heavy cognitive load demands and the constantly changing technology they work with. Hence, it is necessary to pay special attention to engineers' well-being, stress management and resilience. General theories of well-being address several aspects that engineers face. However, due to their specific characteristics, these theories require adaptation to capture the distinct pressures and contextual demands of software engineering work. Moreover, current methodologies require refinement through data triangulation and context-sensitive approaches. Single-source data often falls short in capturing the full experiences, perceptions, and context of engineers.

This thesis aimed to develop a software engineering well-being framework that considers the field's unique circumstances. In addition, it sought to design, test and evaluate interventions targeting engineers' well-being and stress management. Finally, it also investigated a suitable methodological approach that incorporates data triangulation to better capture the complexity of software engineering contexts.

Various empirical methodologies were employed, including interventions, quasi-experiments, experiments, and surveys. The data were analysed using thematic and content analysis for the qualitative data, and descriptive, frequentist, and Bayesian statistics for the quantitative data.

The main outcomes are: First, results provide a context-specific software engineering well-being framework. Second, we present tailored interventions targeting stress and well-being, developed considering engineers' unique circumstances. Third, we propose a data-triangulation approach for data collection and analysis. Finally, they introduce a framework for integrating AI into qualitative data analysis.

The thesis contributions advance the state of the art by offering a framework that explains factors influencing the well-being of software engineers. This framework also offers policy recommendations and interventions to enhance work environments that support well-being. Finally, we advance human factors research with our data triangulation proposal and a hybrid qualitative data analysis framework.

Keywords

Well-being, Resilience, Stress, Software Engineers, Human Factors

List of Publications

This thesis is based on the following publications:

[A] C. Martinez Montes, B. Penzenstadler, R. Feldt "The Factors Influencing Well-Being in Software Engineers: A Mixed-Method Study"
Transactions on Software Engineering and Methodology (TOSEM), 2025. DOI: 10.1145/3770074.

[B] C. Martinez Montes, R. Khojah "Emotional Strain and Frustration in LLM Interactions in Software Engineering"
International Conference on Evaluation and Assessment in Software Engineering (EASE), 2025. DOI: 10.1145/3756681.3756951.

[C] B. Penzenstadler, R. Torkar, C. Martinez "Take a deep breath: Benefits of neuroplasticity practices for software developers and computer workers in a family of experiments"
Empirical Software Engineering Journal (EMSE), 2022. DOI: 10.1007/s10664-022-10148-z .

[D] C. Martinez Montes, B. Penzenstadler "Evaluating the Impact of a Yoga-Based Intervention on Software Engineers' Well-Being"
International Conference on Evaluation and Assessment in Software Engineering (EASE), 2025. DOI: 10.1145/3756681.3756950.

[E] C. Martinez Montes, D. Grassi, N. Novielli, B. Penzenstadler "A Multimodal Approach Combining Biometrics and Self-Report Instruments for Monitoring Stress in Programming: Methodological Insights"
Under submission to special issue on Empirical Software Engineering Journal (EMSE), 2025. https://doi.org/10.48550/arXiv.2507.02118.

[F] C. Martinez Montes, R. Feldt, C. Miguel Martos, S. Ouhbi, S. Premanandan, D. Graziotin "Large Language Models in Thematic Analysis: Prompt Engineering, Evaluation, and Guidelines for Qualitative Software Engineering Research"
Under submission to Transactions on Software Engineering (TSE), 2025.
https://doi.org/10.48550/arXiv.2510.18456.

Other publications

The below publications were written during my PhD as an exploration of tangential areas of my main research. Hence, they were not including in my PhD story line.

[a] C. Martinez Montes, J. Johansson, E. Dunvald "Factors Influencing Gender Representation in IT Faculty Programmes: Insights with a Focus on Software Engineering in a Nordic Context".
In Proceedings of the International Conference on the Foundations of Software Engineering (FSE Companion, Education Track), 2025, pp. 772–782. DOI: 10.1145/3696630.3727234.

[b] B. Penzenstadler, S. Motogna, P. Lago, C. Martinez Montes "Policy Making as Extension of Disseminating Research Results: Policy Influence Plan Canvas".
In B. Penzenstadler, K. Boudaoud, A. Di Marco & S. Caner-Yıldırım (Eds.), Actions for Gender Balance in Informatics Across Europe, 2025 (pp. 383–397). Springer, Cham. DOI: 10.1007/978-3-031-78432-3 16.

[c] S. Chand, C. Li, C. Martinez Montes, B. Cabrero-Daniel, J. Horkoff "Automating Requirements Review in the Automotive Sector: A Tailored AI Approach"
International Requirements Engineering Conference (RE), 2024 (Poster Track), (pp. 492-493). IEEE. DOI: 10.1109/RE59067.2024.00059.

[d] C. Martinez Montes, F. Sjögren, A. Klevfors, B. Penzenstadler "Qualifying and Quantifying the Benefits of Mindfulness Practices for IT Workers"
In 2024 10th International Conference on ICT for Sustainability (ICT4S) (pp. 272-281). IEEE. DOI: 10.1109/ICT4S64576.2024.00035.

[e] C. Martinez Montes, B. Penzenstadler "Piloting a well-being and resilience intervention in a course on digitalization for sustainability."
In The International Conference on Information and Communications Technology for Sustainability (ICT4S), Doctoral Symposium, Demos, Posters (Intl. Workshop on ICT4S Education), (pp. 105-118), 2023.

Research Contribution

In Paper A, my contributions were the conceptualisation, data curation and analysis, investigation, development of the research methodology, project administration, validation, visualisation, drafting, reviewing and editing the original manuscript.

In Paper B, I contributed to the conceptualisation, data curation and analysis, investigation, development of the research methodology, project administration, supervision, validation, visualisation, drafting, reviewing and editing the original manuscript.

For Paper C, I contributed to the formal analysis, data curation, validation and visualisation. I also participated in writing the original draft.

For Paper D, I was responsible for conceptualisation, data curation and analysis, investigation, development of the research methodology, project administration, validation, visualisation, drafting, reviewing and editing the original manuscript.

In Paper E, I was involved in the conceptualisation, data curation and analysis, investigation, development of the research methodology, project administration, validation, visualisation, drafting, reviewing and editing the original manuscript.

For Paper F, I participated in the conceptualisation, data curation and analysis, investigation, development of the research methodology, project administration, validation, visualisation, drafting, reviewing and editing the original manuscript.

In all the articles, my co-authors contributed in several roles, for example, Conceptualisation, Methodology, Validation, Supervision and Investigation. Moreover, particularly in Formal Analysis, papers involving qualitative data required my co-authors to be involved in the data analysis, thereby enhancing the reflexivity of the process. They also provided input reviewing, editing and refining the manuscript drafts.

Table presents my contributions to the appended papers following the CRediT (Contribution Roles Taxonomy) [1] criteria.

Role / Paper	A	B	C	D	E	F
Conceptualisation	✓	✓		✓	✓	✓
Methodology	✓	✓		✓	✓	✓
Software						
Validation	✓	✓	✓	✓	✓	✓
Formal analysis	✓	✓	✓	✓	✓	✓
Investigation	✓	✓		✓	✓	✓
Resources	✓	✓		✓	✓	✓
Data Curation	✓	✓	✓	✓	✓	✓
Writing – Original Draft	✓	✓	✓	✓	✓	✓
Writing – Review & Editing	✓	✓		✓	✓	✓
Visualization	✓	✓	✓	✓	✓	✓
Supervision		✓			✓	
Project administration	✓	✓			✓	
Funding acquisition						

Acknowledgment

A mi **F**amilia, and to all the minds that have accompanied, supervised, inspired, supported, listened, challenged, and motivated me, thank you all, especially those **W**ho were light when it was dark.

Contents

Abstract iii

List of Publications v

Other Publications vi

Research Contribution vii

Acknowledgement ix

1 Introduction 1
- 1.1 Research Focus . . . 3
- 1.2 Background . . . 6
  - 1.2.1 Stress and the Nervous System . . . 6
  - 1.2.2 Mindfulness-based practices . . . 8
    - 1.2.2.1 Breathwork . . . 8
    - 1.2.2.2 Yoga . . . 8
    - 1.2.2.3 Journaling . . . 9
  - 1.2.3 Emotions in SE . . . 9
  - 1.2.4 Qualitative Data Analysis in SE . . . 10
- 1.3 Related Work . . . 10
  - 1.3.1 Understanding Stress and Resilience in Software Engineers . . . 10
  - 1.3.2 Support and Enhancement of Resilience and Well-Being in the Software Engineering . . . 11
  - 1.3.3 Research on the Human Factors in Software Engineering . . . 12
- 1.4 Research Methodology . . . 13
  - 1.4.1 Research Designs . . . 13
    - 1.4.1.1 Quasi-experiments (interventions) . . . 13
    - 1.4.1.2 Survey . . . 15
    - 1.4.1.3 Experiment . . . 15
    - 1.4.1.4 Exploratory Sequential Design . . . 15
  - 1.4.2 Data Collection Methods . . . 16
    - 1.4.2.1 Questionnaires . . . 16
    - 1.4.2.2 Interviews and Focus Groups . . . 16
    - 1.4.2.3 Biometric Data . . . 16
  - 1.4.3 Data Analysis . . . 17

1.4.3.1 Quantitative Analysis Method . . . 17
1.4.3.2 Qualitative Analysis Method . . . 17
1.4.4 Reflexivity . . . 18
1.5 Ethical Considerations . . . 19
1.6 Research Results . . . 20
1.7 Discussion and Answers to the RQs . . . 25
1.7.1 Factors and conditions that influence stress, well-being, and resilience . . . 26
1.7.2 Approaches to foster sustained well-being . . . 30
1.7.3 Multimodal data triangulation and LLM-assisted analysis . . . 31
1.8 Limitations and Threats to Validity . . . 32
1.8.1 Scope of Applicability . . . 32
1.8.2 Internal Validity . . . 33
1.8.3 External Validity . . . 33
1.8.4 Construct Validity . . . 34
1.9 Conclusions . . . 34

2 Paper A: The Factors Influencing Well-Being in Software Engineers: A Mixed-Method Study 39
2.1 Introduction . . . 40
2.2 Background and Related Work . . . 41
2.2.1 Background . . . 41
2.2.1.1 The Conception of Well-being . . . 41
2.2.2 Related Work . . . 42
2.2.2.1 General Population . . . 42
2.2.2.2 Software Engineering Population . . . 43
2.3 Methodology . . . 44
2.3.1 Study design . . . 44
2.3.2 Population . . . 45
2.3.3 Data Collection . . . 45
2.3.3.1 Pilots of the data collection instruments . . . 45
2.3.3.2 Interviews . . . 45
2.3.3.3 Survey . . . 46
2.3.4 Data Analysis . . . 47
2.3.4.1 Interview Analysis . . . 47
2.3.4.2 Survey Analysis . . . 47
2.3.4.3 Reflexivity . . . 47
2.3.5 Ethical Considerations . . . 48
2.4 Results . . . 48
2.4.1 Interviews . . . 48
2.4.1.1 Theme 1: Individual Conception of Well-being . . . 48
2.4.1.2 Theme 2: Personal and Collaborative Factors . . . 50
2.4.1.3 Theme 3: Support and Recognition . . . 52
2.4.1.4 Theme 4: Work Environment and Culture . . . 55
2.4.1.5 Theme 5: Challenges and Stressors . . . 58
2.4.2 Survey Results . . . 60
2.4.2.1 Survey Respondent Demographics . . . 61

2.4.2.2 Associations Between Well-Being and Three-Level Variables . . . 61
2.4.2.3 Likert Scale Questions . . . 62
2.4.2.4 Open Questions . . . 63
2.5 Discussion . . . 69
2.5.1 Alignments in Quantitative and Qualitative Results . . . 70
2.5.1.1 Personal Practices . . . 70
2.5.1.2 Support from the Company and Peers . . . 71
2.5.1.3 Work Environment: Trust, Physical Well-being, and Compensation . . . 71
2.5.1.4 Equality, Equity, Diversity, and Inclusion (EEDI) . . . 71
2.5.1.5 Personal Life Situations . . . 72
2.5.1.6 Workload and Time Constraints . . . 72
2.5.1.7 Social Integration and Loneliness . . . 73
2.5.1.8 Tech Tools and Their Impact on Communication and Productivity . . . 73
2.5.2 Contrasting Quantitative and Qualitative Results . . . 73
2.5.2.1 Influence of Social Interactions on Well-being . . . 73
2.5.2.2 Recognition at Work . . . 74
2.5.2.3 Professional and Personal Growth Support from Companies . . . 74
2.5.2.4 Company Policies and Practices . . . 74
2.5.3 Comparison to Other Theories of Well-being Factors . . . 76
2.5.4 Policy Recommendations . . . 78
2.5.5 Validity Threats . . . 79
2.5.5.1 Internal validity . . . 79
2.5.5.2 External validity . . . 79
2.5.5.3 Construct validity . . . 80
2.5.6 Future Work . . . 80
2.6 Conclusion . . . 80

3 Paper B: Emotional Strain and Frustration in LLM Interactions in Software Engineering 83
3.1 Introduction . . . 84
3.2 Background and Related Work . . . 85
3.2.1 Large Language Models in Software Engineering . . . 85
3.2.2 Emotions Involved When Using Technology . . . 85
3.2.3 Emotions in Software Engineering Tasks . . . 87
3.3 Methodology . . . 88
3.3.1 Target Population . . . 89
3.3.2 Data Collection . . . 89
3.3.3 Data Analysis . . . 90
3.4 Results . . . 90
3.4.1 Respondents Demographics . . . 90
3.4.2 Usage of LLMs in Software Engineering Industry and Academia 90
3.4.3 Emotions During LLM Interaction . . . 91
3.4.4 Expectations When Interacting with LLMs . . . 93

3.4.5 Frustration Triggers in Software Engineering 95
3.4.6 Unmet Expectations' Impact on Motivation 97
3.4.7 User Actions for Improving LLM Interactions 98
3.5 Discussion 101
3.5.1 Frustrations in the Context of LLM Interaction 101
3.5.2 Impact of Frustrating Experiences on Motivation 102
3.5.3 Towards a Less-Frustrating User Experience 102
3.5.4 Threats to Validity 103
3.6 Conclusion 104

4 Paper C: Take a deep breath: Benefits of neuroplasticity practices for software developers and computer workers in a family of experiments 105
4.1 Introduction 106
4.2 Background 107
4.2.1 Context: Stress in Software Development and IT Work 107
4.2.2 Background: Breathing practices 108
4.2.3 Theory: Mindfulness and Mindfulness Attention Awareness . . 109
4.2.4 Related work: Well-being and Resilience for Engineers, Software Developers and IT Workers 110
4.3 Research Design 112
4.3.1 The intervention: Rise 2 Flow program 112
4.3.2 Research Questions 114
4.3.3 Population and Inclusion Criteria 115
4.3.4 Instrument Design 116
4.3.4.1 Entry/Exit Survey 116
4.3.4.2 Daily journal (for RQ1b) 119
4.3.4.3 Weekly survey: WHO-5 (for RQ2d) 119
4.4 Analysis Procedure 120
4.4.1 Demographics 120
4.4.2 Statistical Analysis of Instruments 121
4.4.2.1 The Data and Data Cleaning 122
4.4.2.2 Temporal Analysis of Entry vs. Exit 123
4.4.2.3 Temporal Analysis of Weekly and Daily Trends 124
4.4.3 Qualitative Analysis of Instruments: Thematic analysis 126
4.5 Results 126
4.5.1 Overall Engagement 127
4.5.2 RQ1: Changes in Mindfulness Attention Awareness and Daily Perceptions 130
4.5.2.1 Does the intervention bring about change in the participants Mindfulness Attention Awareness? (RQ1a) 131
4.5.2.2 How did the perceptions of life experience progress over time? (RQ1b) 135
4.5.3 RQ2: Changes in Well-being 137
4.5.3.1 Does the intervention lead to change in the participants' perceptions of positive and negative experiences? If so, how are their experiences affected? (RQ2a) 139

4.5.3.2 Does the intervention lead to change in the participants' psychological well-being? If so, how is it affected? (RQ2b) . . . 140
4.5.3.3 Does the intervention lead to change with regard to their positive thinking? If so, how is it affected? (RQ2c) 140
4.5.3.4 How does the well-being fluctuate and vary over the course of the intervention? (RQ2d) . . . 141
4.5.4 RQ3: Changes in Perceived Productivity and Self Efficacy . . . 142
4.5.4.1 Does the intervention lead to change in the participants perceived productivity? If so, how is it affected? (RQ3a) . . . 142
4.5.4.2 Does the intervention lead to change in the participants' self-efficacy? If so, how is it affected? (RQ3b) . 143
4.6 Discussion . . . 144
4.6.1 Significance . . . 144
4.6.2 Observations and Implications . . . 146
4.6.3 Limitations and Threats to Validity . . . 147
4.6.4 Relating Back to Theory in Psychology . . . 149
4.6.5 Future Work . . . 149
4.7 Conclusion . . . 150
4.8 Data Availability . . . 152
4.9 acknowledgements . . . 152

5 Paper D: Evaluating the Impact of a Yoga-Based Intervention on Software Engineers' Well-Being 153
5.1 Introduction . . . 154
5.2 Related Work . . . 155
5.2.1 The Effectiveness of Yoga in General . . . 155
5.2.2 The Effectiveness of Hatha Yoga . . . 156
5.2.3 Consequences for job performance and outcomes proposed . . . 156
5.2.4 Moderating Factors and Conditions for Mindfulness . . . 157
5.2.5 Well-being Interventions in SE . . . 157
5.3 Methodology . . . 157
5.3.1 Research Design . . . 157
5.3.2 Intervention . . . 158
5.3.3 Company, Population and Inclusion Criteria . . . 158
5.3.4 Data Collection . . . 159
5.3.4.1 Entry and Exit Survey . . . 159
5.3.4.2 Weekly Well-being Scale . . . 160
5.3.4.3 Focus Group . . . 160
5.3.5 Data Analysis . . . 160
5.3.5.1 Statistical Analysis of Instruments . . . 161
5.3.5.2 Qualitative Analysis of Focus Group . . . 161
5.3.5.3 Reflexivity . . . 162
5.3.6 Ethical Considerations . . . 162
5.4 Results . . . 162
5.4.1 Quantitative Analysis . . . 163

5.4.2 Thematic Analysis . . . 164
5.4.2.1 Theme 1: Individual Benefits and Shared Reflections in Practice . . . 165
5.4.2.2 Theme 2: Organisational Support and Logistical Challenges in Implementing the Programme . . . 167
5.4.2.3 Theme 3: Perception and General Feedback . . . 168
5.5 Discussion . . . 168
5.5.1 Results in Context . . . 169
5.5.2 Importance of the Study . . . 170
5.5.3 Lessons Learned . . . 171
5.5.4 Validity Threats . . . 172
5.5.4.1 Internal Validity . . . 172
5.5.4.2 External Validity . . . 172
5.5.4.3 Construct Validity . . . 172
5.5.4.4 Conclusion Validity . . . 172
5.6 Conclusion . . . 173
5.7 Acknowledgement . . . 173

6 Paper E: A Multimodal Approach Combining Biometrics and Self-Report Instruments for Monitoring Stress in Programming: Methodological Insights 175
6.1 Introduction . . . 176
6.2 Background and Related Work . . . 177
6.2.1 Physiological Measures of Stress and Mental Load . . . 178
6.2.2 Using Biometrics for Studying Cognitive and Affective States in Software Development . . . 180
6.3 Methodology . . . 181
6.3.1 Participants and Recruitment Strategy . . . 182
6.3.2 Experiment Setup . . . 182
6.3.3 Data Collection Methods . . . 183
6.3.3.1 Biometric Sensors . . . 183
6.3.3.2 Self-report Instruments . . . 184
6.3.3.3 Post-Task Interview . . . 184
6.3.4 Data Analysis . . . 184
6.3.4.1 Psychometrics . . . 185
6.3.4.2 Biometrics . . . 185
6.3.4.3 Interviews . . . 187
6.4 Results . . . 187
6.4.1 Psychometrics . . . 188
6.4.1.1 Closer Look per Participant . . . 189
6.4.2 Biometrics . . . 190
6.4.3 Thematic Analysis . . . 192
6.4.3.1 Theme 1: Task Impressions: Engagement and Learning 192
6.4.3.2 Theme 2: Emotional Responses to Challenge and Uncertainty . . . 193
6.4.3.3 Theme 3: Sources of Distraction and Discomfort . . . 194
6.4.3.4 Theme 4: Coping Strategies and Adaptation . . . 194

6.5 Discussion . . . 195
6.5.1 Main Contributions . . . 195
6.5.1.1 Evidence Supporting the Alignment of Biometrics and Psychometric Instruments . . . 195
6.5.1.2 Methodological Insights on Recruitment and Experimental Design . . . 200
6.5.1.3 General Lessons for Conducting Stress Studies . . . 201
6.5.1.4 Ethical Reflections on Stress Induction in Research . 202
6.5.2 Threats to Validity . . . 203
6.5.2.1 Internal Validity . . . 203
6.5.2.2 External Validity . . . 203
6.5.2.3 Construct Validity . . . 203
6.5.2.4 Ecological Validity . . . 204
6.6 Conclusion . . . 204
6.6.1 Summary . . . 204
6.6.2 Future Work . . . 204
6.7 Declarations . . . 205
6.7.1 Funding: . . . 205
6.7.2 Ethical approval: . . . 205
6.7.3 Informed consent: . . . 205
6.7.4 Author Contributions [all authors should be mentioned] . . . 205
6.7.5 Data Availability Statement . . . 205
6.7.6 Conflict of Interest . . . 206
6.7.7 Clinical Trial Number in the manuscript. . . . 206

7 Paper F: Large Language Models in Thematic Analysis: Prompt Engineering, Evaluation, and Guidelines for Qualitative Software Engineering Research 207
7.1 Introduction . . . 208
7.2 Background and Related Work . . . 209
7.2.1 Qualitative Data Analysis in Software Engineering . . . 209
7.2.1.1 Thematic analysis . . . 209
7.2.2 Early AI and ML in Qualitative Data Analysis . . . 210
7.2.3 LLMs in Qualitative Analysis . . . 211
7.2.4 Generative AI Tools and Frameworks in QDA . . . 211
7.3 Methodology . . . 213
7.3.1 Dataset . . . 213
7.3.2 Study Design: Mapping TA Phases to Human vs LLM Roles . 213
7.3.3 Prompting Strategy, Application and Evaluation . . . 216
7.3.3.1 Prompting Strategy . . . 218
7.3.3.2 Initial Coding (Phase 2) . . . 219
7.3.3.3 Generating, Reviewing, Refining and Naming Themes (Phases 3-5) . . . 220
7.3.4 Data analysis . . . 221
7.3.5 Evaluators Team . . . 221
7.3.6 Ethical Considerations . . . 222
7.4 Results . . . 222

7.4.1 Evaluation Phase 2: Human vs LLM (1a) . . . 222
7.4.2 Evaluation Phase 2: LLM Codes Rubric-Based Evaluation (1b) 226
7.4.3 Evaluation Phase 3-5 Generating, Refining and Naming Themes (2a) . . . 228
7.5 Discussion . . . 229
7.5.1 LLMs as Analytical Assistants in SE Research . . . 229
7.5.1.1 Human Oversight Remains Essential . . . 231
7.5.2 Strengths and Limitations of LLM Outputs in Engineering Contexts . . . 233
7.5.3 Strategies to Ensure Methodological Rigour . . . 234
7.5.4 Implications for Empirical SE Methods . . . 234
7.5.5 Threats to Validity . . . 235
7.6 Conclusion . . . 236
7.7 Authors' Contributions . . . 237

Bibliography **239**

Appendix **277**

A Appendix - Paper A **277**
A.1 Data Collection Instruments . . . 277
A.2 Survey Instruments . . . 277
A.2.1 The MAAS instrument . . . 277
A.2.2 The instruments SPANE, PWB, and PTS . . . 277
A.2.3 Self Efficacy . . . 279
A.2.4 Perceived Productivity . . . 280
A.2.5 The WHO-5 instrument . . . 282
A.3 Model designs . . . 283
A.3.1 Gaussian Process model . . . 283
A.3.2 Dummy variable regression model . . . 284
A.4 Detailed Findings: Significant Effects of Other Predictors . . . 285
A.4.1 Mindfulness Attention Awareness Scale . . . 285
A.4.2 Scale of Positive And Negative Experiences . . . 286
A.4.3 Psychological Well-Being . . . 287
A.4.4 Positive Thinking Scale . . . 288
A.4.5 Self Efficacy . . . 289
A.4.6 Perceived Productivity . . . 290
A.4.7 Predictor Number of Sessions . . . 290

Chapter 1

Introduction

"*We cannot change the human condition, but we can change the conditions under which people work*" - James Reason.

Over the past two decades, human factors in software engineering (SE) have evolved from a peripheral concern to a central research theme [2–4]. More recently, the focus has been on psychological factors (a core part of the human condition) [5], including well-being.

Research from occupational and cognitive psychology has long established that well-being, stress regulation and resilience influence performance, motivation, and long-term health [6–10]. Prolonged or poorly regulated stress can erode well-being, while resilience processes help maintain functioning under pressure and support motivation and long-term health. In software engineering (SE), these principles require adaptation to the specific characteristics of engineers' work. SE poses different challenges compared to other fields [5] as it encompasses cognitive, emotional, and social aspects. For example, engineers' activities require a combination of creativity and autonomy [11], high cognitive load [12], problem solving [13], group dynamics [14], long periods of focused attention [15], and collaborative workflows [16]. Moreover, engineers engage in sustained problem-solving under conditions of uncertainty, time pressure, and rapid technological change [17, 18]. **These characteristics make software engineers particularly vulnerable to stress, emotional exhaustion, and burnout** [19]. Additionally, this specific combination of factors makes work conditions in SE distinct from those in general occupational contexts, creating a need to adapt psychological models and measures to this high-risk population.

Furthermore, the recent and rapid adoption of artificial intelligence (AI) tools, particularly large language models (LLMs), presents new concerns for well-being. These tools change several engineers' tasks. For example, how they search for information, design solutions, and code and debug [20]. Its use potentially reduces some forms of effort. However, it also introduces new challenges, such as overreliance on generated outputs, the need to be vigilant in detecting subtle errors, and the requirement to learn new workflows. Understanding how these technologies influence cognitive load, affect, and well-being is important for the future of sustainable software development.

At the same time, practical change requires going beyond identifying and measuring these factors. Improving well-being in practice requires developing and implementing interventions that address them. Promoting well-being is not merely an ethical imperative but a condition for sustainable development, retention, and innovation [21]. Nevertheless, there is limited empirical evidence on which approaches can be effectively developed, evaluated, and sustained to foster well-being in software engineering. Such approaches can only be effective if they are grounded in a precise understanding of the problem they seek to address and are evaluated using appropriate empirical methods.

Despite this growing recognition and research, **empirical evidence on the factors influencing stress, well-being, and resilience in software engineers remained fragmented**. This fragmentation limits theoretical integration and constrains the design of effective interventions. Existing studies have predominantly focused on single constructs such as happiness [8], sentiments and emotions [22], motivation [23] burnout [24] or productivity [25]. However, an integrated view of these factors is essential, since the interaction between them influences how software engineers experience and manage their work-related challenges.

Additionally, **current study methods often rely on a single data collection point**, typically surveys [3], and rarely integrate multiple data sources into a single analysis. This limits the comprehensiveness of the problem. Quantitative surveys capture correlations but lack context, while qualitative studies offer depth but are challenging to replicate and scale [26]. Some studies have started to apply data triangulation by incorporating biometric data. However, methodological standards for such multimodal research are still in development.

Similarly, **interventions to enhance well-being (e.g., mindfulness-based or resilience-training programs) are rare and typically limited to short-term pilots** with small sample sizes [27]. More importantly, improving well-being in software engineering requires approaches that go beyond isolated individual practices. Moreover, prior research suggests that supporting well-being in knowledge-intensive work requires broader approaches that consider organisational conditions [9]. However, there remains a limited empirical understanding of how such multi-level approaches (spanning individual and organisational factors) can be designed, combined, and evaluated within SE contexts.

To effectively address these limitations, a qualitative approach is necessary. Consequently, this raises another question: how to rigorously and consistently analyse rich qualitative data as studies grow in scale and complexity? Recently, researchers have been exploring whether LLMs can support qualitative analysis (or parts of it). For example, to generate deductive codes [28] or create themes [29]. However, as this line of inquiry is still emerging, its methodological foundations remain unsettled. The implementation of LLMs to assist qualitative analysis presents opportunities and epistemological risks. Partial automation can augment analytical rigour, but it also threatens interpretive validity if used uncritically. Hence, clear methodological strategies are needed to integrate LLM-assisted analysis without compromising rigour and transparency.

This thesis addresses these gaps by studying the well-being of software engineers through an integrated, empirical, and reflexive research agenda. It combines psychological theory, human factors research, and software engineering methodologies.

1.1 Research Focus

The thesis follows a progression argument, moving from explanation to action and, ultimately, to methodological contributions. Hence:

In-depth problem analysis → Approaches (actions) targeting the problem → Methodological proposals to study the problem

The first step was to explore and **study the problem (RQ1)**, so we developed an empirical framework to explain the factors that influence well-being in software engineers. Then, with a more precise understanding of the problem, we could suggest actions. In this thesis, those **actions (RQ2)** were translated into interventions, policy recommendations, organisational guidelines, and design proposals for chatbots and LLMs. It is important to note that the interventions served beyond just testing tailored programmes for stress management. They also informed the need to strengthen and improve measurement, completeness, and rigour in qualitative research, specifically in human factors. Finally, we addressed and proposed **strategies for studying human factors in SE (RQ3)**. We suggested and tested the inclusion of biometric data in mixed-methods studies and proposed a hybrid framework to integrate LLMs as research assistants.

The work is organised around three overarching Research Questions(RQs):

RQ1.What are the factors and conditions that influence stress, well-being, and resilience in engineers?

This question aims to identify the main contributors and hindrances to stress, well-being, and resilience. It attempts to directly address the current trend of treating each factor in isolation. It is driven to detect and integrate these multiple factors, explaining their interaction at different levels (individual, team, and organisational) and their influence on the well-being of software engineers. As stated in the introduction, much of the current research studies general psychosocial factors without distinguishing between fields or professions. Given the specific cognitive, social, and organisational characteristics of software engineering activities, this thesis aims to explore and compare these factors in a context-sensitive and empirically grounded manner. One of the goals is to develop a framework grounded in the characteristics, practices, and environments of the software engineering population. The resulting framework is intended to provide a structured view of existing and newly identified factors. Our vision is for the framework to guide future empirical research and to inform the design of interventions, educational practices, and organisational policies that are better aligned with the realities of software engineering work and learning contexts.

RQ2.What approaches can be developed and evaluated to foster sustained well-being among software engineers?

The goal of this question is to propose informed mindfulness interventions based on the data and results from the previous question. We aimed to develop approaches grounded in the specific stressors, work practices, and contextual constraints identified within SE settings. Mindfulness-based interventions are selected as the primary focus because they directly target attentional regulation, cognitive reactivity, and emotional awareness. These aspects are particularly relevant to SE work, which is characterised by prolonged cognitive effort, high mental load, frequent interruptions, and persistent problem-solving demands. Moreover, mindfulness interventions can be applied at

the individual and team level and are adaptable to diverse work contexts, including remote and time-constrained environments. Additionally, these types of interventions are a low-cost and scalable alternative that can be integrated into existing practices with minimal disruption.

Beyond the design, this question also addresses how these interventions can be more effectively measured and evaluated, as well as the challenges that may hinder their implementation and success. With RQ2, we want to bridge the gap between theory and well-being interventions applicable in real-world SE environments.

RQ3. How can multimodal data triangulation and LLM-assisted analysis be used to develop rigorous methodological strategies for studying human factors in software engineering?

This question examines how triangulating multiple types of data (psychometric instruments, interviews and physiological measures) can strengthen the study of human factors in SE. It focuses on improving validity, depth, and interpretive robustness. Each data modality captures different aspects of stress, well-being, and cognitive experience, and their combined use allows the identification of convergent, complementary, or conflicting evidence that would be hard to observe through single-method approaches.

It also investigates how LLMs can be integrated as analytical assistants to support qualitative data analysis while maintaining methodological rigour, transparency, and ethical safeguards. We want to assess if LLMs can augment human analysis while preserving the subjective and reflexive nature of qualitative data analysis. This question seeks to examine strategies for maintaining transparency, traceability of analytical decisions, and consistency with established qualitative methods. One of the goals is to critically evaluate the benefits and limitations of multimodal triangulation and LLM integration. This RQ aims to develop practical methodological strategies that are empirically robust and suitable for human factors research in SE.

To answer these questions, six empirical studies (Papers A–F) were conducted:

- Paper A (Well-Being Factors): Mixed-method exploration of the determinants of software engineers' well-being.
- Paper B (Emotional Strain by AI): Investigation of emotional strain in human–LLM interaction.
- Papers C (Breathwork Intervention) and D (Yoga Intervention): Quasi-experimental mindfulness interventions using breathwork, yoga, and journaling.
- Paper E (Multimodal Methodology): Multimodal stress study combining biometric, self-report measures and interviews.
- Paper F (AI 4 Thematic Analysis): Advancing human–AI collaboration in qualitative data analysis.

These studies propose a multi-level (considering individual, team and organisation) and multi-modal (integrating different data sources) perspective on well-being in SE. Figure 1.1 shows how the papers group to answer the previous RQs. Each paper is presented with an icon with its main contribution written below.

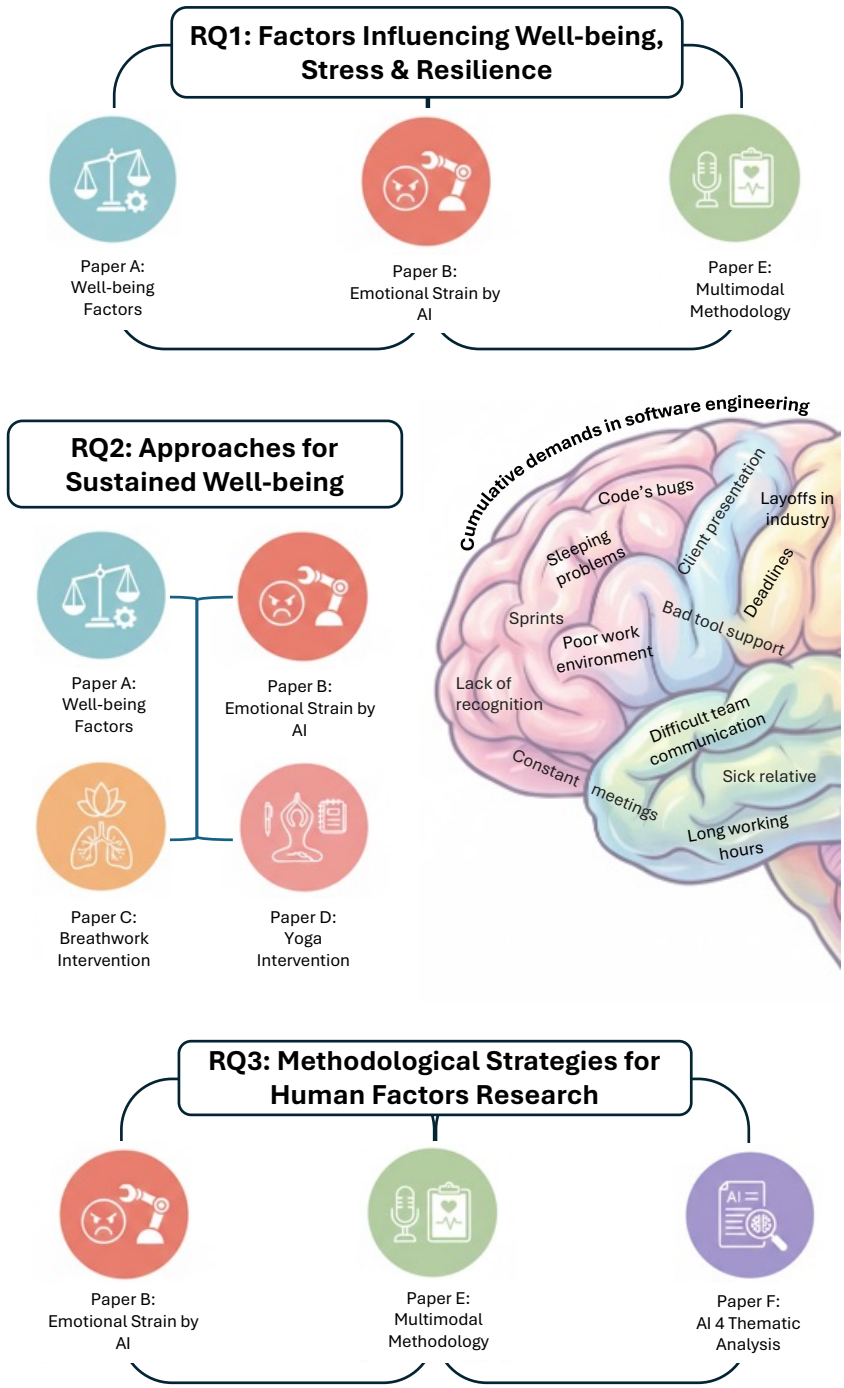


Figure 1.1: Thesis overview. It shows how the papers answer each RQ. The brain in the centre simulates what triggers stress in software engineers.

1.2 Background

This section provides definitions of the main concepts used in the thesis, an introduction to the biology involved in stress, and a brief overview of mindfulness practices.

Table 1.1 presents the definitions based on the American Psychological Association (APA) [30].

Concept	Definition
Awareness	- Being consciously able to notice, recognise, or understand something, and being able to describe it accurately. It refers to the state of being conscious of what is happening around or within oneself.
Mindfulness	- A state of enhanced awareness of the present moment, including one's sensations, thoughts, bodily states, consciousness, and environment, while fostering an attitude of acceptance without judgment or reaction.
Neuroplasticity	- The ability of the nervous system to change its structure or function in response to experience or environmental stimulation.
Resilience	- The process and outcome of successfully adapting to difficult or challenging life experiences, primarily through mental, emotional, and behavioural flexibility and adjustment to external and internal demands.
Stress	- The body's response to events that demand an individual to adjust or employ coping strategies. These events can arise from external situations or internal factors. It influences almost every system in the body, shaping how people behave and feel.
Well-being	- A state of happiness and contentment, with low levels of distress, overall good physical and mental health and outlook, or good quality of life

Table 1.1: Summary of the most important terms used in the thesis.

1.2.1 Stress and the Nervous System

This subsection provides a brief explanation of the biological aspects of stress in the human body [31].

Stress is a biological process that prepares the body to respond to demands. When the brain perceives a situation as challenging or threatening, either physical (e.g., danger, injury) or psychological (e.g., work pressure, fear), the amygdala detects it and sends a distress signal to the hypothalamus. The hypothalamus acts as a control centre that activates the sympathetic nervous system (SNS). Then, the SNS triggers the adrenal medulla (a part of the adrenal glands located on top of the kidneys) to release adrenaline (epinephrine) into the bloodstream. When this happens, several physiological changes occur in the body, including increased heart rate, elevated blood pressure, faster breathing, redirection of blood flow to the muscles, and heightened alertness. If the challenge continues, the hypothalamic–pituitary–adrenal (HPA) axis releases cortisol, which helps the body sustain attention, mobilise energy, and remain prepared.

Under short-term conditions, the stress response is adaptive. The parasympathetic

nervous system (PNS), largely via the vagus nerve, restores the body once the demand has passed by reducing heart rate, supporting digestion, and lowering stress-related neuroendocrine activity [32].

However, modern environments create a different kind of demand. In knowledge-intensive work, such as in SE, stress comes from continuous cognitive load, task switching, ambiguity in roles, interruptions, overload of work and prolonged mental effort [33, 34]. These situations require the brain to regulate attention, energy, and physiological resources repeatedly. This ongoing regulatory activity is described as allostasis: the process of adjusting bodily systems to meet current and anticipated demands [35–37].

When these adjustments are triggered too often, or recovery is insufficient, the body accumulates allostatic load, a measurable strain on physiological systems [36]. Elevated cortisol levels over time can weaken immune function, impair memory and decision-making, and increase inflammation [38, 39]. Repeated sympathetic activation can contribute to hypertension, sleep difficulties, and anxiety, while excessive epinephrine exposure can strain the cardiovascular system [38, 39]. Brain areas involved in focus, planning, and emotional regulation (such as the prefrontal cortex and hippocampus) also become less efficient under chronic load [40].

Since recovery depends heavily on parasympathetic activity, the vagus nerve is essential in helping the body return to baseline. Higher vagal tone is associated with quicker physiological recovery, improved emotional regulation, and better stress resilience [32].

In this thesis, the implemented interventions targeted the vagus nerve.

Technostress

Technostress is defined as a particular type of stress caused by the use of information and communication technologies (ICT) [41]. Software engineers are particularly prone to face technostress due to their close daily interaction with technology [42]. This interaction is defined by continuous development, use, and modification of complex, rapidly evolving technological systems [19, 43]. While technostress affects many technology-intensive professions, software engineering is characterised by persistent cognitive demands, frequent interruptions, tight deadlines, and ongoing pressure to adapt to new tools, frameworks, and programming paradigms [44, 45]. These characteristics make software engineers especially susceptible to technostress and its specific dimensions, including techno-overload, techno-complexity, and techno-uncertainty. These dimensions have been associated with increased psychological tension and emotional exhaustion [41]. Hence, besides the average stressors of daily life, there are more and closer risk factors for the engineering population.

Moreover, the increased use of AI in SE tasks adds to the current equation. Only a few works have specifically examined the negative impact of AI tools. However, existing studies report that constant use can cause emotional dysregulation and social withdrawal [46, 47]. Since AI use also produces positive effects, such as increased productivity, it becomes challenging for users to notice when the advantages shift into pressure or emotional strain. This makes the risks of technostress harder to detect in daily work.

1.2.2 Mindfulness-based practices

Mindfulness-based practices are activities or interventions where the individual intentionally observes their own body and mind in the present moment [48, 49]. It implies an attitude of openness, curiosity, acceptance, and non-judgment.

Practising mindfulness has been shown to have a positive effect on emotionally stressful situations and to increase immune system response [50–52]. For this thesis, we selected specific practices that could be adapted to the formats and organisations in which they were implemented. I elaborate on each practice, breathwork, yoga and journaling, in the following sub-sections.

1.2.2.1 Breathwork

Breathwork is the practice of regulating the way one breathes [53]. It encompasses specific breathing rhythms and patterns focused on promoting mental, emotional, and physical well-being [54]. By adjusting their breathing patterns, individuals can quickly influence the interaction between the respiratory system and the brain regions that regulate behaviour, thinking, and emotions [55]. Moreover, controlled breathing rhythms can promote harmony in brainwave activity. Intentionally slowing the breath aligns with brain electrical patterns, enhancing communication across different brain regions [53].

Research in psychiatry has found that breathwork can improve obsessions, depression, anxiety, trauma, inattention and compulsions [56]. Additionally, specific breathing exercises have proven to help reduce stress levels [57].

Breathwork has historical and traditional roots, including Tibetan Buddhism, yoga, and ancient practices such as Pranayama. Over time, numerous contemporary methods and exercises have been developed, yet the traditional and modern approaches remain widely practised today. As it has gained attention, several studies have examined its effects in different populations. In this thesis, the focus was on IT workers.

1.2.2.2 Yoga

The word “Yoga” has its origin in the Sanskrit root yuj, which means “to join” or “to unite”. This refers to the union of mind and body. It originated in ancient India, seeking to achieve the union of the individual self and the transcendental self [58].

The Western adaptation and manifestation include physical exercises and postures combined with the regulation of respiration and meditation [59]. The focus is mostly on isometric exercise and stretching.

Several studies have demonstrated the positive effects of yoga on overall health, particularly on stress regulation. For example, a study by Ross [60] comparing the benefits of exercise and yoga found that, in healthy and diseased populations, yoga was more effective than exercise in improving various health-related measures. Furthermore, some studies have gone beyond and researched its effects on specific brain regions. For instance, Li et al. [61] found that in different phases of practising yoga, long-term yoga practitioners showed higher blood oxygen levels in the dorsolateral prefrontal cortex compared with short-term practitioners. Participants also reported better task mastery and showed lower ventrolateral prefrontal activation. Finally, they

exhibited higher blood oxygen levels in the orbitofrontal and ventrolateral prefrontal regions compared to short-term practitioners.

This lays a more scientific ground for the benefits of practising yoga.

1.2.2.3 Journaling

Journaling is the practice of writing down thoughts, feelings, experiences, or reflections. When done reflectively, it encourages introspection and emotional discharge. Studies have found that it increases self-awareness, self-exploration and release of pent-up emotions [62]. It also improved individuals' physical and emotional well-being [63].

Journaling is commonly used in nursing studies to enhance reflective practice and as a technique for active learning. Among the reported benefits are the development of critical thinking, making connections between experiences, and discovering meaning [64]. Researchers have adapted journaling to digital formats, for instance, e-journaling. This format has the same effects as traditional journaling [65]. King and LaRocco [65] implemented e-journaling, achieving positive results with their students and demonstrating that the transition from paper to electronic does not hinder outcomes. In this thesis, we used e-journaling, as it was implemented in an online course; therefore, it was not possible to gather paper journals from around the world.

1.2.3 Emotions in SE

In this thesis, we used the American Psychological Association's (APA) definition of emotion [66]. Although numerous theoretical perspectives define emotions in different ways, this formulation was chosen because it explicitly integrates these experiential, behavioural, and bodily components into a single, coherent construct. The APA describes emotion as *"a complex reaction pattern, involving experiential, behavioural, and physiological elements, by which an individual attempts to deal with a personally significant matter or event"*.

Software engineering is an intensely human, cognitive, and social activity; hence, emotions are present in it. Engineers continuously engage in problem solving, learning, communication, and coordination, all of which are influenced by emotional states. At the individual level, emotions have been linked to other constructs such as productivity [67], personality [68], and collaboration [69], showing that emotions permeate to team and organisational levels. SE relies heavily on collaboration through activities such as code reviews, meetings, and issue discussions, where emotional expressions influence trust, conflict, and decision-making. Hence, understanding emotions in this specific context contributes to healthier teams, more sustainable work practices, and improved engineering outcomes.

With the previous in mind, we incorporated existing emotion frameworks into SE tasks. For example, for Paper B specifically, we used the Emotions Wheel created by Gloria Willcox [70] to guide the participants' emotions classification. This tool, frequently employed in therapeutic and self-reflective contexts, assists individuals in naming and differentiating their feelings with greater specificity. The wheel is organised in concentric layers: at its centre lie six foundational affective categories (happy, sad, angry, scared, strong, and calm). Progressing outward, each of these core

categories branches into increasingly fine-grained descriptors, offering a structured way to capture the nuance and complexity of emotional experience.

1.2.4 Qualitative Data Analysis in SE

Qualitative data is crucial in SE for examining the human and organisational dimensions of engineering work [71]. Through the close interpretation of rich, non-numerical evidence, researchers can uncover themes, explanations, and relationships that are not accessible through quantitative methods [71, 72]. SE qualitative datasets frequently combine technical records with materials centred on human experience. For instance, code review discussions, architecture decision logs, or incident communication channels alongside interviews or observational notes. Working with these types of sources explains the need for careful handling of dataset size, consistent analytic reasoning across researchers, and documented chains of interpretation to ensure the trustworthiness of the findings.

In SE studies, thematic analysis is one of the most popular methods to analyse qualitative data [73, 74] (see steps in methodology section 1.4.3.2). We chose the Braun and Clarke guidelines [75] for their clear and simple steps to develop a hybrid framework in Paper F. The framework proposes a collaboration between human researchers and Large Language Models (LLMs) to automate steps 2 to 5 of the analysis process.

1.3 Related Work

This section provides an overview of related work as a big scope og the main problem. The detailed related work is discussed in each paper separately.

1.3.1 Understanding Stress and Resilience in Software Engineers

In the general literature, primarily from psychology, various frameworks aim to explain the study, development, and measurement of well-being and resilience. Examples of these well-being frameworks are: Gallup's Five Elements of Well-being [76], Seligman's Five Pillars of Well-being [77] and Michaelson's Five components [9]. However, these frameworks are not population-specific. Software engineers, like many other subgroups, have specific characteristics that influence how well-being and resilience are experienced and maintained in their context. They also have specific stressors that need to be taken into account. Factors such as high cognitive demands, frequent task switching, collaborative yet often distributed work environments, and rapid technological change shape their stressors and coping mechanisms in unique ways. Consequently, applying general psychological frameworks without adaptation may overlook critical occupational, social, and organisational dynamics.

Previous studies, SE population-specific, have focused on specific areas. For example, emotions of: (un)happiness [8], feeling overwhelmed [78], frustration [79], reasons for negative emotions in agile contexts [80] and emotions recognition in software development [81]. Productivity, such as successful environments on software

teams [82], satisfaction and perceived productivity [83]. Additionally, there are other sub-areas, for example, inclusivity, empathy, and supportive work environments [84,85], everyday interpersonal challenges [86], and burnout [24]. However, there were no integrative frameworks that consider the interaction of these aspects when this thesis project began; currently, there are two besides our proposal, which are discussed next.

A comprehensive understanding of the well-being and resilience of software engineers, as well as their main stressors, requires moving beyond isolated constructs. It is essential to study different dimensions of interaction and include specific contexts. Wong et al. [87] proposed one of the first integration models explaining mental well-being, considering different levels (individual, team, and organisation) of interaction. Nevertheless, this framework is US-focused, which limits generalisability and fails to integrate different cultural contexts. In one step forward, Godliauskas and Šmite [88] conducted a literature review analysing 44 studies that reached populations from 42 countries. They proposed a theory about the Predictors and outcomes of software engineers' well-being. While this represents a significant step towards a multidimensional conceptualisation, the study was based solely on secondary data, relying predominantly on quantitative, cross-sectional surveys. As acknowledged by the authors, this limits causal inference and overlooks the nuanced, context-dependent experiences of software engineers. The analysis has limitations in capturing the lived experiences, interpretations, or mechanisms underlying the relationships identified.

Understanding well-being is a complex task that requires context-sensitive and empirically grounded insights. To address the previous limitations, this project's thesis conducted a mixed-methods study (Paper A) combining interviews and surveys. It also included participants from different contexts and countries. The data was triangulated to obtain statistical generalisability and in-depth contextual understanding. Moreover, it was then complemented and updated with a study (Paper B) focusing on the use of AI in daily SE tasks [89].

Addressing this gap enables the development of context-sensitive models and interventions that more accurately reflect the lived experiences of software engineers and support their sustainable well-being in the workplace.

1.3.2 Support and Enhancement of Resilience and Well-Being in the Software Engineering

A limited number of interventions within software engineering have explored well-being enhancement through mindfulness-based practices, yet the available studies consistently report positive effects. For instance, Heijer et al. [90] examined mindfulness in agile software teams through a two-month intervention involving short, three-minute mindfulness exercises during daily stand-up meetings. Conducted across eight companies and involving more than sixty participants, the study reported enhanced perceived effectiveness, decision-making, and listening skills. One limitation noted by the authors, however, was the use of non-standardised questionnaires. Research by Bernardez et al. [91, 92] has found that mindfulness interventions positively influence the mental well-being and self-perception of software engineers. Complementary to these findings, Romano et al. [93] investigated the effects of an eight-week Mindfulness-Based Stress Reduction (MBSR) programme among software developers from a multinational company in Italy. Participants who underwent MBSR reported

reduced stress and improved focus.

Collectively, these studies suggest that mindfulness-based interventions hold considerable potential for enhancing well-being and focus within software development contexts. Nevertheless, no more interventions have been done in SE contexts. One important characteristic of the previous interventions is the limited number of participants. In the appended studies of this thesis, we discussed the same challenges. Participation rates tended to decline over time, which, although common in longitudinal or behavioural interventions, still poses difficulties in maintaining engagement and ensuring sufficient statistical power. This attrition can influence the robustness and generalisability of findings, as participants who remain engaged may differ systematically from those who discontinue. To address the challenges, it is essential to design interventions that are flexible, minimally intrusive, and better integrated into existing work routines. These aspects, among others, are discussed in our yoga intervention (Paper D) study [93].

1.3.3 Research on the Human Factors in Software Engineering

Research on human factors in SE focuses on understanding how cognitive, affective, social, and organisational factors influence software development activities. Prior work has studied engineers' motivation [94, 95], personality [96, 97], and job satisfaction [83, 98]; cognitive load and program comprehension [99, 100]; collaboration and communication in teams [101, 102]; decision-making [103] and expertise development [104], creativity [105] among others.

Recently, researchers have started exploring stress, burnout, well-being [19, 27, 106, 107] and empathy [108]. These factors have been shown to influence individual productivity, software quality and long-term sustainability of development teams.

Studies in this area have increasingly adopted mixed-methods approaches to gain a deeper and more complete understanding of the subject.

Quantitative studies, such as large-scale surveys, have been widely used to identify relationships between psychological factors and productivity or well-being [8, 83]. However, these studies often provide limited contextual understanding. To address this, several authors have integrated qualitative methods, for instance, interviews and observations, to interpret developers' subjective experiences.

A growing body of work also incorporates physiological and behavioural measures to complement self-reported and qualitative data. Studies have combined electroencephalography (EEG), electrodermal activity (EDA), and heart-rate monitoring with surveys and interviews to investigate cognitive load, affect, and engagement during programming tasks [109–111].

The triangulation of data produces more comprehensive and ecologically valid insights into developers' experiences and perceptions. A central part of this triangulation is the analysis of qualitative data (QD). QD enables an in-depth examination of the non-technical dimensions of software development [71]. Through systematic interpretation of rich, non-numerical data, researchers uncover patterns, meanings, and insights [71, 72]. QD is especially valuable for understanding software processes, tool adoption, and organisational or technical contexts.

Given the potential of Large Language Models (LLMs) to process substantial amounts of textual data, several authors have explored their use for the analysis

of QD [28, 29, 112–114]. However, these studies face several challenges, including limited transparency and explainability, a lack of systematic evaluation on SE data, insufficient methodological rigour, and a narrow model scope. To address these limitations, this thesis proposes a human–LLM collaborative framework for thematic analysis (TA). This is the first study (Paper F) to incorporate tailored rubrics for evaluating the quality of codes and themes and to compare LLM-generated results with human-coded themes.

Thus, our contribution is to consolidate a triangulated approach that combines surveys, behavioural/physiological signals, and qualitative data (Paper E). Additionally, to extend the triangulation with a transparent implementation of LLMs as assistants in qualitative data analysis.

1.4 Research Methodology

This section explains the research methodology used in the studies presented in this thesis. Each study employed a mixed-methods approach to obtain a comprehensive, coherent picture of the researched phenomena.

Mixed methods research involves collecting quantitative and qualitative data and integrating them to get a comprehensive analysis of the research problem [115–117]. Scholars like Cresswell [116] claim that all methods have biases and weaknesses, and by combining more than one method, these biases can be neutralised. In this thesis, we applied a **triangulation strategy**, where numerical data from surveys and biometric data are enriched by the lived experiences captured in participant journals and interviews. Table 1.2 shows how each paper is linked to the thesis research questions, as well as the design and study methods. Each design, data collection and analysis method is explained next.

1.4.1 Research Designs

The research design serves as a broad structure or strategic framework that connects the research questions to the empirical data collection and analysis. It is based on the kind of explanation the researcher wants to deliver from the study [117].

1.4.1.1 Quasi-experiments (interventions)

Quasi-experiments are a type of experimental design used to investigate whether a direct causal link can be established between the independent variable (in this thesis context, an intervention) and the dependent variable [116, 118]. Quasi-experiments are positioned between the strict control of true experiments and the great flexibility of observational studies. It is often used when randomisation of groups and a control group cannot be implemented [119]. Hence, participants self-select into the treatment group. This presents challenges for internal validity, as the lack of randomisation means the groups may not be equivalent at the outset [119]. Pre-existing differences between the groups, known as selection bias, can confound the results, making it difficult to attribute any observed effect solely to the intervention [119]. Therefore, while quasi-experimental designs are efficient for real-world research, their conclusions

Table 1.2: Overview of the included papers in the thesis, the research questions they answer, their design, and their data collection and analysis methods. All the papers followed a mixed-methods approach.

Paper	RQ	Design	Data Collection	Data Analysis
Paper A: Well-being Factors	1: Factors 2: Approaches	Exploratory Sequential Design	Questionnaire & Interviews	Statistics & Thematic & Content Analysis
Paper B: Emotional Strain	1: Factors 2: Approaches 3: Methodology	Survey	Questionnaire (Open & Closed)	Descriptive Statistics & Content Analysis
Paper C: Breathwork Intervention	2: Approaches	Quasi-Experimental	Questionnaire & Journals	Bayesian Inference & Thematic Analysis
Paper D: Yoga Intervention	2: Approaches	Quasi-Experimental	Questionnaire & Focus Groups	Descriptive Stats & Thematic Analysis
Paper E: Multimodal Approach	1: Factors 3: Methodology	Experiment	EEG (Biometric), Questionnaire & Interviews	Statistics & Thematic Analysis
Paper F: LLM for Qualitative Analysis	3: Methodology	Experiment	Questionnaire (Open & Closed)	Statistics & Content Analysis

about causality must be interpreted with caution, acknowledging the potential for alternative explanations.

We used two interventions in the form of quasi-experiments in this thesis. To increase validity and rigour in our quasi-experiments, we followed the implementation guidelines by Maciejewski [119]. Both studies (Papers C and D) implementing quasi-experiments used a one-group pre-test and post-test design. That is, data was collected from the same single group at three time points: before the intervention (pre-test), during its implementation, and after its completion (post-test). The difference lay in the psychometric instruments used for data collection and the methods employed for data analysis. Both intervention programmes had a mindfulness practice as a core practice. Each programme is explained next:

Online Intervention Rise 2 Flow (R2F). It was designed to help build mental and emotional resilience and enhance well-being. It was based on a yogic breathing practice called Pranayama. This practice was combined with two other mindfulness practices, journaling and meditation. The technique is a three-part breath through the mouth, practised while lying down. A certified facilitator guided sessions.

We implemented two rounds of the R2F, lasting 12 and 8 weeks, respectively. All sessions were held online once a week. The participants were IT workers who joined from 26 countries. The data collected were quantitative, coming from day ratings and psychometric instruments (such as entry and exit surveys), and qualitative from participants' journals. For the quantitative analysis, we used Bayesian analysis, and for the qualitative data, we used thematic analysis by Braun and Clark's [120] guidelines. The results are published in Paper C.

In-person Industry Intervention. The yoga intervention was done in a software development company. It was a weekly practice for eight weeks. The target population were software engineers. Participants had a 45-minute Hatha yoga session every Wednesday from 8:00 to 8:45, taught by a yoga instructor. These sessions focused on the principles of Hatha yoga, incorporating physical postures, breathing exercises, and relaxation techniques (5 min). The data collection was done using psychometric instruments to create a survey and obtain quantitative data. Additionally, for qualitative data, we organised a focus group. The results were published in Paper D. The data analysis was conducted using descriptive statistics and thematic analysis, as described by Braun and Clarke [120]. The results were published in Paper D.

1.4.1.2 Survey

Surveys offer a quantitative description of trends, traits, attitudes, and opinions of a population [116, 121]. By systematically collecting responses from a defined sample, surveys enable the aggregation and statistical analysis of self-reported data. This approach is particularly suitable for examining patterns across participants and for comparing responses across predefined variables or conditions. In this thesis, a survey research design was employed to gain an understanding of participants' perceptions, opinions, and psychological constructs in various study contexts. By using a survey design, we could identify patterns, averages, and correlations.

1.4.1.3 Experiment

To examine the relationships between variables, we employed controlled experiments. In experimental studies, one or more independent variables are deliberately manipulated in order to observe their effects on dependent variables while holding other factors constant [122]. We carefully controlled task conditions, standardised instructions, and consistent evaluation procedures. These guidelines informed the selection of variables, the structuring of experimental conditions, and the interpretation of results. For the experiment involving human participants, we followed established experimental design principles as outlined by Brysbaert [123], including careful control of task conditions, standardised instructions, and consistent evaluation procedures. For the experiment involving large language models (LLMs), we aimed to compare human and LLM outputs in the same task. The goal was to identify and measure similarities and differences in the performance of a specific analysis method.

1.4.1.4 Exploratory Sequential Design

This design employs a two-phase mixed-methods research approach, where the researcher collects and analyses qualitative data first to explore a topic, identify key themes or variables, and develop a theory, framework, or instrument [116]. Then, the qualitative results directly inform and guide the following quantitative phase (a survey or experiment) to test, generalise, or validate those initial findings on a larger scale. In this design, it is essential to focus on the appropriate qualitative findings to build a solid and useful foundation for the second phase (quantitative), and to select the correct sample and analysis methods.

1.4.2 Data Collection Methods

In this section, the specific techniques and instruments used to gather evidence as dictated by the research design are explained.

1.4.2.1 Questionnaires

We primarily used two different types: the first was creating the questionnaire from scratch, following Stol and Fitzgerald guidelines [124]. This first type employed open-ended, closed-ended, and Likert questions and was for exploratory purposes. The second type was a questionnaire made of psychometric instruments. This type was mainly used to measure changes after an intervention.

Psychometric instruments are tools designed to measure psychological constructs, attitudes, and behaviours in a systematic and quantifiable manner [125]. The main reasons to use them in this thesis were: validity, which refers to the accuracy with which an instrument measures the intended construct; reliability, reflecting the stability and reproducibility of measurements over time; and responsiveness, indicating the instrument's sensitivity to detect meaningful changes [126].

1.4.2.2 Interviews and Focus Groups

To gather views, opinions and perceptions of our study's participants, we used semi-structured interviews [116]. We chose the semi-structured format to allow our participants flexibility in expressing themselves and capturing unexpected yet valuable insights. Two different semi-structured interviews, open questions in questionnaires, and one focus group were used. Since the studies were exploratory, aiming to understand participants' experiences, perspectives, and emotions, **interviews** were the most suitable approach [117]. The **focus group** was chosen since we had two goals for that study. First, we wanted to know our participants' experiences in the intervention. Second, we aimed to investigate the interaction between the intervention's organisers. This method enabled a dynamic exchange of ideas, reflections, and shared experiences [127]. It provided deeper insights into the collective understanding of the intervention's design, delivery, and perceived impact.

We designed semi-structured interview guides to conduct the interviews and the focus group. This type of interview guide allowed respondents to expand on their answers, and me, as an interviewer, to ask follow-up questions and explore topics in depth.

1.4.2.3 Biometric Data

Biometric data is unique information about a person's physical (fingerprints, face, iris), physiological (heart rate, DNA), or behavioural (voice, typing rhythm, gait) characteristics [128]. We collected Electrodermal Activity (EDA) and Heart Rate Variability (HRV) using a wearable wristband and Electroencephalogram (EEG) using a Neurosity Crown device for one of the studies.

Human factors and their characteristics are a core part of all the studies in this thesis; hence, obtaining reliable and objective data is essential for our studies [129]. Therefore, we decided to collect biometric (EDA, EEG and HRV) information to go

beyond self-reported answers. Our goal was to understand participants' responses, emotions, and cognition in a more in-depth and accurate manner.

1.4.3 Data Analysis

This section explains how the two types of data, qualitative and quantitative, were analysed in the included studies.

1.4.3.1 Quantitative Analysis Method

We used the quantitative data in two ways: first, to identify trends, traits, and overall perceptions, which were presented as means, medians, modes, and standard deviations. For this goal, we used descriptive statistics to summarise and describe our datasets. We also used these summaries to create visualisations that give readers an overview of our results. Second, we applied inferential statistical methods to analyse changes in responses over time at multiple temporal points(entry vs. exit, daily, and weekly trends).

We employed non-parametric and parametric frequentist tests to seek differences between groups. Similarly, we also studied the relationship among variables (Mann–Whitney U, Kruskal–Wallis and Spearman). We also performed Bayesian analyses to examine responses in three ways: (1) temporal analysis for each instrument at t0 versus t1 (entry vs. exit), (2) temporal analysis of daily trends, and (3) temporal analysis of weekly trends.

1.4.3.2 Qualitative Analysis Method

For the qualitative data, we used two methods: **Reflexive Thematic Analysis** and **Content Analysis**.

For **Reflexive Thematic Analysis**, we followed Braun and Clarke's [75] guidelines, which consist of six steps, as explained below.

1. Familiarisation with the Data: Reading and re-reading the data.
2. Generating Initial Codes: Systematically going through the data to label meaningful segments (codes) relevant to the RQs.
3. Generating Initial Themes: Grouping and organising the codes into broader, potential themes that reflect meaningful patterns across the dataset.
4. Reviewing Themes: Evaluate the potential themes against the coded data, the entire dataset and the RQs. Refine, combine, split, or discard themes to ensure they are coherent and distinct.
5. Refining, Defining and Naming Themes: Identify the themes' central idea, define and name each theme and evaluate the general structure.
6. Writing the Report: Write the story following the themes' narrative and using quotes to answer the study's RQs.

Similarly, for **Content Analysis**, we followed the guidelines by Kuckartz and Radiker [130]; the seven steps are explained below.

1. Initiating Text Work: Read and highlight important passages and write "case summaries" to grasp the overall context.
2. Developing Main Categories: Create broad codes based on the study's RQs.
3. First Coding Cycle: Code the entire material using these broad main categories.
4. Inductive Sub-categorisation: Create sub-categories within the main categories, focusing on the central categories for the study.
5. Second Coding Cycle (coding data with sub-categories): Re-code the entire dataset using the now-complete system of main and sub-categories.
6. Simple and Complex Analysis: We chose to create visualisations and data display in this step.
7. Analysis and Presentation: Compare categories across cases and write up your results.

1.4.4 Reflexivity

I am a PhD student with a background in behavioural and social sciences. My background shaped my epistemological positioning, choice of research questions, preference for mixed-methods approaches, and emphasis on participants' subjective experiences.

I acknowledge that my lack of a software engineering background made interpreting domain-specific terminology, situating findings within software engineering practices, and articulating technically grounded practical implications challenging at the beginning of my PhD. I relied on my supervisors' technical knowledge, professional experience and insights to support interpretation and contextualisation.

Through prolonged engagement with software engineering research communities, repeated interaction with practitioners, and sustained immersion in the domain throughout the PhD, my positionality shifted. While I remain professionally grounded in social and behavioural sciences, I have developed substantial domain familiarity and sensitivity to software engineering practices, norms, and constraints. As a result, I came to occupy a hybrid insider–outsider position, functioning as a domain insider regarding software engineering culture and concerns, while retaining an external disciplinary perspective.

This hybrid positionality had methodological implications. My background supported sensitivity to affective, cognitive, and well-being-related aspects of participants' accounts. Meanwhile, my developing domain knowledge helped me with a more nuanced interpretation of software engineering-specific practices and constraints. At the same time, maintaining an external disciplinary stance supported critical questioning of the field's assumptions. Reflexive engagement with this shifting positionality was therefore central to data interpretation, analytic decision-making, and the development of practical implications.

1.5 Ethical Considerations

This thesis topic core was human factors; hence, having humans in all the studies was imperative to gather data and evidence to create a basis for informing the development of guidelines, policies, interventions, and programmes. To ensure the safety and privacy of participants in studies involving humans, oversight by an Institutional Review Board is necessary. For the included studies, we consulted the Swedish Ethics Review Authority (etikprövningsmyndigheten). We obtained ethical approvals from etikprövningsmyndigheten [131]. We also followed the guidelines of Chalmers University for ethical research and The National Institutes of Health's seven principles of ethics for human subjects research [132]:

1. Social and clinical value. Our research aims to improve the well-being of our target population, software engineers.

2. Scientific validity. We employed rigorous and appropriate scientific methods that contributed to the body of evidence for the SE field.

3. Fair subject selection. We invited a diverse range of software engineers to participate in our studies. We did not limit participation based on age, sex, race/ethnicity, or sexual orientation.

4. Favourable risk-benefit ratio. We tried to minimise any risks or discomfort for our participants. For example, in the interventions, we informed participants of potential discomfort, such as a dry throat, and provided them with advice on what to do after the intervention.

5. Independent review. We review the interventions and their corresponding activities. We also piloted all survey and interview guides to prevent any misunderstandings or identify activities or questions that could be considered socially, racially, and/or ethnically inappropriate.

6. Informed consent. For all studies, we collected written informed consent from all participants. They received an explanation of the study's objectives, methods, and potential risks, as well as their right to withdraw from the study at any time.

7. Respect for potential and enrolled subjects. To keep our participants' privacy, all data was anonymised and securely stored. During the interventions and experiments, we monitored our participants to ensure they were not experiencing any discomfort.

We also offered compensation without undue inducement. Participants in Paper C received a donation to a charity of their choice as a token of appreciation. Similarly, participants in Paper E received a voucher for a meal or drink, while those in the control group of Paper D received a gift card. We followed the arguments by Wikilson and Moore [133] and concluded that gift cards did not bias or put our participants at risk.

1.6 Research Results

This section presents the results in the form of a summary of the included papers. Each paper starts with the main goal or motivation, then an overview of the findings and finally the implications. The full papers are in the coming chapters. The papers follow the argument presented in the Research Focus Section 1.1.

Paper A - Well-being Factors

The primary motivation of this paper was to empirically explore the factors, from the software engineers' perspective, that influence their well-being. We wanted to gather experiences from engineers from various parts of the world to gain a comprehensive picture and also to make our model contextually relevant.

The methodology that best suited this paper's goal was a mixed-method combining surveys and interviews. We interviewed 16 software engineers in Sweden to get deep insights and experiences in their daily work. Then, we created the survey in three languages (English, Spanish and Portuguese) to reach a large number of respondents to examine whether the themes identified in the interviews were reflected more broadly across the international sample.

Findings: A framework that identifies the main factors shaping well-being, that considers individual perceptions of well-being, interpersonal and collaborative relations, workplace support and recognition, organisational culture, and stressors arising within software engineering. The framework is presented in Figure 1.2.

Implications: This paper provides policy guidelines and recommendations for organisations to support the well-being of engineers. For research, our framework can inform the design of well-being interventions and future empirical studies.

Paper A's findings align partially with other well-being frameworks, such as Gallup [76], the Perma model [77], and Michaelson's framework [9]. However, Paper A focuses on software engineers and considers the characteristics of their working context. Regarding previous works on the SE context, Paper A considered engineers' answers from different countries, in contrast to Wong et al. [87], whose work focuses on the USA. Furthermore, Wong et al. examined internal self-reported well-being experiences, while Paper A also considers external factors such as company culture and peer support. Moreover, Paper A's results are based on survey and interview data, unlike the work of Godliauskas and Šmite [88], which relied on literature reviews. This difference is substantial because primary data offers a more direct and contextually grounded basis for developing a well-being framework than secondary synthesis alone.

This paper contributes to answering RQ1 by defining the first of the proposed frameworks, which explores the factors influencing the well-being of engineers as a distinct population. By explicitly grounding these influences in a SE context, the framework points to what should be considered relevant when studying well-being-related phenomena in this domain. It also answers RQ2 by proposing guidelines for organisations and policy recommendations to support the well-being and good practices.

Paper B - Emotional Strain in LLM Interactions

Paper B extends the conceptual framework of Paper A into the technological

sphere, demonstrating how AI-based tools can introduce an additional layer of stress. Although the technological aspect was mentioned in Paper A, given the current updates and state of AI, it was necessary to explore this area in greater depth. In this paper, the source of strain shifts from workload or interpersonal tension to the interaction between human expectations and machine behaviour.

We used a survey to gather engineers' experiences interacting with LLMs. We obtained 62 answers and analysed them using content analysis. We used Wilcox's Emotions Wheel to categorise participants' answers and conceptualise their emotions.

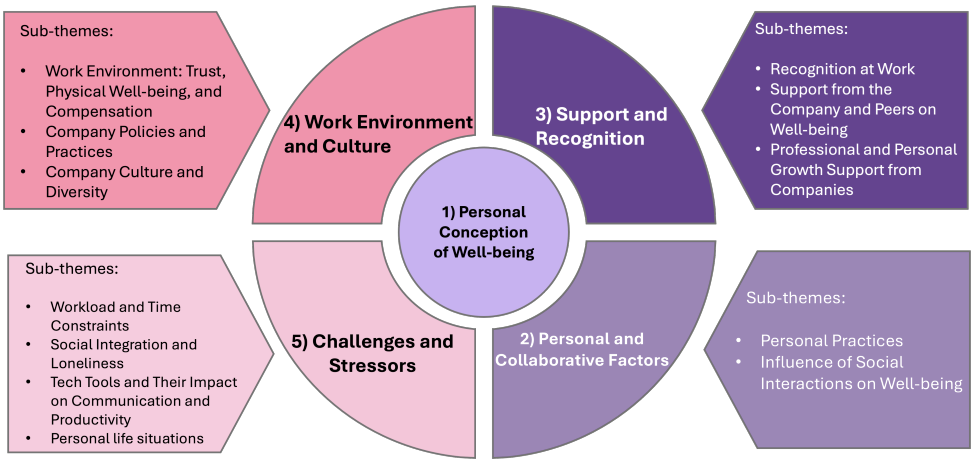


Figure 1.2: Themes from Paper A showing the framework of factors influencing the well-being of software engineers. These factors were later compared to the survey answers.

Findings: Software engineers using LLMs in their daily tasks reported distinct emotional responses. These emotional answers ranged from curiosity and satisfaction to frustration, disappointment, and guilt when these systems produced incorrect, misleading, or verbose outputs. Several participants' experiences were manifestations of techno-frustration [134], a specific form of techno-stress driven by perceived inefficacy, cognitive dissonance, and loss of control during digital interactions. Importantly, participants displayed adaptive resilience strategies, such as refining prompts, cross-verifying outputs, or switching tools. However, repeated failures or "hallucinations" led to cumulative strain.

Implications: Based on our results, we proposed recommendations for designing tools that reduce stress associated with user interaction. Similarly, we argue that companies need to prioritise employees' emotional intelligence training to cope with techno-stress.

The psychological effects of using LLMs are a relatively new area of research. Most studies focus on how LLMs are used at work [20, 135, 136], and how they influence workers' efficiency and efficacy [137, 138]. However, their impact on emotions remains largely underexplored.

Paper B is the first study to investigate emotional responses in LLM–human

interaction specifically. Recently, only the work by Maitipe [139] explored the psychological impact of LLMs on IT professionals. However, the emotions involved in this interaction are not yet well studied, and their influence on workers' well-being and performance remains largely unexplored. Emotions have a direct impact on stress, satisfaction, and overall mental health. Understanding how LLMs influence feelings like frustration, anxiety, or confidence helps design systems that support users rather than harm their well-being [89].

Paper B contributes to answering RQ1, RQ2, and RQ3 by extending the investigation of well-being and stress into the technological domain through the lens of human–LLM interaction. It identifies how AI-based tools introduce new forms of techno-frustration and cognitive demand. For RQ2, Paper B proposes design-oriented and organisational recommendations to mitigate emotional strain and support sustainable tool use in practice. Furthermore, Paper B contributes to RQ3 by exemplifying the adaptation of a psychological instrument for emotion classification to the SE context.

Paper C - Breathwork Intervention

This paper presents the results of a breathwork intervention. The intervention was the implementation of R2F p.14, a programme designed to teach breathwork to IT workers in weekly online meetings. The goal was to help participants manage stress, increase well-being, and develop resilience. R2F's Thursday sessions followed this structure: 1) It started with participants answering questions on the weekly self-development topic. They received the topic on Monday and had time to reflect and answer the questions. 2) The breathing practice for three rounds of seven minutes. 3) 20-minute relaxation. 4) Aftercare suggestions (e.g., to hydrate well) and time for participants' questions. The data collection was done before (with the entry survey), during (with a written journal and a weekly survey) and after the intervention (with an exit survey). The entry and exit survey was created with the following psychometric instruments:

- Mindfulness Attention Awareness Scale (MAAS) [140]
- The Scale of Positive and Negative Experience (SPANE) [141]
- The Psychological Well-Being scale (PWB) [141]
- The Positive Thinking Scale (PTS) [141]
- Perceived Productivity instrument (HPQ) [142]
- Self-Efficacy instrument [143]

Thematic analysis was used to analyse the qualitative data, and a temporal analysis was conducted for the quantitative data (for each instrument, comparing entry vs. exit surveys, as well as daily and weekly trends).

Findings: The results indicated that the R2F programme may help improve participants' mindfulness attention awareness, well-being, and self-efficacy.

Implications: We identify three types of implications. For policy, organisations need to create concrete actions on mental health awareness. For research, our programme proved to be effective; more modalities need to be tested and adapted to in-person interventions. For practice, modelling healthy well-being habits is more effective than only talking about them. Hence, managers and teachers should

demonstrate these habits in their daily work to foster a healthier and more supportive environment.

Paper C presented an online intervention, which enabled participants from various locations around the world to be reached. Unlike other in-person interventions [92, 93, 144, 145], the online format allowed participants to practise breathwork to manage stress more effectively and increase their well-being and resilience. This was particularly important since the interventions were done during pandemic times. In the software engineering context, no other study has targeted this population in a large-scale, remote intervention that combines stress-management techniques with outcomes related to well-being and resilience.

This paper addresses RQ2 by empirically evaluating R2F to support stress regulation and well-being among software engineers and IT workers. It provides evidence on how structured, recurring practices can foster individual resilience and emotional regulation. The findings inform how such interventions can be designed, measured, and implemented in real-world settings, explaining their potential benefits and practical considerations for sustained engagement.

Paper D - Yoga Intervention

This paper presents the second intervention appended in this thesis. To measure the effectiveness of yoga in improving general well-being among software engineers, we implemented an eight-week yoga programme. The intervention was done in collaboration with a Swedish software company. We collected quantitative data using the following psychometric instruments as entry and exit surveys:
- The Schutte Self-Report Emotional Intelligence Test (SSEIT) [146]
- The 14-Item Resilience Scale (RS-14) [147]
- Short Form Self-Regulation Questionnaire (SSRQ) [148]
- Self-Transcendence Scale (STS) [149]
- The Flourishing Scale (FS) [141]
- Brief Resilient Coping Scale (BRCS) [150]

Additionally, we used the WHO-5 Well-being Index [151] as a weekly survey and a focus group with the organisers to collect qualitative data. One of the objectives was to have a control group, which was not possible due to the small number of volunteers.

Findings: Results from the psychometrics did not reveal any statistically significant differences between the entry and exit surveys. However, the qualitative results showed participants experienced positive effects after the sessions. Our conclusions focused on how contextual factors, in this specific case, layoffs, critical deliveries, and non-work demands, time pressure, emotional intensity, and schedule disruptions associated with Christmas celebrations can mitigate the positive effects of yoga.

Implications: We shared lessons learned that can inform future mindfulness interventions in the workplace. Particularly, the importance of tailoring interventions to consider the context and unique needs of participants is one of the main implications.

Paper D is the first study involving software engineers practising yoga. Previous interventions in software engineering have mainly used mindfulness as the primary practice. For example, Bernardez et al. [92] ran three controlled experiments with students practising mindfulness and focusing on conceptual modelling. Paper D

instead involved engineering practitioners and investigated general well-being, making the context and population markedly different. One more study by Romano et al. [93] introduced a Mindfulness-Based Stress Reduction (MBSR) program. However, they only focused on collecting qualitative data. In contrast, Paper D employed a mixed-methods design, combining quantitative data from validated psychometric instruments with qualitative insights from focus groups. This provided a more robust and comprehensive assessment of the intervention's impact. Finally, Bernardez et al. [144] continued their experiments by having software workers participate in mindfulness practices. In Paper D, yoga included mindful awareness but extended beyond it through physical postures and breath-based exercises, offering a more comprehensive approach than mindfulness alone.

Paper D also contributes to RQ2 by complementing Paper C's findings by illustrating how contextual and organisational factors can mediate the effectiveness of individual-level well-being interventions. The lessons learned from this paper aim to strengthen approaches to fostering sustainable well-being. They also refine the understanding of the conditions necessary for well-being interventions to succeed in practice.

Paper E - Multimodal Methodology

The motivation of this paper was to introduce a physiological dimension by utilising biometric data to measure stress, mental workload and emotional responses during programming tasks. We designed an experiment to expose participants to stress associated with limited time constraints while programming.

We combined three data sources to collect data during and after the programming tasks:

- Biometrics: EEG (electroencephalography), EDA (electrodermal activity), and HRV (heart rate variability) sensors.
- Validated psychometric scales: Perceived Stress Scale (PSS-10) [152], Short Stress State Questionnaire (SSSQ) [153] and NASA Task Load Index (NASA TLX) [154].
- Interviews.

The data was analysed using thematic analysis for the interviews and descriptive analysis and T-tests for the quantitative data.

Findings: The psychometric results did not show any differences when comparing the tasks with and without time limitations. However, while participants claimed to feel relaxed or neutral, EDA data showed micro-level spikes in arousal, especially during time-constrained tasks.

Implications: Based on our findings, we proposed guidelines and considerations for research in stress, mental workload and emotions in SE. Paper E's findings challenge the reliability of self-report instruments in isolation. It also emphasises that stress in SE can be non-conscious or cumulative, accumulating over repetitive micro-stressors like debugging or code compilation delays. Hence, it also evidences the importance of multimodal methods for accurately capturing human experiences in technical contexts.

Paper E contributes to answering RQ1 and RQ3 by advancing the understanding of stress-related phenomena, revealing discrepancies between self-reported experiences and physiological indicators, and demonstrating that stress may be subtle, cumulative,

or non-conscious. Methodologically, the paper addresses RQ3 by proposing guidelines for integrating biometric data, psychometric instruments, and qualitative interviews, thereby strengthening the rigour and interpretive depth of empirical human-factors research in SE.

Paper F - LLMs as Analytical Assistants

Paper F aimed to advance qualitative data analysis methodology by integrating LLMs as analytical assistants. To achieve this, we designed an experiment to compare human and LLM-generated steps 2 (creating initial codes) to 5 (creating, naming, and refining themes) of thematic analysis by Braun and Clarke. We used 15 interviews from a previous study that had already been coded and analysed by human researchers. First, we created a prompt for the LLM to create initial codes. We compared those codes with the human-made and asked external experts to evaluate them using a tailored rubric. Then we had the LLM create themes and also compared them with those of the human researchers. These themes were also systematically evaluated using a rubric designed based on Braun and Clarke's guidelines.

Findings: Evaluators preferred LLM-generated codes 61% of the time over the human ones. They found them analytically useful for answering the research question. However, evaluators also pointed out the limitations of LLM codes and themes.

Implications: - A reproducible approach integrating refined, documented prompts with an evaluation framework to operationalise Braun and Clarke's reflexive TA.
- An empirical comparison of LLM- and human-generated codes and themes in software engineering data.
- Clear guidelines for integrating LLMs into qualitative analysis, preserving methodological rigour.

Figure 1.3 illustrates our proposal for integrating LLMs in thematic analysis, specifically delimiting LLM and human activities. LLMs act as assistants in steps 2 (creating initial codes) and 3-5 (creating, naming, and refining themes). Researchers supervise, evaluate and refine the steps done by LLMs. Steps 1 (familiarising with the data) and 6 (writing the final report) remain entirely done only by the researcher.

Despite several studies aiming to automate qualitative analysis methods, either partially or fully [28, 29, 112], Paper F extends the literature by evaluating LLM-generated outcomes in terms of interpretive depth, theme coherence, and alignment with the RQs. In this paper, we also provided rubrics to assess the quality of codes and themes, which can be applied to data generated by LLMs or by human researchers.

Paper F addresses RQ3 by advancing methodological strategies for qualitative data analysis in SE through the integration of LLMs as analytical assistants. The study contributes concrete guidelines for maintaining methodological rigour, transparency, and reflexivity when incorporating AI into qualitative research workflows.

1.7 Discussion and Answers to the RQs

This thesis used mixed-method and multi-modal studies to integrate psychological aspects, organisational perspectives, and SE practice to explain and theorise how stress, resilience, and well-being are experienced and shaped in contemporary software

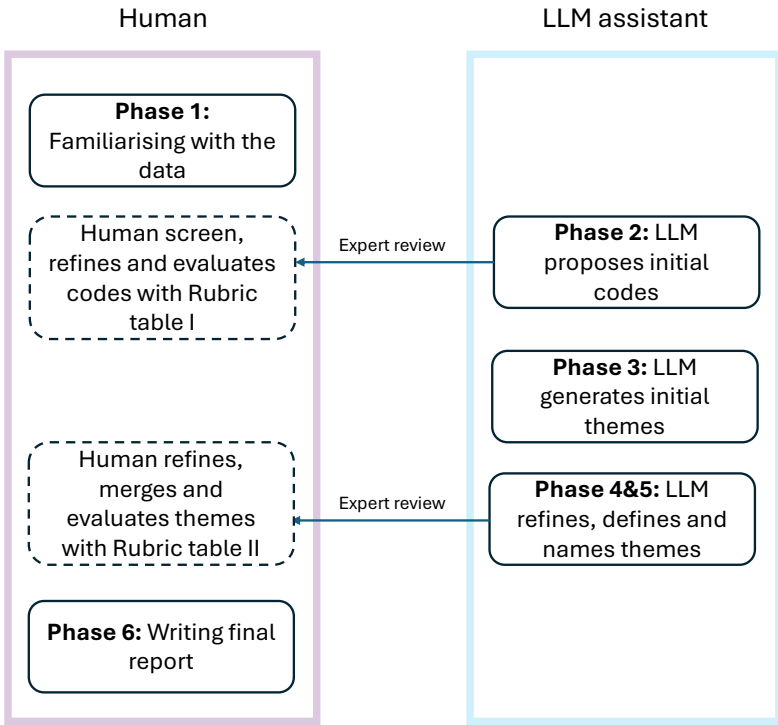


Figure 1.3: Proposal of implementation of LLM in thematic analysis (TA). LLM contributes to TA Phases 2–5 as an assistant; the human leads Phases 1 and 6 and gates progression using rubric-based evaluations. Dashed boxes indicate areas that require human evaluation and refinement.

engineering contexts. It conceptualises software engineers' well-being as an emergent property of interacting individual, social, and technological systems. Our results from empirically examining mindfulness-based interventions, emotional strain in human–LLM interaction, and the limitations of single-source self-report data show the potential and constraints of current approaches to supporting engineers' well-being.

The next paragraphs elaborate on the thesis's contributions to the state of the art in software engineers' well-being, resilience and research on human factors.

1.7.1 Factors and conditions that influence stress, well-being, and resilience

Results from papers A, B, and E answer RQ1 and show that well-being and stress in software engineers are multifactorial. Several factors from different contexts and systems interact and shape how software engineers feel and cope with stress. These factors operate at various levels: individual, interpersonal, and organisational. This makes them a bioecological and sociotechnical phenomenon, where individual

characteristics and coping strategies are continuously influenced by social relations, organisational conditions, and the technological ecosystem in which work is taking place [155, 156]. Figure 1.4 shows the Individual–Social–Technological System (ISTS) integrated framework proposed in this thesis, which presents the factors influencing well-being and stress management in software engineering, derived from Papers A and B.

The ISTS framework (in Figure 1.4) shows well-being as the outcome of a balance between stressors and resources arising simultaneously from three interdependent systems: individual resources, social and organisational environment, and technological ecosystems.

We believe that well-being is a product of continuous interaction between personal capacities, social structures, and technological conditions. It is dynamic, interactive, cyclical, and domain-crossing, and it changes under the influence of any of the three spheres. Hence, any changes in any one system can ripple into the others, reshaping the overall balance. Several elements belong to more than one system. For example, mental workload is both an individual resource and part of the technological ecosystem, as both areas add to how engineers experience and manage their cognitive demands.

The ISTS framework's spheres are:

Individual Resources: This domain captures personal factors (habits, physical and emotional practices, life circumstances and the individual's conception of well-being) that shape people's starting point for managing stress. This sphere is also present in several other well-being frameworks; however, we give more importance to the work-related aspects, for example, we consider the mental workload.

Social and Organisational Environment: This sphere addresses the relationships, mainly at work, of engineers. Specifically, it considers social interaction and integration, company policies and culture, company and peers support, and recognition. Our framework considers these aspects as direct influences on well-being, rather than as background context or indirect influences, as in Michaelson's [9] and Seligman's [77] model.

Technological Ecosystem: For this sphere, we consider task demands, digital tools, automation and AI/LLM interactions, and technology in general as mediators in workflows. They also shape cognitive load, time pressure, attention fragmentation, and the overall stress in engineers' work. The technological ecosystem introduces constraints and affordances that can either amplify or reduce stress, making it an essential component of any well-being framework in software engineering. We use the term 'ecosystem' to emphasise the interrelation and interdependency of the system's component elements. Despite its relevance, this sphere is not addressed in key well-being frameworks such as those proposed by Seligman [77], Gallup [76], and Michaelson [9].

Previous well-being frameworks [9, 76, 77] focus on the general population and view well-being as either a psychological or life domain outcome, but rarely as an integration of both. We consider the interactions of the primary systems present in software engineers' lives.

Moreover, these frameworks are also context-independent. In contrast, our framework considers SE-specific context characteristics, such as sustained cognitive load, sociotechnical collaboration, rapidly evolving technologies, chronic time pressure, and

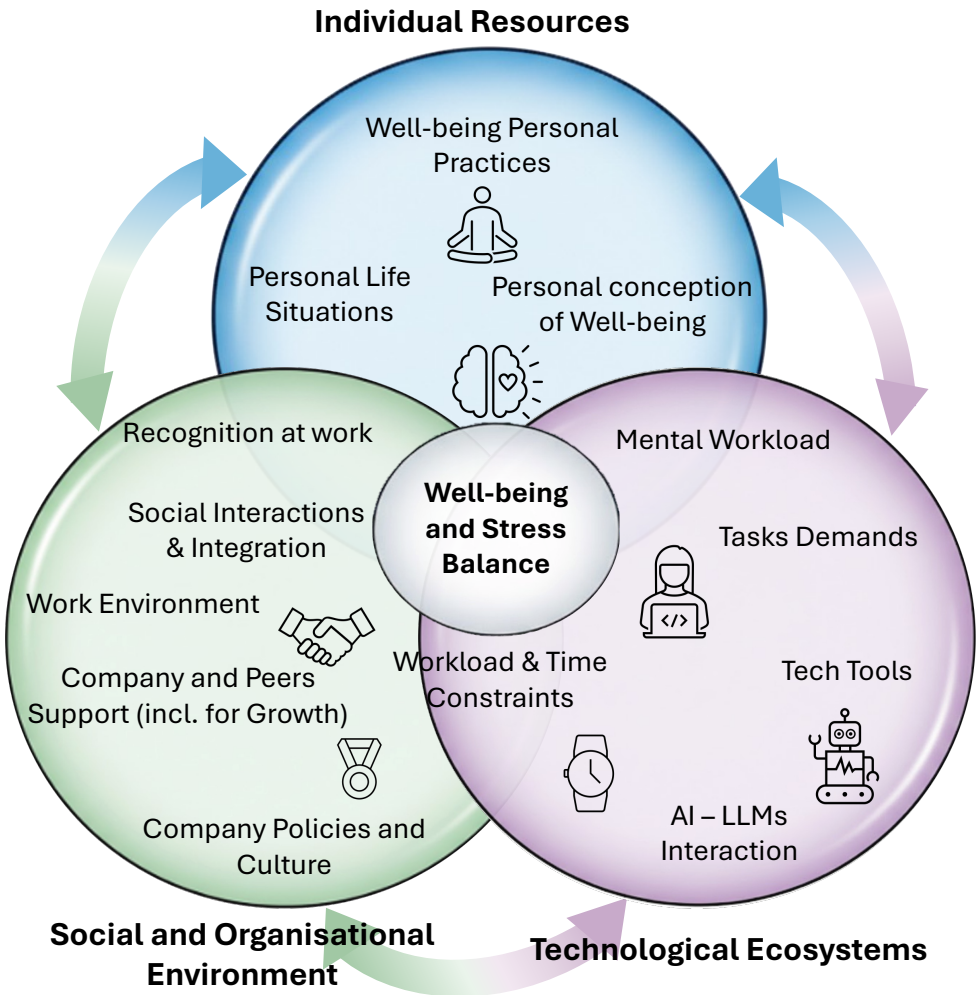


Figure 1.4: The Individual–Social–Technological System (ISTS) integrated framework of well-being and stress balance in software engineering. The framework illustrates how individual resources, social and organisational environments, and technological ecosystems jointly shape well-being and stress. The arrows indicate dynamic, cyclical interactions, whereby changes in one domain can propagate across the system.

particular cognitive demands.

Our individual-social-technological system views well-being as an emergent property, shifting the trend in traditional frameworks from seeing it as a set of dimensions [157]. For example, in our framework, individual habits influence how technology is used. At the same time, technology shapes organisational expectations. This explains how well-being arises within an individual-social-technological system, and not

only what well-being looks like for engineers.

At a theoretical level, the ISTS framework proposes the following claims to explain how well-being emerges from the configurations and interactions among its spheres.

Specifically, the framework implies that (i) the effectiveness of individual regulation strategies is contingent on organisational enabling conditions, (ii) technological systems influence well-being indirectly by reshaping the distribution of cognitive and temporal demands across levels, and (iii) stress and resilience outcomes are shaped by alignment or misalignment between individual capacities, organisational structures, and technological practices. These claims articulate boundary conditions and interaction logics that can be operationalised and tested in future empirical research.

In Paper A, we explained that well-being depends on how these conditions converge to either balance or overload the engineer's mental and emotional resources [106]. For example, engineers are prone to experience higher well-being when they are recognised at work, have clearly defined roles, and work within inclusive and transparent leadership [158, 159]. However, when the situation is different, with deadlines tightening, interruptions multiplying, and context switching becoming continuous, the influences of this particular context can challenge individual resources to cope with stress [106]. Furthermore, the constantly changing technological context adds to the previous overview. The use of AI, particularly the increased interaction with LLMs, introduces new stressors that differ from traditional forms of technical frustration [89, 160]. Errors, hallucinations, or mismatched outputs require engineers to rephrase prompts, verify content, and correct the system's responses. This verification work becomes an additional source of demand that fragments attention and increases effort, especially when it must be performed under time pressure. The accumulation of such microfailures generates sustained emotional friction, contributing to fatigue over time.

Different factors influence at various levels, yet their effects are not simply additive. Individual traits such as resilience or self-awareness may buffer stress, but their effectiveness depends on whether organisational and social contexts enable their expression [106]. For instance, supportive leadership or psychologically safe environments can transform individual coping efforts into collective resilience. On the contrary, rigid hierarchies or conflicting demands can neutralise even strong personal resources. This interconnectedness implies that well-being in software engineering cannot be fully supported through individual-level interventions alone.

Structural and cultural aspects, such as unclear expectations, unstable workflows, and performance-driven norms, can perpetuate stress regardless of personal coping abilities [155]. Therefore, improving well-being and resilience requires systemic adjustments that align organisational practices, interpersonal dynamics, and individual strategies [106]. Stress recovery and coping are thus a manifestation of how well the surrounding system supports sustainable human functioning within the technological and social realities of software development.

One more factor to consider is the "not so visible" manifestation of stress [161]. In Paper E, the results indicated that stress, as measured by self-reported instruments, may not be accurately captured. However, electrodermal activity showed more frequent phasic peaks, indicating subtle physiological arousal. Since software engineers often work under high mental workload, they may not consciously recognise their stress. Additionally, it is important to consider variability across individuals. In Paper E's

experiment, not all participants displayed the same physiological sensitivity. Hence, stress cannot be inferred solely from uniform group-level trends.

Well-being and resilience are maintained when organisational systems and technologies work together to stabilise demands [106, 162]. This prevents mental workload from escalating into sustained physiological strain. Resilience, then, is the visible outcome of an environment. This environment spans individual, organisational, and team levels, providing opportunities for recovery. It also enables software engineers to sustain performance without incurring hidden physiological costs.

Understanding this interdependence is essential for designing work environments and technologies that foster sustained well-being and resilience in software engineering.

1.7.2 Approaches to foster sustained well-being

Building on the previous question, a coherent program for sustained well-being must consider three key areas: individual regulation capacity, organisational context [163] and AI interaction.

At the individual level, Papers C and D investigated the effectiveness of mindfulness-based practices in cultivating resilience and emotional regulation. Paper C provided evidence that breathing practices had a positive effect on participants when done weekly. Notably, the study also demonstrated that sustained, structured engagement is essential. Short-term exercises produce transient effects, whereas weekly facilitated sessions promote the internalisation of well-being habits.

However, the results from Paper D (Yoga Intervention) were not as positive as those from Paper C. This study exposed a central challenge in workplace interventions, contextual interference from ongoing work stressors and organisational culture. Participants worked on their individual well-being but were overcome by the work context. They reported positive subjective experiences in qualitative data. These perceived benefits were strongly influenced by organisational and temporal context, including workload pressure and external stressors. Hence, sustained well-being also depends on organisational integration [164] to create organisational conditions that support individual participation.

Paper A complements these findings and suggests organisational strategies and policies that strengthen interventions. Multi-level interventions are more effective than isolated wellness initiatives. Furthermore, Papers B (LLM Frustration) and E (Biometrics) point toward technological design interventions. Paper B argues for "emotionally intelligent" digital tools that recognise user frustration and adapt accordingly. Paper E supports biofeedback-informed environments, where physiological data can alert individuals or teams to early signs of stress overload.

In the big picture, sustained well-being in software engineering requires a three-pronged strategy:

- Individual empowerment through evidence-based practices (e.g., breathwork, mindfulness, yoga).
- Organisational commitment to supportive structures (e.g. recognition, fairness, manageable workload).

- Technological empathy, where tools and systems are designed to minimise frustration and cognitive overload.

To conclude and answer RQ2, based on my results, approaches that address more than one level, with a particular focus on the individual, are more effective in enhancing well-being. Structured programmes with mindfulness-based practices at the individual level can be feasible to implement in SE contexts [27,145,165]. However, their effectiveness appears to be context-dependent, varies across practices, and is not consistently captured by quantitative measures. Implementation time (time of day and season of year) is a crucial factor to consider, as is the inclusion of supporting activities that enhance the effects of mindfulness-based interventions. Additionally, at the organisational level, practices and policies should be tailored to foster sustained well-being.

1.7.3 Multimodal data triangulation and LLM-assisted analysis

Papers B, E and F provide a multi-modal, mixed, rigorously transparent methodology for investigating human factors in SE. The combinations of results answer RQ3. These papers' methodologies propose that to better understand human aspects, it is imperative to cover the depth and breadth of the object of study. It is also important to collect subjective and objective data observations. Gathering qualitative and quantitative data will provide a complete overview, helping to understand subjective experiences, measurable indicators, and contextual factors.

By combining mixed methods, each paper contributed a layer of study: the emotional and experiential dimension, the physiological dimension, and the analytic reproducibility of qualitative interpretation. For example, in Paper B, I used the Willcox Emotion Wheel [166] to map engineers' emotional states, adapting a psychological tool to SE contexts, which allowed quantification of affective states without oversimplification. The adaptation of a qualitative instrument from psychology helped us to translate complex emotional states into a structured format suitable for quantification. This helped us to integrate experiences with measurable data, allowing us to write more informed recommendations.

In Paper E, I implemented a multi-modal data triangulation. The experiment combined psychometric instruments, biometric sensing (EEG, EDA, HRV), and qualitative interviews. The design aimed to overcome the limitations of self-reports, offering a more objective and continuous measure of stress. Part of the goal was to test whether the tailored survey could reliably measure stress and mental workload, and whether the results were supported by the biometric data. However, the experiment was not conclusive and surfaced important ethical and interpretive challenges. It demonstrated the need to utilise more sensitive tools for measuring stress and mental workload. In this paper, psychometric instruments alone underestimated subtle strain, since they were not able to capture it. EDA captured variations that occur below conscious awareness [167], and the interviews showed that stress was present in the participants. The results point to the need to implement a measurement that accounts for individual variability and subtle traits of the object under study. Regarding the ethical aspects, creating higher levels of stress in participants to the point that the

instruments could capture it was not permissible, as deliberately inducing harmful stress would violate research ethics. We raised this point in Paper E [161] and discuss its implications and limitations.

Finally, Paper F advanced qualitative methodology by integrating AI into the analytical process. Qualitative data forms a central component in exploring and understanding human factors. In this paper, we introduced an LLM-assisted collaborative framework designed to enhance transparency, reproducibility, and scalability in qualitative analysis. We also developed rubrics to systematically evaluate the quality of generated codes and themes. By documenting each step of the analytical process (from data preparation and coding to theme development), Paper F addressed a persistent challenge in QD analysis: the opacity of qualitative reasoning and replicability [168, 169]. Making this process publicly available, following clear guidelines rather than only presenting final results, further strengthens external validity.

We acknowledge the recent methodological critiques and concerns regarding the use of LLMs in qualitative analysis, particularly within reflexive approaches such as thematic analysis. Scholars, including Braun and Clarke, have cautioned that uncritical or automated use of such tools risks undermining reflexivity, interpretive depth, and researcher accountability, potentially reducing qualitative analysis to a mechanistic coding exercise rather than an analytic process grounded in meaning-making [170]. The hybrid framework proposed in this thesis explicitly aligns with these critiques by positioning LLMs as analytical assistants rather than analysts [171]. In Paper F, LLMs are deliberately constrained to support specific phases of thematic analysis under continuous human supervision. At the same time, interpretive authority, reflexive judgement, and theoretical sensitivity remain the responsibility of the researcher.

The study of human factors encompasses multiple variables that can compromise the robustness of empirical findings. For example, even when triangulating data, biases still exist, instruments can fail to identify nuances, and it is not possible to control for all confounding variables [172]. Recognising these boundaries is itself a methodological principle: validity in human-factors research depends as much on documenting what the data cannot show as on what it reveals.

To summarise, these studies implemented and propose a multi-level triangulation strategy: behavioural and emotional data, physiological indicators, and interpretive rigour in thematic analysis. A methodology that is a system of cross-validation, where each data type complements the interpretation of the others and is reflexively documented.

1.8 Limitations and Threats to Validity

This section presents a general scope and overview of the threats and biases encountered during our studies. In addition, each paper has a detailed section explaining its specific threats.

1.8.1 Scope of Applicability

The findings and frameworks proposed in this thesis should be interpreted in light of the sampling and contextual characteristics of the empirical studies. Across the

included papers, participation was largely voluntary, which may have led to an overrepresentation of individuals already interested in well-being, mindfulness-based practices, or reflective approaches to work. Consequently, the ISTS framework (and the other frameworks in the papers) may place greater emphasis on resources and coping strategies salient to such populations. At the same time, these frameworks might underrepresent experiences of software engineers who are sceptical of, constrained from, or disengaged from well-being initiatives.

This limitation is particularly relevant for the intervention studies (Papers C and D), where self-selection and attrition may bias outcomes toward participants who were able or motivated to sustain engagement over time. Similarly, the stress responses observed in the experiment in Paper E were collected in a controlled academic setting and with a limited organisational range. Stress manifestations and physiological sensitivity may differ in environments characterised by chronic time pressure, lower psychological safety, or different organisational cultures.

As such, the findings are examples of how subtle or non-conscious stress responses can emerge under specific task conditions. These examples may not represent stress in all SE contexts.

This thesis's contributions are likely to transfer to SE contexts that share comparable characteristics: knowledge-intensive work, sustained cognitive load, and partial organisational openness to reflective or preventive well-being approaches. In contrast, contexts marked by extreme workload, limited employee autonomy, or low tolerance for non-production activities may require additional organisational or structural interventions beyond those examined here.

1.8.2 Internal Validity

Internal validity concerns whether the observed effects in each study can be confidently attributed to the intended intervention or condition rather than to confounding factors. Across the studies in this thesis, several potential threats were identified. In the intervention studies (Papers C and D), it was difficult to isolate the effects of the breathwork and yoga practices from the influence of group interaction or the novelty of participating in a community activity. Some participants emphasised the social component as a significant contributor to their perceived well-being, which complicates causal interpretation. Furthermore, the absence of control groups (considering also Paper E) limited our ability to distinguish intervention effects from spontaneous changes or placebo-like influences. In the biometric experiment (Paper E), individual differences in physiological reactivity introduced additional internal validity concerns, as not all participants exhibited comparable sensitivity to stress induction. To mitigate these issues, the studies used pre–post comparisons, triangulation of data, and followed standardised measurement protocols to enhance the credibility of observed effects. We also acknowledge that controlling for confounding variables was not possible given our quasi-experimental settings.

1.8.3 External Validity

External validity refers to the extent to which findings can be generalised beyond the studied samples, contexts, and tasks. The thesis collected data from industrial

and academic settings, with participants ranging from professional software engineers to computing students. However, several limitations affect generalisability. Recruitment relied on personal and professional networks, academic mailing lists, and online channels, which may have resulted in an overrepresentation of participants who were already interested in mindfulness or well-being. Similarly, self-selection bias may have favoured individuals with positive attitudes toward the interventions, potentially inflating observed benefits. In the multimodal experiment, participants were purposively sampled, which may not represent the broader population of software developers. Because of the previous, the results may not translate to the "average" sceptical software engineer. To strengthen external validity, findings were interpreted with caution, and the triangulation of diverse samples, covering organisational, experimental, and educational contexts, was employed to identify consistent patterns across settings.

1.8.4 Construct Validity

Construct validity addresses whether the empirical indicators used in the studies truly capture the theoretical concepts of interest, such as stress, well-being, and resilience. Several steps were taken to safeguard construct validity. Several studies employed validated psychometric instruments recommended in occupational and psychological research. In the other cases, for example, for Papers A and B, we included the definitions of the concepts being explored. This helped participants to develop a shared understanding of key constructs such as frustration and emotional strain, thereby reducing the risk of misinterpretation. Nonetheless, measurement limitations were noted: participants occasionally reported survey fatigue, which might have reduced response accuracy, and psychometric instruments sometimes failed to capture subtle, non-conscious stress reactions. Paper E explicitly addressed this by integrating physiological measures (EEG, EDA, HRV) with self-reports and qualitative interviews, revealing discrepancies between subjective and objective indicators.

1.9 Conclusions

This thesis analysed and theorised the well-being of software engineers. It considered stress and resilience as central determinants for exploration and intervention. Through six empirical studies employing mixed, qualitative, and physiological methods, it contributes to a psychological, organisational, and methodological understanding of human factors in SE. It proposes evidence-based strategies to foster sustained well-being at individual and systemic levels.

Across the mixed-method studies (Papers A, B, and E), we identify and delimit the **factors influencing stress, well-being, and resilience of software engineers (RQ1).** The findings demonstrate that well-being is an emergent property of a sociotechnical ecosystem. Individual coping capacities, interpersonal and team relations, and the broader organisational and technological environment influence stress and recovery processes. The integrated model proposed in Figure1.4 illustrates the interactions between these environments.

Our goal is that by having a comprehensive picture of the conditions that sustain,

hinder, and enhance well-being, organisations can implement empirically informed guidelines, policies, interventions and strategies to improve engineers' well-being.

For research in this area, we aim for our framework to offer direction for future studies to explore specific areas deeper and have a point of reference to confirm, contrast, and challenge our results.

Finally, by offering a framework that integrates several spheres beyond the work environment, we aim to raise awareness that interpretations of workplace stress and resilience must account for influences that originate in the broader life domain, which can significantly alter how engineers respond to pressures at work.

This thesis results inform **approaches to foster sustained well-being (RQ2)** in the form of interventions, policies and organisational recommendations. All of these have one common goal: to support the well-being of software engineers. The intervention studies provide empirical evidence that mindfulness-based practices (breathwork, yoga, and journaling) can support stress regulation and emotional balance in software professionals. The results confirm improvements in several areas, while also indicating that the workplace context strongly mediates these outcomes. Important to note that the absence of statistically significant quantitative effects in some intervention settings (Paper D) does not indicate failure, instead points to the sensitivity of well-being outcomes to organisational context, timing, and measurement choices. Individual-level practices are most effective when accompanied by supportive organisational policies and cultures.

Multiple stakeholders can use RQ2 results to inform practice and design decisions. Researchers and companies can apply the findings to design, implement, or adapt interventions to support engineers' well-being. Specifically, we provide recommendations of DOs and DON'Ts for tailoring mindfulness-based interventions to SE contexts, as well as organisational guidelines and policy-level recommendations. Additionally, designers and developers can utilise our findings to inform the design of chatbots with smoother interaction patterns, thereby reducing the triggers of frustration.

Regarding the **methodological strategies for studying human factors (RQ3)**, this research integrates triangulated, multi-modal evidence data (combining psychometrics, physiological data, and qualitative insights) to overcome the limitations of single methods. We made visible that this triangulation, although essential, comes with challenges that must be explicitly acknowledged and analytically addressed. Part of our goal was to report and analyse misalignments across data sources explicitly and to give visibility to null results. Future studies can consider the lessons learned from our experiences and implement experiments that avoid our struggles. Similarly, our results seek to raise awareness among researchers and practitioners of the importance of considering acute and long-term stress when planning studies or interventions.

Moreover, we also proposed an LLM-supported framework for qualitative analysis and created guidelines to improve the transparency, reproducibility, and scalability of the integration. Qualitative researchers can implement the framework and use the rubrics as metrics. The rubrics can even be used alone to guide manual coding and theme development, particularly among new users of thematic analysis.

RQ3 results are particularly useful for making visible the ethical implications of stress (and health-related topics) research, especially when combining intrusive measurements in non-heavy health-related fields, such as SE.

To conclude, with RQ3, we aimed to provide researchers with concrete method-

ological tools and considerations for studying human factors in SE more rigorously and responsibly.

Future Work

Building on the findings of this research, future work will test and refine the proposed ISTS integrated framework in different contexts. For example, although we attempted to include engineers from several countries, we acknowledge that there is still a need to explore other regions. Future studies will therefore extend data collection to underrepresented regions to examine the robustness and transferability of the framework across different sociotechnical environments and organisational cultures. The application of the framework to design and evaluate assessments and interventions for the organisational environment is also a next step in our studies. Coming studies can also explore the addition of new concepts at the individual level, such as emotional intelligence, need for cognition, and empathy. The social and organisational environment, as well as the technological ecosystem, can be further amplified in future studies by adding new layers and elements that adapt to specific contexts and companies.

Another direction is the investigation of role-specific experiences within software engineering. The studies included in this thesis intentionally adopted a broad view of the SE population. Future work could examine how the framework manifests differently for developers, testers, technical leads, product owners, and engineering managers. These roles are embedded in distinct sociotechnical positions, with different cognitive demands, accountability structures, and exposure to interruptions. Understanding these differences would enable more targeted interventions and refine the framework's sensitivity to organisational role structure.

About the interventions, we aimed to implement more follow-up studies, with a primary focus on exploring diverse mindfulness-based interventions. We also want to recruit larger cohorts and include control groups. For the implemented interventions, we will use follow-up measurements to assess their sustained effects on stress-related outcomes. In addition, future work can explore adaptive interventions that respond to contextual signals, such as workload peaks or project deadlines. Future work could also investigate individual differences as moderators of intervention effectiveness. For example, to study how factors such as prior experience with mindfulness practices, baseline stress regulation capacity, cognitive style, or attitudes toward technology influence outcomes.

Regarding the immediate work pipeline, for our methodology work, we plan to experiment with adapting content analysis using AI and eventually testing more qualitative data analysis methods. We also plan to test our hybrid framework with new interviews, in a way that we analyse them following our guidelines without comparing them to human outcomes. This will give us a good picture of how the framework works when we do not use a human benchmark. Another direction is to study the epistemic risks and boundary conditions of AI assistance in interpretive research.

For data triangulation, we aim to include EEG data in SE tasks and collect it in real-world software development scenarios. Further studies can align biometric data with self-report instruments and recruit larger cohorts. One of the goals of this

comparison was to assess how accurately questionnaires (particularly psychometrics) capture the variability in the software engineering population. We argued in this thesis that theories need to be adapted; there is also the possibility that psychometric instruments need to be adapted to the SE population. Currently, we are exploring the collection of blinking data to test its effectiveness as a proxy measure for identifying stress in SE human tasks.

Chapter 2

Paper A:

The Factors Influencing Well-Being in Software Engineers: A Mixed-Method Study

C. Martinez, B. Penzenstadler, R. Feldt

In Transactions of Software Engineering and Methodology (TOSEM), 2025.

Abstract

The well-being of software engineers is increasingly under strain due to the high-stress nature of their roles, which involve complex problem-solving, tight deadlines, and the pressures of rapidly evolving technologies.

Despite increasing recognition of mental health challenges in software engineering, few studies focus on the factors that sustain or undermine well-being. Existing research often overlooks the interaction between personal, collaborative, and organisational influences on this unique population. This study fills this gap by investigating the specific factors affecting the well-being of software engineers. We conducted 15 qualitative interviews and complemented them with a survey with participants from multiple countries to validate and extend our findings to a broader population. Our mixed-methods approach provides a robust framework to identify key factors influencing well-being, including personal perceptions of well-being, interpersonal and collaborative dynamics, workplace support and recognition, organisational culture, and specific stressors inherent to software engineering.

By offering a detailed, context-specific exploration of these factors, our study builds on existing literature and provides actionable insights for improving well-being in software engineering. We conclude with policy recommendations to inform organisational strategies and develop targeted interventions that address the specific challenges of this field, contributing to more sustainable and supportive work environments.

2.1 Introduction

Software development is fundamentally a human activity that relies on engineers' skills, creativity, and well-being. Developers' mental and emotional states significantly impact their productivity and the quality of their work [5]. Good levels of well-being enhance cognitive function and job satisfaction, fostering engagement and innovation [9]. In contrast, stress and burnout can lead to decreased performance, more errors, and reduced creativity, ultimately affecting the success of software projects [19]. Several studies have explored how various factors such as personality traits [97], feelings [8], sentiments and emotions [22] influence software development. Additionally, research has examined the relationship between job satisfaction and perceived productivity [25] and the effects of stress on software engineers [19, 107]. Specific contexts, such as the COVID-19 pandemic, have also been studied to predict well-being and productivity fluctuations under global stressors [173].

Despite these contributions, well-being within software engineering (SE) remains only partially understood, with significant gaps in how individual, team, and organisational factors shape it. Well-being can be seen as a dynamic process that allows people to evaluate how their lives progress based on the interaction of their circumstances, activities, and mental resources, often called 'mental capital' [9]. To accurately assess well-being, it is essential to consider both objective factors and personal perceptions. Wong et al.'s [87] study on mental health (an interrelated contributor to well-being) addresses this gap; however, it focuses on a single country and concentrates primarily on the individual level. Our study extends this perspective by incorporating input from software engineers across several countries and examining how intersecting factors influence well-being at individual, team, and organisational levels.

By having a comprehensive view of the factors influencing software engineers' well-being, we aim to raise awareness about mental health issues in SE and contribute to the literature focusing on the software field. At the same time, we aim to add to the global discussions on improving the workplace.

With this study, we aim to answer the question:

What factors influence the well-being of software engineers? We wish to understand what in their environment, on a personal as well as a team level, contributes or takes away from software engineers' well-being. We break this overarching question down into two research questions:

RQ1: What are the **contributors** to software engineers' well-being?

RQ2: What are the **hindrances** to software engineers' well-being?

To achieve our goal, we conducted 15 interviews and later compared the insights with a survey of 76 participants from multiple countries. We compare our results to work on well-being factors from other fields.

The paper has the following structure: Section 7.2 presents the background and related work. Section 7.3 explains our mixed-method research, including participant recruitment, data collection, and analysis procedures. Section 7.4 presents the qualitative and quantitative results, and Section 6.5 discusses the findings, limitations, and implications for practice. Finally, Section 2.6 introduces future research and concludes our study.

2.2 Background and Related Work

This section presents the background and related work for the study at hand. We first present central concepts around well-being and then give an overview of the most relevant related work.

2.2.1 Background

2.2.1.1 The Conception of Well-being

> We must practise the things which produce happiness, since if that is present, we have everything, and if it is absent, we do everything in order to have it. — Epicurus

According to Magyar and Keyes [174], the two most common lines of well-being research distinguish between well-being as the presence of something positive and the absence of something negative. These approaches define well-being either in terms of positive feelings or positive functioning. The first line, hedonic well-being, focuses on the degree of positive feelings (e.g., happiness) and overall life satisfaction and is commonly referred to as emotional well-being [175]. The second line, eudaimonic well-being, emphasises positive functioning, experienced when individuals realise their human potential and is typically described in terms of psychological well-being [174].

Complementing these perspectives, McNaught [176] proposes a definitional framework of well-being, in which well-being is understood as an objective and subjective assessment of a desirable human state. This framework identifies society, community, family, and the individual as its central pillars.

For clarity, we present the main terms used in this article, *well-being* and *mental health* in Table 6.1. These definitions also clarify that the concepts are closely related, but mental health has a more specific scope limited to perceptions of the mind. Although our work mentions mental health, the construct we are investigating is well-being.

Table 2.1: Definition of main concepts in this paper

Concept	**Definition**
Well-being	A desirable human state involving subjective and objective assessments. It includes positive feelings and life satisfaction (hedonic dimension) and positive functioning and self-realisation (eudaimonic dimension). It is grounded in individual, family, community, and societal contexts [174–176].
Mental Health	"A state of mind characterised by emotional well-being, good behavioural adjustment, relative freedom from anxiety and disabling symptoms, and a capacity to establish constructive relationships and cope with the ordinary demands and stresses of life " [177].

Further, to frame the relevance of **intellectual stimulation** for well-being, Anjali and Anand [178] find that intellectually stimulating work increases job contentment and employee commitment in IT.

Finally, to point out the relevance of **creativity**, Sokol and Figurska [179] confirm creativity as one of the core competencies of knowledge workers, and that it requires space (mentally and on the schedule) to come to fruition.

2.2.2 Related Work

2.2.2.1 General Population

Several works investigate well-being through quantitative assessments and specific factors that contribute to or limit perception.

Diener and Seligman presented the most established frameworks: Diener [175] looks at individual or subjective well-being and was the first to establish psychometric instruments for measuring the construct, for example, the subjective well-being (SWB) scale. Seligman [77], one of the central figures of positive psychology, conceptualises well-being around the five pillars of positive emotion, engagement, relationships, meaning and accomplishment. We explored some of these pillars in this study.

According to Gallup researchers Rath and Harter [76], well-being consists of five essential elements: career well-being (how you occupy your time or liking what you do every day), social well-being (having strong relationships and love in your life), financial well-being (effectively managing your economic life), physical well-being (having good health and enough energy to get things done daily), and community well-being (the sense of engagement you have with the area where you live). While the framework includes career well-being, its focus remains individual, mainly, with limited attention to the organisational and contextual factors that shape workplace well-being. We expanded this category by including team and organisational contexts.

Michaelson et al. [9] offer a more internally focused perspective; they propose a framework for personal well-being comprising five components: emotional well-being, satisfaction with life, vitality, resilience and self-esteem, and positive functioning. Complementing this individual focus, Nielsen [180] shows that, in employee settings, well-being in self-managing teams depends strongly on supportive management. These views motivate our study's multi-level lens in software engineering, examining how individual resources, team dynamics, and organisational practices jointly shape software engineers' well-being.

Beyond frameworks that emphasise personal or organisational factors, some studies examine well-being's **limitations** and **social utility**. For example, Leifels and Zhang [181] investigated cultural factors and found that a significant predictor of well-being impairments was a lack of trust and accountability in only mono- and bicultural teams, not in multicultural teams. Misunderstanding and disagreement were positively associated with well-being impairments only in multicultural work teams. At the societal level, Michaelson et al. [9] argue that national governments should directly measure people's subjective well-being (as in their experiences, feelings and perceptions of how their lives are going) to guide social development. They call for these measures to be collected on a regular, systematic basis and published as National Accounts of Well-being, and argue that the measures are needed because the economic indicators which governments currently rely on tell us little about the relative success or failure of countries in supporting a good life for their citizens. In a similar vein, but more oriented towards companies, Harter et al. [182] propose measuring the social utility of subjective well-being in business profitability, productivity, and employee retention. Together, these perspectives show how societal and organisational factors can influence well-being. In our study, we integrate these dimensions and research how they interact in SE.

While these frameworks have significantly advanced our understanding of well-

being in the general population, they are rarely examined in the context of specific professions or work settings. As such, it remains unclear how well these theories capture the lived experiences of occupational groups with distinct demands, such as software engineers. In the following section, we explore previous work investigating well-being within the software engineering domain and position our study within this emerging body of research.

2.2.2.2 Software Engineering Population

Several important contributions from the last few years show the various impacts on well-being within the software engineering domain. For example, regarding **emotional states and negative experiences** in software development, frustration [79], burnout [24], and (un)happiness [8] have been analysed at different levels. Madampe et al. [80] investigate reasons for negative emotions in agile contexts and propose several solutions to overcome the causes. One such negative emotion, the experience of feeling overwhelmed, was explored by Michels et al. [78] in a qualitative psychology study that identifies seven distinct categories: communication-induced, disturbance-related, organisational, variety, technical, temporal, and positive overwhelm. Similarly, Santana et al. [86] identified everyday interpersonal challenges that point to a lack of psychological safety in software development practices, challenges such as reluctance to admit mistakes, avoiding seeking help, and fear of sharing negative feedback. While these studies identify symptoms and contexts of emotions, our work traces some of these emotions to systemic drivers (e.g., how organisational culture or lack of recognition influences well-being).

On the **intervention** side, Bernardez et al. investigated mindfulness interventions for conceptual modeling [92], and Penzenstadler et al. conducted studies on the impact of breathwork interventions [19]. Similarly, Montes et al. [27] compared different mindfulness practices to improve stress management. Our study supports these interventions, advocates for acknowledging systemic factors, and proposes policies to develop and implement tailored interventions in software engineering companies.

Concerning **productivity and well-being** relationships, Leme et al. [183] developed an approach based on the GQM (Goal, Question, Metric) methodology to collect, measure, and monitor mental health and productivity metrics. They found a positive correlation between the two. Furthermore, Hicks et al. [82] presented a research-based framework for measuring successful environments on software teams for long-term and sustainable socio-cognitive problem-solving that was tested across 1282 full-time developers in 12+ industries, predictive of developers' self-reported productivity. One more framework by Sghaier et al. [184] was designed to assess AI-driven software engineering tasks to customise the tools to improve developers' efficiency, well-being, and psychological functioning. Finally, Storey et al. [83] developed a theory with a bi-directional relationship between software developer job satisfaction and perceived productivity that identifies what additional social and technical factors, challenges, and work context variables influence this relationship. These studies established correlations between well-being and productivity or job satisfaction. Our study complements these findings by examining how specific stressors (e.g., tight deadlines) erode well-being and output, and proposes how organisational policies can break this cycle.

Studies also look at specific work conditions, such as **remote and hybrid work settings**. Russo et al. [173] and Ralph et al. [185] examined the effects of remote work during the pandemic on well-being and job satisfaction, a very particular circumstance. Correspondingly, De Souza Santos et al. [186] investigated how hybrid work influences the well-being in the software industry. Their findings indicate that hybrid work positively affects overall well-being and has challenges like infrastructure issues and reduced interaction with co-workers. In contrast, our work is agnostic of particular work context circumstances. It focuses on the overall factors for software engineers' well-being, specifically the contributors (RQ1) and the hindrances (RQ2). We use qualitative data from interviews to establish well-being factors and quantitative data from a confirmatory survey to triangulate.

Inclusivity, empathy, and supportive work environments have been studied by Dwomoh and Barcomb [84], who explored three ways in which organisations and individuals interested in improving representation can make tech careers more inclusive: by (1) supporting networking, (2) cultivating inclusive leadership, and (3) promoting the development of self-efficacy. Aligned with the previous, Cerqueira et al. [108] recommend that team members practice empathy by being mindful, being open, understanding others, and taking care, which can reduce blame, improve job motivation, prevent burnout, and create a better work environment. Moreover, Singh et al. [187] worked with women software engineers and provided a prototype that employs an emotion detection approach to generate Mental Health Scores called SOFTMENT (SOFTware sector MENTal well-being support system). Our study investigates how the companies' initiatives in DEI promotion influence well-being and uses the results to advocate for supportive and inclusive work environment policies.

Most closely related to this article, Wong et al. [87] analysed 14 interviews with software developers to explore how mental well-being should be addressed at individual, team, and organisational levels, highlighting the need to integrate mental well-being into the technologies employees use at work. The authors focus mostly on personal experiences with mental well-being in the workplace and their approaches to managing it in the US context. Our study integrates a Europe-centric perspective from the interviews with a global outlook from the survey, enabling us to uncover broader patterns related to mental well-being and workplace dynamics.

2.3 Methodology

2.3.1 Study design

This study adopted a mixed-methods approach utilising surveys and interviews to explore the factors influencing software developers' well-being comprehensively. Interviews were conducted with software engineers working in Sweden, examining the cultural, social, and contextual factors shaping well-being within that specific context. Subsequently, surveys were distributed to software developers across several other countries.

Combining interviews and surveys allows for a nuanced understanding of the diverse factors contributing to well-being. Surveys provide quantitative data to analyse trends, while interviews offer qualitative insights into developers' unique experiences and challenges within a particular cultural milieu.

Inspiration was taken from the Bioecological Model (BM) by Bronfenbrenner [156] to design the data collection instruments and to later analyse the data. As an ecological approach, the BM embraces holistic views, recognising that biological, psychological, sociocultural, and physical environmental factors collectively influence well-being [188]. This approach values physical and social environments in health creation: physical aspects encompass architecture, geography, and technology within a context, while the social environment includes the cultural, economic, and political dynamics at play [189]. Hence, the questions in the interview and the survey explored the different systems that the subjects interact with aiming to make connections between personal situations and explain how these intersect with those other systems (team, company and culture).

2.3.2 Population

Our target population was software engineers currently working in IT. We specifically looked for engineers living and working in Sweden for the interviews. However, we aimed to have software engineers answer the survey from anywhere in the world. We wanted to compare and contrast our results from Sweden with those from other countries, but a systematic comparison was not possible since the samples from other countries were much smaller than those from Sweden. We do point out noted specifics of Swedish versus other answers in the discussion.

2.3.3 Data Collection

We collected data from interviews and surveys. The following subsections elaborate on each instrument and its corresponding pilots and adjustments.

2.3.3.1 Pilots of the data collection instruments

We used an interview guide and a survey to collect our data. Both instruments were piloted before being applied to our target population. The interview guide was tested twice to ensure the questions were clear and to measure the estimated time. The first author corrected the guide based on the interviewee's feedback. The survey was piloted at the Eclipse Developer Conference, which took place in Ludwigsburg, Germany, in October 2023, and we received 20 answers. Participants gave feedback on the questions, and corresponding changes were made.

2.3.3.2 Interviews

Qualitative data was collected through 15 individual semi-structured interviews using an interview guide with open questions to gather in-depth information [190]. The interview guide was designed at three levels, plus the demographic data. The first questions gathered information about the background and experience of the interviews, and the following sections explored the factors that influence well-being at the individual, team and organisational levels. As mentioned in Section 2.3.1,

the interview aimed to holistically explore the participants' context, considering the systems the interviewee interacts with. See online appendix for interview guide [191].

To recruit interviewees, we used social media posts such as LinkedIn, X, and Facebook groups, direct emails to software companies, and the personal networks of the three authors and the university contacts. We targeted software engineers living and working in Sweden.

The interviews lasted between 40 and 75 minutes. We allowed the participants to join online or in person, so we had thirteen interviews in person and two online. The interviews were performed by the first author with the aim of consistency. The first contact with the participants was to explain the interview's goal and share the informed consent. The informed consent explained the goal of the interview, the voluntary and anonymous participation, and the interviewees' right to withdraw their participation at any time. During the interview, the first step was establishing rapport and presenting and signing the informed consent. All interviews were audio-recorded with the consent of the interviewees (see informed consent in the online appendix [191]) and later transcribed and denaturalised (this removes involuntary vocalisation), focusing only on the content of the interview [192].

2.3.3.3 Survey

We designed the survey in a similar way to the interview. The first page of the online survey showed the informed consent with an explanation of how the data would be handled and let the participants know that participation was voluntary and anonymous. We provided our contact information in case participants had questions or wanted to contact us.

The survey had 33 questions in total. The first questions collected demographic information, while the following sections explored how (i) the perception of well-being, (ii) the influence of equality, equity, diversity and inclusion, and (iii) the relationship with managers and peers, companies' culture and physical environment influence software engineers' well-being. The survey had open, multiple-choice, and Likert scale questions. We tailored the scales based on the questions and answer options to capture the participants' perceptions and opinions better. For example, the overall well-being scale differs from the scale measuring how heard and respected the participants feel. Therefore, each item was treated as an independent measure of a particular aspect of well-being.

Relevant questions were identified from existing research (e.g. [87], [175]) and the preliminary results of the interviews.

The survey was available in three languages, namely English, Spanish and Portuguese. The survey was posted on LinkedIn, X, and Facebook. We contacted several software companies to ask for support in sharing the survey. Similarly, personalised emails were sent to software engineers inviting them to answer. We targeted software engineers from anywhere in the world.

2.3.4 Data Analysis

2.3.4.1 Interview Analysis

The interviews were transcribed and checked against the original recordings for accuracy. We analysed the transcripts using reflexive thematic analysis following Braun and Clarke's six steps [193]. After reading the transcript several times to familiarise ourselves with the data, the first and second authors coded three interviews (20% of the total data) to assess coding reliability. We compared our results, and we aligned labels, definitions, and examples for each code. Later, we coded the rest of the transcripts. Then, we continued with the rest of the steps: combining codes into themes, reviewing and refining themes, and reporting findings.

2.3.4.2 Survey Analysis

The first step was to clean and organise the data. Every survey answer was read to ensure the respondents were among our target group. Answers from people not working in the software field were deleted. Next, the answers in Spanish and Portuguese were translated into English to create a single database for analysis. We used graphs to visualise the answers based on the type of question. The demographic data was analysed and summarised to understand participants' age, gender, area and years of expertise, and geographical distribution.

The Likert scale questions were analysed using descriptive statistics and visually represented with diverging stacked bar charts. The open questions were analysed using content analysis. We ran Spearman's correlation, Mann-Whitney U and Kruskal-Wallis tests to explore the relationships between specific variables and assess group differences in our data.

2.3.4.3 Reflexivity

Here, we outline the backgrounds and perspectives of each study author, examining how our unique experiences might have influenced both the research process and its outcomes. This reflexive approach [73] is critical in qualitative research, helping to identify and mitigate biases that might shape the interpretation of findings.

The first author, with a bachelor's degree in psychology and a master's in social work, offers a strong foundation in human behaviour and social dynamics, supporting exploring well-being factors like stress, coping strategies, and interpersonal relationships within software engineering environments. In contrast, the second author holds a PhD in Software Engineering, paired with training as a yoga instructor and embodied mindfulness coach. This brings a unique balance of technical and mindfulness insights to the study. Their background informs an understanding of work-related challenges, such as workload, deadlines, and technology's role in daily tasks. The third author, with dual expertise in psychology and software engineering and over two decades of consulting experience, provides an integrative perspective on organisational processes and team dynamics, bridging our research's human and technical aspects. Together, we share a view that human factors in software engineering are often undervalued and deserve greater attention for creating healthier, more effective organisations.

This blend of interdisciplinary perspectives has shaped our approach. The first author's insights into psychological and social dynamics, grounded in practical community work, enriched the analysis. The second author's expertise in technical and therapeutic fields contributed to a holistic perspective, integrating rigorous software engineering with mindfulness. Meanwhile, the third author offered a broad organisational view, emphasising the impact of culture, context, and individual attributes on well-being.

Throughout the study, the first and second authors led the qualitative analysis reflexively, regularly evaluating assumptions and biases through open dialogue. The third author acted as an external reviewer, critically examining methodological choices and interpretations. This approach aimed to enhance the study's credibility and to represent participants' experiences with integrity. Nevertheless, despite efforts to remain objective, our shared belief in the importance of human factors in software engineering may have influenced our interpretations.

2.3.5 Ethical Considerations

This research followed the recommendations of the ethical research study guidelines of Chalmers University. Further, this study was approved by the Swedish Etikprövningsmyndigheten[1]. Informed consent was obtained from all participants.

Participants were thoroughly briefed on the study's objectives, methods and potential risks. They were also informed of their right to withdraw from the study at any time without facing any consequences.

All personal identifying information was kept strictly confidential to protect participants' privacy. Each interview participant was assigned a unique code as an identifier, and all collected data, including transcripts and audio recordings, was anonymised and securely stored. Access to the information was restricted to authorised researchers only.

2.4 Results

This section presents the results from the interviews and the survey.

2.4.1 Interviews

We conducted 15 interviews with software engineers (SErs). Table 2.2 presents the respondents' positions and years of experience.

From the thematic analysis, five themes emerged. See Figure 2.1 for an overview of themes and sub-themes. In the following sections, every theme is explained with its corresponding sub-themes.

2.4.1.1 Theme 1: Individual Conception of Well-being

This theme explains how software engineers conceive their well-being, with most seeing it as multifaceted.

[1] https://etikprovningsmyndigheten.se

Table 2.2: Demographic and professional characteristics of interview participants (N=15), including job positions, years of experience, and industry domains.

ID	Position	Years of Experience	Domain
P1	Systems Engineer	7	Government
P2	Product Test And Integration Engineer	2	Government
P3	Software Developer	7	Government
P4	Software Developer	5	Government
P5	Software Developer	23	Government
P6	Configurations And Test Methods	20+	Government
P7	Embedded Software Engineering	10	Automotive
P8	Software Developer	5	Transport
P9	Software Developer	12	Automotive
P10	Software Developer	12	Automotive
P11	Computer Vision Specialist	7	Computer Vision
P12	Scrum Master And Developer	3.5	Android apps
P13	Requirements Engineer / Research Project Leader	6	Automotive
P14	Software Application Developer	6	Airlines
P15	Back-End Developer	15	Fintech

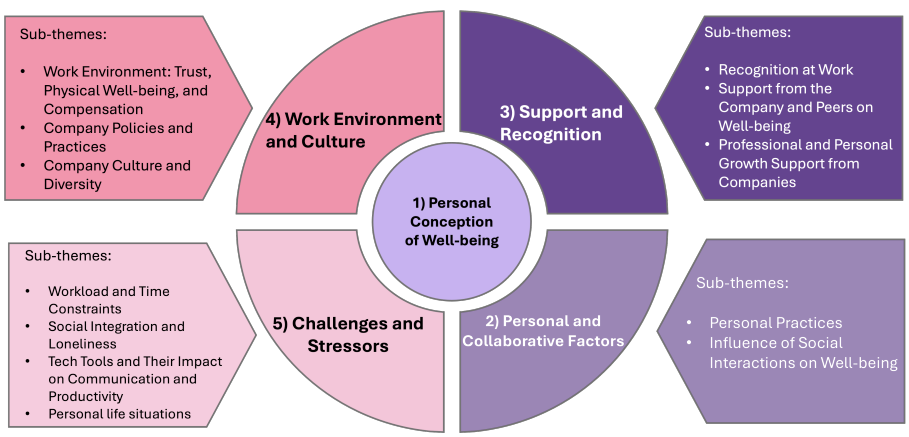


Figure 2.1: Framework derived from qualitative analysis of interviews, categorising factors affecting software engineers' well-being into five primary themes with associated sub-themes.

Well-being for software engineers, according to the interviews, comprises several aspects. It involves feeling happy, content, motivated to perform daily activities, and supported by a healthy work environment. It includes balancing personal and professional life without interference, ensuring mental and physical health, and having safety and access to fundamental human rights. Well-being also entails mental and emotional aspects such as the absence of stress and anxiety, sleeping well, not feeling overly tired, and lack of suffering.

Additionally, it encompasses meaning and accomplishment, including having meaningful tasks, feeling accomplished, and being able to help others. It is about having peace of mind and not being stressed about work deadlines. It comprises physical wellness, which involves feeling physically well, being active, and not getting out of breath easily. Finally, social aspects are crucial, including having supportive relationships and a positive work culture.

In conclusion, software engineers conceive their well-being as a multidimensional concept encompassing emotional, physical, and social aspects. This holistic approach to well-being is reflected in the coming themes.

2.4.1.2 Theme 2: Personal and Collaborative Factors

Starting from an individual point of view, this theme elaborates on the various well-being practices and how SErs integrate them into their routines. Physical activity, from gym sessions to yoga, is a prevalent practice.

Beyond individual practices, and considering the immediate context, social connections significantly influence well-being. Open communication, trust, and mutual respect create positive interactions that foster emotional well-being and reduce stress. Conversely, a lack of support or negative interactions can have detrimental effects. Overall, SErs' well-being is shaped by personal efforts and collaborative factors.

Sub-theme 1: Personal Practices. Several key activities and their regularity were identified. Physical exercise, including gym attendance and sports, is frequently mentioned, with some participants going to the gym three to five times weekly. Yoga and breath work are cited as regular practices.

The quote below shows the emphasis of one participant on physical activity, although they do not perceive any intentional or specific actions aimed at directly addressing their mental well-being.

> *"For physical health, I go to the gym but I don't think I'd do anything special for mental well-being."* — P7

From a different perspective, the participant mentioned below how, for them, physical and mental well-being are connected and taken care of at the same time. They view physical activity as a foundational aspect of their overall well-being.

> *"Number one, foremost and having the opportunity to move or maybe I'll keep repeating this over and over again. But taking care of my physical well-being is like one of the best ways I know of taking care of my mental well-being."* — P10

The high regularity with which participants engage in physical exercise, three to five times a week, shows the significance of this practice in their well-being routines.

Social interaction is also considered a well-being practice, and many respondents regularly go to the office to socialise with colleagues, live with partners, and frequently meet friends to foster well-being. As mentioned by one participant:

> *"I like to come into the office quite often. I can work at home some days, but mostly I want to be at work because I gain something from the social interaction with colleagues."* — p 4

Social engagement is essential for enhancing mood and overall well-being.

Participants also mentioned several activities they practice regularly, such as taking walks, yoga, breathwork (for example, box breathing), hobbies, meditation, and positive affirmations, showing a holistic approach to well-being.

One participant mentioned when asked what they do as a well-being practice:

> *"Not really. Nothing specific at least for that purpose, other than general you know, hobbies and everything, nothing specific for well-being"* — P3

While they may not engage in specific practices targeted at well-being, by incorporating activities that bring joy and fulfilment into their lives, individuals enhance their psychological resilience and cultivate a sense of work-life balance.

Additional activities include going to the mall, sleeping well, meditation and acro-yoga, singing in a choir, cognitive behavioural therapy (CBT), participating in marathons, and seeing a psychologist. These have been a regular practice for other participants for a few years.

> *"Being active is like, I feel like I get more dopamine when I am more active, including going to the mall, maybe going for a walk. And also, as I said, hanging out with friends and just going out instead of staying indoors."* — P7

These findings show how participants integrate physical, mental, and social activities with varying regularity. It also gives an idea of the different angles of well-being. Participants tailored their activities to their individual preferences and needs.

Sub-theme 2: Influence of Social Interactions on Well-being Participants reflected on how their well-being influences relationships and interactions with others. Participants see social connections, both at work and in personal life, as crucial for well-being. The interactions are characterised by open communication, trust, and mutual support, which provide emotional support and a sense of belonging. Conversely, challenges such as communication barriers or conflicts can create stress and negatively affect individuals' mental health.

One participant mentioned how impactful it is for them to have friendly colleagues, highlighting at the same time the role of positive social interactions in the workplace, where friendly relationships contribute to an individual's emotional well-being and overall satisfaction at work.

> *"I love having friends. Like hanging out with people that I like... So it is important for me, having friendly colleagues, that I can talk to them freely."* — P7

Participants mentioned activities that promote emotional sharing and help resolve conflicts, enhancing team cohesion and reducing stress. For instance, open communication seemed to flow better during team events and after work. Furthermore, personal

relationships outside work also play a crucial role in our participants' well-being. One participant mentioned the importance of the people around them:

> *"So I would say the people around me really matters for me. So, if they're bringing negative vibes, it really affects me. The people is the main factor that makes me feel mentally well. So, if I feel alone or if I feel you know, left out, I definitely feel down and I'm sad."* — P14

This quote shows how positive and negative personal relationships can foster positive and negative feelings and influence one's overall sense of well-being, which, in turn, influences work performance and satisfaction.

In conclusion, this sub-theme showed that the well-being of software engineers is significantly influenced by their social interactions and personal relationships characterised by open communication and mutual respect. Conversely, negative interactions and a lack of support can increase stress and decrease general well-being.

Theme 2 takeaway: Software engineers achieve well-being through personal practices and social connections. Regular physical activities are essential to physical and mental health. Additionally, positive social interactions enhance emotional well-being, while negative or unsupportive relationships increase stress.

2.4.1.3 Theme 3: Support and Recognition

Participants mentioned two aspects of the work environment that are important for their well-being: support and recognition. This theme elaborates on them and explains how respondents perceive their company to provide support through team collaboration, managerial assistance, resource access, and whether recognition is present or not.

Sub-theme 1: Recognition at Work Participants mentioned recognition at work as a factor influencing their well-being and job satisfaction. They elaborated on what recognition at work entails for them and its significance. They also stressed the need for positive feedback and the sense of being part of a team. Recognition, for our participants, involves acknowledging hard work and achievements, providing feedback, and ensuring employees feel integrated. Feeling valued and acknowledged for contributions can significantly enhance motivation and engagement.

Conversely, the absence of recognition can lead to dissatisfaction and even the consideration of leaving the job. Interviewees shared their varying experiences and perceptions regarding recognition at their workplace. One of them mentioned:

> *"I think I've earned my way into people, at least into my company, and into my peers, and I feel everyone respects me and listens to me when I have something to say."* — P1

This interviewee felt valued for their contributions and believes they have earned the respect of their colleagues and peers. The quote shows that recognition is not only about formal acknowledgements but also about everyday interactions where one's input is valued.

In contrast, other participants mentioned the absence of recognition and how it made them feel. The participant below mentioned that their lack of acknowledgement for their efforts prevents them from being fully happy at work.

> *"Well, I'm missing the recognition. That would make me fully happy."* — P13

This quote emphasises recognition's significance in an employee's emotional well-being, job satisfaction and career decisions. Further, it shows that without it, even other positive aspects of the job may not suffice to ensure complete job satisfaction. The interviewee mentioned they were considering changing their job since they still needed recognition. Recognition at work is something to consider when planning actions to influence retention and engagement.

Sub-theme 2: Support from the Company and Peers on Well-being Support from the company can manifest through various initiatives aimed at promoting mental and physical health and fostering a positive and inclusive work environment. It is common for companies in Sweden to provide allowances catering to mental and physical health. Employees can choose activities that help them manage stress and maintain a healthy work-life balance, as the quote below shows:

> *"If you want to, they have these programs you can participate in; different activities. So if you're interested in a sport, you can participate in clubs. But I mean, it's nothing that you know, unless you look for it or went in the portal search for it. But it's there."* — P2

Companies support various activities; however, the employees are the ones who take proactive steps to maintain their overall health. Nevertheless, the effectiveness and perception of this support vary among employees, as the quote below shows.

> *"My first thing that I want to say is that there isn't much support from them. Apart from what is in the collective agreement that they need to provide this free sports (wellness allowance) and things like that, which I think is just the bare minimum. They do the bare minimum."* — P10

This contrasting quote provides a critical perspective, indicating that not all employees feel adequately supported by their company. The respondent perceives the company's efforts are limited to the minimum requirements stipulated by collective agreements without going beyond to offer something meaningful.

While some employees feel that support is minimal and meets basic requirements, others appreciate different forms of support their companies provide, such as creating a positive and engaging work environment. Participants commented that their companies focus on teambuilding activities and cultural events to strengthen interpersonal relationships and foster a sense of community among employees.

> *"Yeah, so company's trying to be in the best workplaces in the industry in the city. So they're promoting let us... teambuilding and, you know, a lot of cultural balance events every month, so they are trying to have a positive work environment for everyone."* — P14

This quote illustrates the company's actions to create an engaging and supportive workplace culture. Initiatives such as regular events aimed at cultural balance can foster inclusive work environments and promote that employees feel valued and included.

In addition to company-led initiatives, peer support adds to the collaborative and positive work environment by creating and fostering an environment where team members can rely on each other for assistance, feedback, and camaraderie. Interviewees elaborate on the impact of their peer network.

> *"Oh, I don't have anything negative to say because our team is really friendly and we can talk to each other without any hesitation. They all are reachable, even though people are not working in the same office."* — P7

This participant commented on the importance of open communication and accessibility among team members. The respondent stresses the friendliness and approachability of their peers, which creates an environment where individuals feel comfortable sharing their thoughts and seeking help. It is notable, too, that the quote mentions that the approachability applies to even team members who work in a different location, so peers feel supported regardless of their location.

More positive attributes of the teams were mentioned, including friendliness, supportiveness, and reasonableness. Interviewees, in general, commented on how impactful peer support is on the individual's overall well-being and job satisfaction.

Sub-theme 3: Professional and Personal Growth Support from Companies
This sub-theme focuses on opportunities and support provided by the company for employees to develop professionally and personally. Participants' view of their companies' support showed a complex picture. On one hand, interviewees expressed a potential disconnection between their desires for more growth opportunities and the current company offerings.

> *"I don't feel my company supports so much the personal development and the professional development. But I would like it to. I would like to be part of a company that talks more about personal development and professional development. Right now I don't feel it."* — P12

On the other hand, some participants perceive support from their companies via efforts to provide opportunities for learning and development through platforms and goal-setting. The quote below is an example of that perception.

> *"The company invest on us, like for our day to day learnings. They have different platforms to learn and there is a platform we can go and learn from there and do the examination and improve our qualifications. Also they have this yearly milestone plannings for the each employee so that they review them by every six months."* — P14

This interviewee sees the company's provision of learning platforms as a significant factor in professional growth. They also value the company's investment in resources that enable them to continuously learn and stay updated in their field.

In the cases where the companies were not supportive, participants commented on some managers' significant role in taking the initiative in employee growth despite

the lack of a structured system. One participant mentioned their manager actively supports personal and professional development through regular meetings.

> *"My manager is actually a really busy person when I look at his calendar, it's always full. But still he finds his time to talk to each each of us. Like we have, like official one on one meetings every two weeks. Other than that, still he talks to us even though he is not involved in what we're doing. He tries to talk to us and see if we face any issues and like not micromanagement, but he is so supportive."* — P7

This quote presents the potential impact of good leadership and a personalised approach to growth. Further, some aspects of the work environment might indirectly contribute to growth, even if not explicitly designed for it. For example, a supportive manager with open communication and a focus on work-life balance can create a positive environment for learning and development. Similarly, opportunities for interaction and support within the team can foster knowledge sharing and a sense of community, which can contribute to personal and professional growth. One participant shared about their colleagues:

> *"They're always supportive people. They are always helpful. When you ask someone for help you get your help. I always get help from people."* — P15

While some companies might not have a robust growth support system, there are lines of support from some managers and colleagues that can contribute to employee growth.

Theme 3 takeaway: Support and recognition are essential for employee well-being and satisfaction. Recognition, both formal and informal, boosts motivation, while its absence may lead to dissatisfaction. Peer, managerial, and company support enhance well-being through mental, physical, and professional growth initiatives.

2.4.1.4 Theme 4: Work Environment and Culture

This theme explores higher levels in the BM, focusing on the work environment and culture of the participants' company.

Sub-theme 1: Work Environment: Trust, Physical Well-being, and Compensation Several participants mentioned trust as an essential aspect they find and want to keep in their work environment. They mentioned that they are likelier to thrive and contribute positively when they feel trusted. One interviewee emphasised the importance of feeling trusted and having flexibility, stating:

> *"I don't think I would thrive in an environment where they tell me - you need to work from eight in the morning to five in the afternoon every day. Because things happen in life and sometimes you need to be a bit more flexible. So for me, that's really important, flexibility and the trust that comes with that flexibility."* — P1

Moreover, trust extended beyond mere sentiment for the interviewees, reflected in management's actions and policies. They pointed out that when upper management conveys a sense of trust in their abilities, it permeates the organisation. As one employee noted:

> *"They (managers) promote this hybrid work, so we have to go two days a week, even that's not necessary, it's recommended, and they do have the trust, and you feel that they don't micromanage you, you have your own, partially at least, freedom to do yourself. Yeah, really positive culture for sure."* — P9

By trusting participants to manage their time and tasks effectively without needing constant oversight or micromanagement, the company cultivates a flexible and autonomous work environment, fostering a sense of empowerment and accountability.

One more important aspect mentioned by interviewees was the physical work environment. They commented on the physical well-being tied to the physical workspace, including ergonomic and standing desks, chairs and natural light. One interviewee expressed how important it is to consider several factors to create a conducive and comfortable workspace.

> *"We have nice desks and nice chairs and things like that. The desks raise and lower but the general open office area is catastrophic. It's bad light, we don't get any daylight at all."* — P10

This quote illustrates a disparity between the physical comforts provided by the office, such as nice desks and chairs with adjustable heights, and the overall ambience of the workspace, particularly the open office area. Despite ergonomic furniture, the environment is described as "catastrophic," primarily due to poor lighting and the absence of natural daylight.

Finally, the salary and benefits were also considered crucial during the interviews. Several participants expressed contentment with their compensation and benefits, not necessarily because it is high but more due to being happy with other company factors, such as the work environment. A few commented that their salaries need improvement, such as salary transparency and equitable distribution of benefits across job levels.

Sub-theme 2: Company Policies and Practices This sub-theme presents diverse participants' perspectives on company policies and practices, highlighting how these influence their experiences, well-being, and organisational engagement. Interviewees expressed value for well-being programs and initiatives provided by the company, such as wellness allowance, lunch walks, and opportunities for physical activity. Conversely, some others expressed dissatisfaction with the adequacy of these initiatives, suggesting the need for more comprehensive well-being support.

> *"We have asked for higher wellness allowance. The company says no, we will not increase it even though that benefit make the employee feel better or exercise more. They don't promote any well-being efforts or activities. It feels that the company wants to pull in every different cost. I don't think they mind if someone, for instance, sent out and hit the wall. It's not like we have any active prevention of being too stressed; sadly, I'm missing that."* — P12

This quote shows how some interviewees think the company could do more to prevent stress and promote mental well-being. This feeling was shared by several participants, concluding that the provision of wellness programs and health-related benefits are insignificant.

Another essential aspect mentioned by participants was effective communication and collaboration, which were pointed out as crucial components of company culture. Further, participants said they value open dialogue, feedback mechanisms, and teamwork and peer support opportunities. One participant highlighted the importance of these elements by saying:

> *"I feel like they're supporting it by giving quite a lot of room to express my opinions and also be able to affect how we do things."* — P4

The quote shows the importance of a work environment where employees feel heard and empowered to contribute to decision-making processes. Other participants noted the significance of a collaborative atmosphere, indicating that a supportive culture is vital for personal and professional growth. Moreover, structured team events and informal practices such as open-door policies and peer support were commented to play a significant role in fostering a collaborative environment. One employee mentioned:

> *"They tried always to make this mix. To make the people communicate with each other. They remind people in meetings to talk to each other."* — P15

In conclusion, effective communication and collaboration add to a positive work culture. Participants see it as crucial to foster practices that make them know their opinions are considered, feedback is constructive, and there are ample opportunities for teamwork and support.

Sub-theme 3: Company Culture and Diversity Participants commented on various aspects of company culture regarding diversity, touching upon openness to different races, genders, and backgrounds and efforts towards inclusion and equal opportunities. They shared observations and experiences regarding diversity initiatives, policies, the composition of teams, the impact of cultural diversity on workplace dynamics and societal norms regarding diversity and inclusion. While some saw progress and positive steps towards inclusion, others highlighted challenges such as gender imbalances and the persistence of glass ceilings.

One recurring aspect was the participants' acknowledgement of efforts made by their companies to embrace diversity, such as actively recruiting employees from various backgrounds and promoting inclusivity in hiring practices.

> *"We have a lot of employees from different parts of the world, different countries. And we also work with people from other countries."* — P4

Another aspect expressed in the interviews was the impact of cultural diversity on workplace dynamics. Participants shared their opinions on working in multicultural teams and the value they see in having colleagues from different backgrounds. They recognised that diversity brings different perspectives, enriching discussions and problem-solving processes. However, they also acknowledged the challenges that can arise, such as language barriers or cultural differences in communication styles. Despite these challenges, many believed in the importance of diversity and its positive impact on team dynamics and overall organisational culture.

Moreover, interviewees commented on company policies and practices in shaping diversity initiatives. While some employees perceived their companies as actively promoting diversity through recruitment strategies and inclusive policies, others expressed scepticism about the effectiveness of these efforts, and others mentioned they do not mind diversity in their workplace.

One participant shared their experience as a minority and how intersectionality plays a role in broadening the issue of diversity.

> *"So I work in the aviation sector, and that's very male-dominated, very old male-dominated, so it's not so it's not only a sex it's also an age."* — P1

This quote exemplifies the challenges faced in industries with entrenched gender and age biases. More participants also shared stories of feeling alienated or marginalised due to their background, while others expressed gratitude for working in environments where diversity is celebrated. One story is the quote below:

> *"We have really good diverse teams, and about inclusion. Let me tell you one thing. One day, three of my Swedish colleagues were talking to each other and I was there, I was not actively involved in that conversation but these three were talking in English. So I just asked, why are you speaking in English? You can speak in Swedish. So they said, because you're besides us. And if you feel like joining our conversation you can join, if we talk in Swedish then you don't understand. So we have that kind of culture."* — P7

This quote highlights how crucial it is to create spaces where individuals from all backgrounds feel valued and included and have the opportunity to integrate into their workplace.

Theme 4 takeaway: A supportive work environment and inclusive culture significantly impact employee well-being, engagement, and retention. Participants value trust, flexibility, quality workspace, fair compensation, and policies encouraging open communication, collaboration, and wellness. Diversity efforts are appreciated for enriching teamwork. However, challenges like language barriers, gender imbalances and biases still persist.

2.4.1.5 Theme 5: Challenges and Stressors

This theme focuses on the different challenges and factors that contribute to stress in our participants.

Sub-theme 1: Workload and Time Constraints Various factors, including deadlines, customer demands, and the allocation of responsibilities, influence the workload of our participants. They commented that the pressure on them to perform escalates due to the organisation trying to meet delivery targets, particularly when client expectations clash with the organisation's internal capacity. The lack of proper planning leads to a backlog of tasks and increased stress among the interviewees.

For the interviewees, having a sense of control over their workload is essential since it gives them the feeling of handling responsibilities without feeling overwhelmed

by stress. However, they also commented that an overload of tasks and unhappy clients can bring down their motivation and make it hard to get things done. In busy times, organised workplaces provide relief. **Good planning, structured work environments and support from managers** were mentioned as facilitators of handling workload and avoiding feeling overwhelmed.

> *"One thing that I have seen that the company, or at least the department, has done that is quite negative in my point of view is that there have been people agreeing on deliveries with customers while not having first checked that we have the capacity to fulfil that."* — P5

This quote expresses the discussion of workload dynamics, the pressure to meet delivery targets, and the consequences of a lack of proper planning.

> *"When they get frustrated and when people leave. When I started, one guy had just quit without having a new job. Just he needed to get away. It was horrible, apparently. We have that still to some extent, the frustration within the organisation can be... the levels can be high."* — P6

Sub-theme 2: Social Integration and Loneliness One important aspect that directly influenced interviewees' well-being was their social integration and feelings of loneliness and exclusion. Participants expressed that they face challenges when integrating socially into their teamwork and making friends. Several of them have struggled to feel included and build meaningful connections. Feelings of shyness, difficulty initiating conversations, and the absence of a close-knit social circle contribute to loneliness and isolation. Despite being immersed in work environments, participants expressed a longing for deeper connections beyond professional interactions.

> *"I do have these problems with finding the right people, like, the right friends."* — P15

As expressed in this quote, some participants feel lonely at the workplace and in their private lives.

Sub-theme 3: Tech Tools and Their Impact on Communication and Productivity The role of tech tools was mentioned as another factor that can lead to stress, frustration, and delays among interviewees. They commented that they face issues with tools that crash and slow IT department responses. One participant noted:

> *"We have tools that crash a lot, and the IT department needs to be involved because they are so slow. I have software that I need now to do one specific job within one project, and it's the 3rd week, and it took, I don't know how long, it's a standard software that is available on the web, and it took forever to get access to it."* — P6

This participant highlighted the recurring frustration of dealing with unreliable technology. Such problems slow the workflow and cause a ripple effect on project timelines. Furthermore, interviewees also mentioned that restrictive IT policies and outdated tools further hinder productivity, making routine tasks unnecessarily cumbersome and time-consuming.

Another reason mentioned was the inefficiencies in workplace communication, such as unnecessary meetings, that tools like Zoom or Teams promote that could be replaced by emails. Some participants commented that they felt frustrated and preferred face-to-face interactions over virtual meetings for collaboration.

Sub-theme 4: Personal Life Situations Interviewees explained the main factors from their personal life that influenced their work performance and overall well-being at work. One primary concern was that managing personal responsibilities, such as family issues and tasks, added to the stress burden, making it difficult to maintain a healthy work-life balance. Some participants deal with specific situations, such as conditions like ADHD.

One participant shared a scenario when they had to deal with different responsibilities at the same time and how they perceived it affected their mental health.

> *"When we have a lot to do at work and also personally, when there are things I need to take care of, help someone, family, something like that, it can be anything. Sometimes it can be stressful and it affects our mental health."* — P2

Factors such as sleep quality, health issues, seasonal effects like reduced daylight hours in winter, and time spent on social media influence work performance and negatively impact mental health. Furthermore, participants commented that the physical environment and daily routines, such as lengthy commutes, also contribute to stress levels.

Another factor mentioned by participants was **financial pressures**; with inflation rising, managing financial responsibilities, such as mortgages, has become increasingly challenging. On the professional front, interviewees expressed **feelings of inadequacy and pressure** exacerbated by working alongside highly talented colleagues. They noted that a competitive environment can lead to self-doubt and increased stress as they strive to match the perceived performance of their peers. Finally, one participant commented on the **agile way of work**; for those who like structure and clear responsibilities, working in agile negatively impacts their well-being.

Theme 5 takeaway: Participants face multiple pressures, including workload, social integration, technology issues, and personal life demands (e.g. family responsibilities and financial pressures), which collectively impact their well-being and job satisfaction. Additionally, the work environment's competitive nature and agile workflows can exacerbate feelings of inadequacy and add to participants' stress.

2.4.2 Survey Results

This section presents the results of the survey organised by the type of questions, first the demographics, then the Likert scales and finally, the open questions.

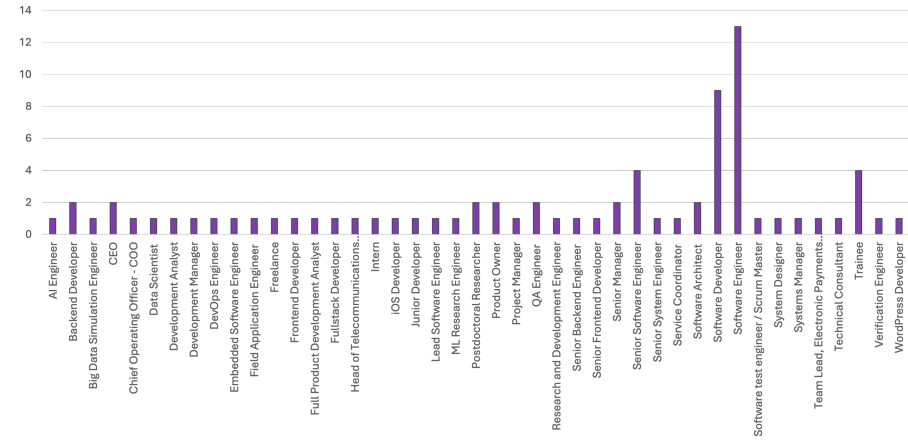


Figure 2.2: Distribution of Survey Respondents by Job Position. Most respondents identified as Software Developers or Software Engineers.

2.4.2.1 Survey Respondent Demographics

We received 83 responses, of which 76 were valid after data cleaning. Respondents came from 14 countries, with the largest groups from Sweden (33), Mexico (17), and Brazil (11). Other countries included Germany (2), Hungary (2), the United States (3), and several countries with one respondent each (Austria, Ecuador, Spain, Ghana, Italy, Netherlands, Poland and South Korea).

Regarding pronouns, 57 respondents prefer "he/him" pronouns, 14 prefer "she/her," 4 opted for no pronouns, and one is comfortable with both "she/her" and "he/him".

Respondents' professional experience varied: 4 (5.26%) have less than 1 year of experience, 11 (14.47%) have 1-2 years, 17 (22.37%) have 2-5 years, 24 (31.58%) have 5-10 years, and 20 (26.32%) have over 10 years. Figure 2.2 shows the job positions of the survey respondents, with software engineers and developers forming the majority.

2.4.2.2 Associations Between Well-Being and Three-Level Variables

We explored the relationship between participants' self-rated overall well-being and variables at the individual (age, gender and years of experience), team (quality of communication and team challenges) and company level (company's culture). Results are presented in Table 2.3.

Concerning age, we found a non-significant association between age and overall well-being, $p = 0.004$. Hence, we concluded that age was unrelated to participants' well-being ratings. Similarly, the variables "Experience" presented no significant differences between groups, "Team challenges" and "Company's culture" showed a weak and non-significant association with well-being.

Conversely, the one variable that showed differences was gender. Since the people who chose "no pronoun" and "both pronouns" were small (4 and 1 sample, respectively), we considered only she/her and he/him to look for differences. Overall

well-being differed between men and women (Wilcoxon rank-sum test, W = 583.5, p = 0.003), with men (median = 5.0) reporting higher well-being than women (median = 3.5). Men reported moderately higher well-being, aligning with prior research [194] showing gendered differences in self-reported mental health. Additionally, "Quality of communication" presented only a moderate positive correlation.

Table 2.3: Summary of Statistical Tests Examining Associations with Overall Well-being. All variables were tested in relation to self-rated overall well-being. Spearman correlations were used due to ordinal or non-normal data distributions. Mann-Whitney U and Kruskal-Wallis tests were used for group comparisons.

Variable	**Test Applied**	**Test Result**	**p-value**	**Interpretation**
Age	Spearman cor	$\rho = 0.004$	.973	No meaningful association
Gender	Mann-Whitney U	$W = 583$	.003	Statistically significant difference; men reported higher well-being
Experience	Kruskal-Wallis	$\chi^2(4) = 5.27$	.260	No significant group differences
Quality of communication	Spearman cor	$\rho = 0.40$	.0004	Moderate positive correlation
Team challenges	Spearman cor	$\rho = 0.14$	.228	Weak, non-significant association
Company's culture	Spearman cor	$\rho = 0.10$	.388	Weak, non-significant association

2.4.2.3 Likert Scale Questions

The results from the Likert scale questions are presented in Figures 2.3, 2.4 and 2.5. The overall well-being of our survey respondents is, in general, good; 38% assessed it as high, 33% as good. Meanwhile, only 8% qualified as low, and we did not get answers with very low.

Due to the nature of the questions, we used different scales for each question. Figure 2.4 shows the answers with their corresponding scale and percentages.

Our results revealed that most participants, 71%, practice activities related to physical health and 51% practice activities for mental health. When asked how often they face challenges with their teams, 67% mentioned that occasionally and frequently, and 33% answered that they experienced negative impacts on their well-being due to colleagues or supervisors.

Most participants, 87%, are satisfied with their work environment and 78% with their compensation. Similarly, most feel respected (91%) and heard (79%). Further, participants' perception of support in general (82%), personal (74%) and professional (64%) was overall high. The quality of communication with their managers and peers was also mostly (83%), rated positively. Finally, 48% of participants commented that their company's culture has an important influence on their well-being.

Figure 2.5 illustrates participants' views on whether their companies promote equality, equity, diversity, and inclusion (EEDI) and whether this promotion or its lack affects their well-being. Most respondents indicated that their companies actively support EEDI initiatives and that these efforts positively impact their well-being.

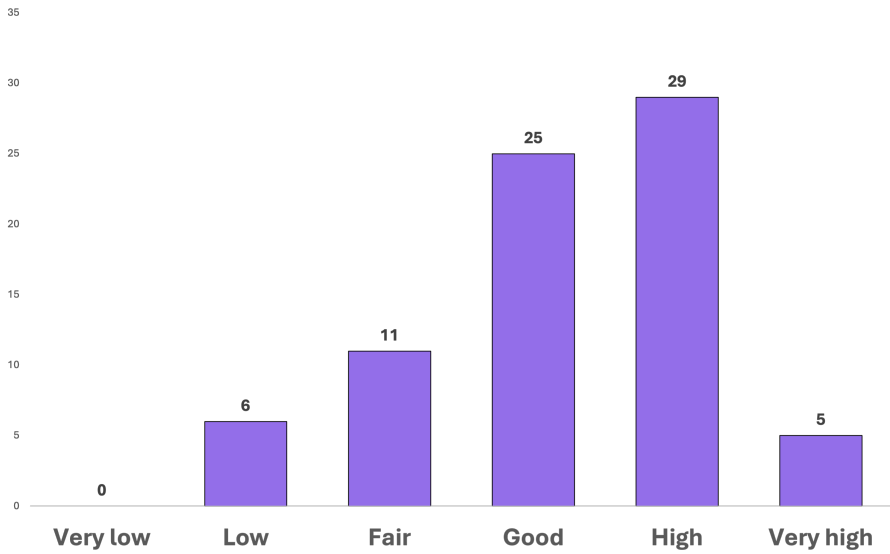


Figure 2.3: Distribution of self-reported overall well-being levels among software engineer respondents (n = 76).

Two questions were about the factors contributing positively and negatively to the respondents' well-being (Q9 and Q10), and neither used a Likert scale. The questions were closed and participants had to choose specific answers, see Figure 2.6, 2.7 for the result of Q9 and Table 2.4 for Q10. For the results in Figure 2.6 and 2.7, participants had to rank from 1 to 7, a list of factors that contribute positively to their well-being. An average of the responses was made to obtain a visualisation, hence, the lowest average was the factor that was closest to 1 (most important), Flexible Work Environment. Furthermore, Figure 2.7 shows how many times each aspect was ranked as number 1, Personal Well-being Activities.

Regarding the factors or challenges they face in their workplace. Table 2.4 shows their answer in order of frequency. Personal life stress was chosen most times, followed by a high workload. Excessive screen time and seasonal affective factors were mentioned the least.

2.4.2.4 Open Questions

The open questions are presented in the coming subtitles. These questions were optional, hence, the number of them was less in comparison with the Likert questions.

Other factors that Influence Respondents' Well-being When asked about other factors besides the ones in Table 2.4 that negatively impact their well-being, participants mentioned that at an individual level, extended periods of isolation and

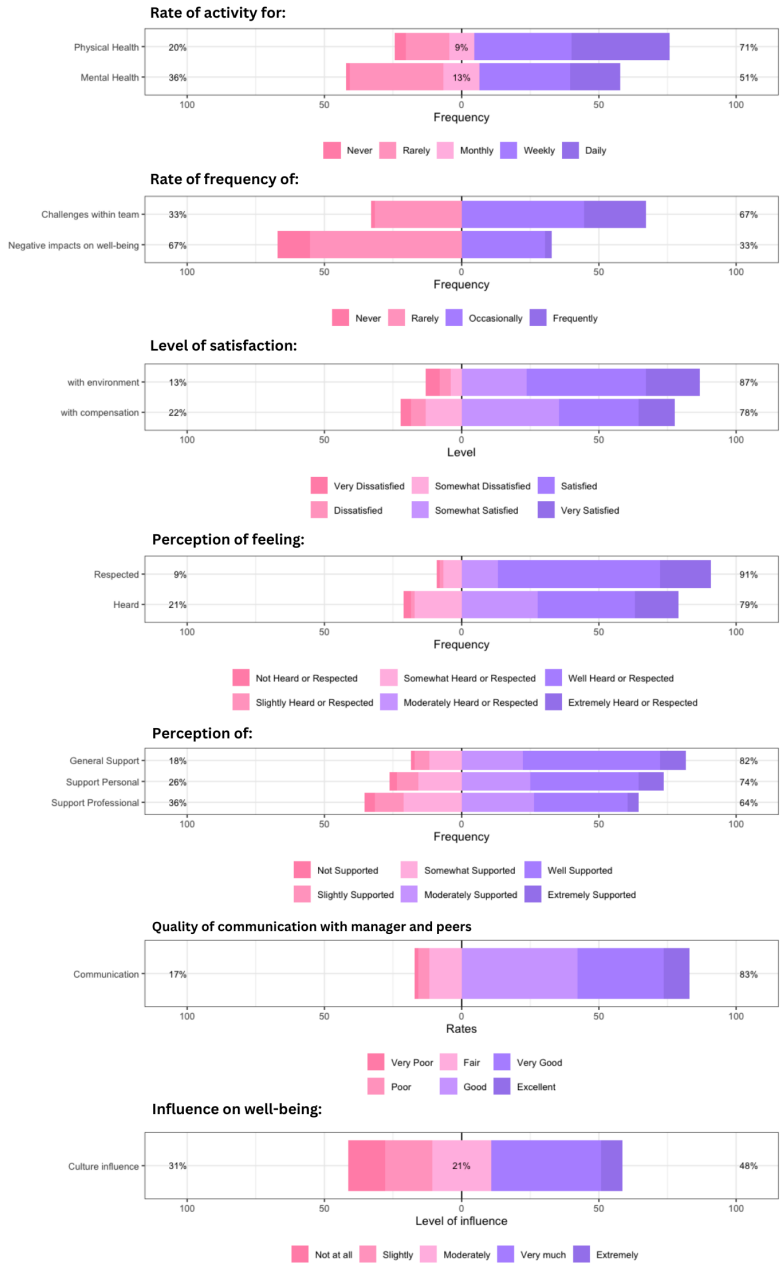


Figure 2.4: Diverging stacked bar chart showing response distributions (in percentages) for multiple Likert-scale survey questions (n = 76). Questions addressed distinct aspects of workplace experiences.

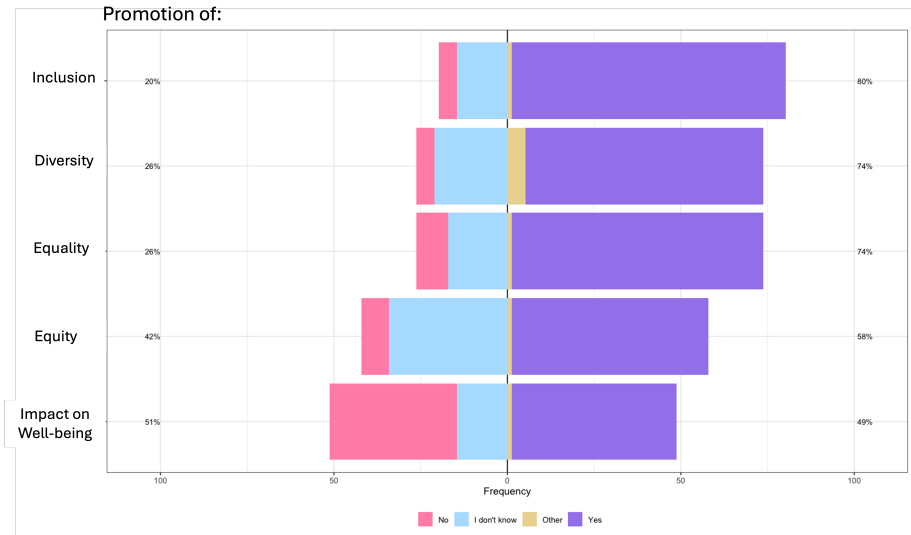


Figure 2.5: Employee perceptions of company-driven equality, equity, diversity, and inclusion (EEDI) initiatives and their perceived impact on well-being.

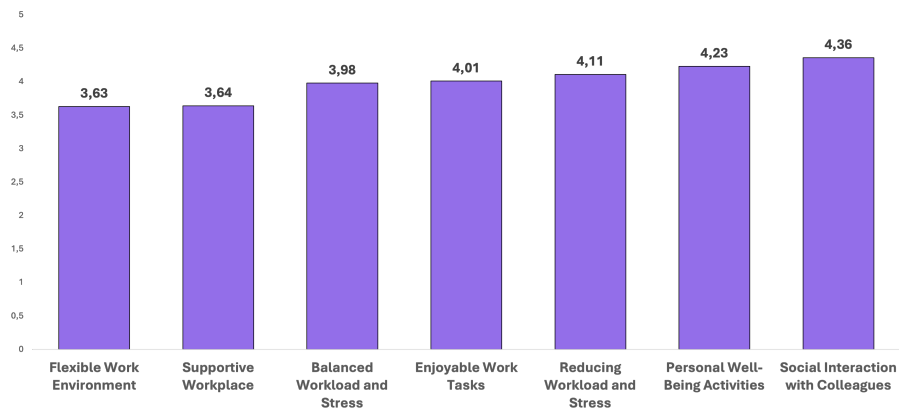


Figure 2.6: Factors positively influencing software engineers' workplace well-being, with average scores derived from survey responses. Lower values (close to 1) indicate stronger perceived benefits.

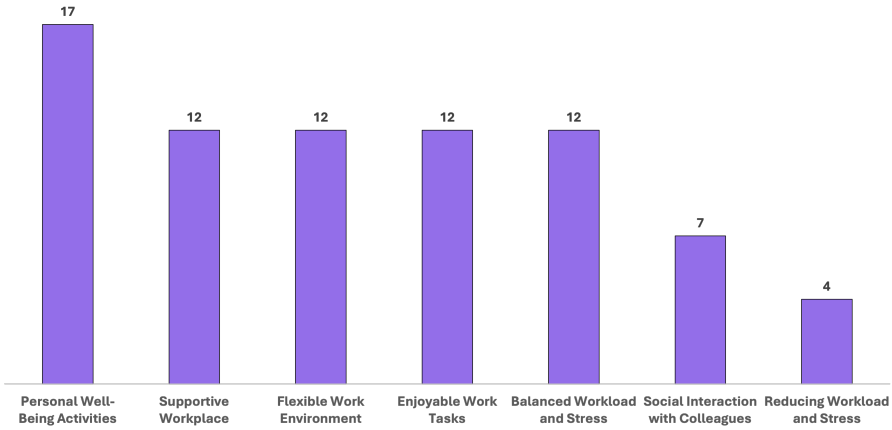


Figure 2.7: Factors contributing positively to respondents' well-being in the workplace. The graph shows the number of times each factor was chosen as the main factor contributing to their well-being.

Table 2.4: Workplace challenges affecting software engineers' well-being, with mention frequency from survey data (n = 76).

Factor	**Num of Mentions**
Personal life stress	47
High workload	40
Tight deadlines	35
Challenges related to workplace communic. with managers	26
Challenges related to workplace communication with peers	22
Pressure to keep up with rapidly changing technology	21
Seasonal affective factors, especially during winter	18
Excessive screen time	15

issues related to migration are additional stressors. One respondent, a startup co-founder, feels a profound effect on their well-being based on the company's successes and failures. Concerning social interactions, participants mentioned peer pressure, the mental health issues of coworkers, boring relationships in the workplace, a hostile work environment and communication issues, particularly with clients, as important factors. Regarding the company level, two participants cited traditional work environments with rigid schedules and resistance to hybrid or remote work as unnecessary and detrimental. An overwhelming workload, especially in areas outside one's expertise, micromanagement and the routine nature of work further contribute to a negative sense of well-being in the workplace.

On the opposite side, there were various answers regarding factors that positively

influence respondents' workplace well-being besides the ones reported in Figure 2.6. One participant mentioned powerlifting as the only thing that works for them. Technological tools, specifically GPT-4, were also highlighted as beneficial, with one respondent expressing a positive impact from interacting with this AI. Food availability and quality play a crucial role; one participant commented that having a reasonably priced cafeteria and snack bar on-site allows them to not worry about meal preparation. Social interactions and recognition within the workplace were part of the answers, too; respondents cited the enjoyment of talking with friends at work and the positive effects of feeling listened to by management. Opportunities and recognition also emerged as key to enhancing well-being.

Influence of Company Culture on SE Well-Being The company culture significantly impacts SE's well-being, influencing various aspects such as work-life balance, inclusivity, engagement, support, management, mental and physical health, social interaction, motivation, and growth opportunities. Positive cultures that emphasise flexibility, support, inclusivity, meaningful work, and transparent management contribute to higher employee satisfaction and well-being.

On the other hand, cultures that lack these elements can lead to stress, demotivation, and a negative impact on overall well-being. For instance, seven respondents mentioned that effective management and leadership are critical. The positive side includes transparency, collaborative environments, and a no-blame culture fostering safety and growth. The negative side includes poor management, a lack of understanding from leaders, and hostile treatment towards employees.

Engagement is driven by meaningful work and alignment with personal values, based on five responses. Employees feel demotivated when their work seems pointless or disconnected from their values. Conversely, having a say in decision-making and understanding the company's goals enhances engagement. Similarly, five other participants agreed that the culture around work-life balance significantly affects their well-being. They appreciate flexible work hours, support for remote work, and the absence of micromanagement, all of which contribute to a comfortable and stress-free work environment. Further, a supportive environment, characterised by fun projects, the ability to change assignments, opportunities for continuous learning, a growth mindset, and group activities, were recognised as crucial for five respondents. Social events and team-building activities help employees build personal connections, which currently need to be improved in some companies.

Regarding diversity, participants mentioned that a welcoming and inclusive culture, with representation of different people, positively impacts them. However, a lack of inclusivity, such as language barriers, can lead to fewer opportunities and feelings of exclusion. While some individuals feel unaffected by these initiatives, others report significant negative or positive impacts on their professional and personal lives. Respondents highlighted feelings of exclusion, frustration, and demotivation in environments that fail to promote EEDI. They emphasised the importance of feeling included, respected, and valued in the workplace. The mixed nature of the feedback suggests that while EEDI is a crucial factor for many, its importance varies widely depending on individual circumstances, work environments, and personal values.

Finally, three more answers talked about how having excessive meetings, high

pressure, and a lack of understanding from management, as well as a focus on speed over quality, can lead to burnout and decreased motivation.

Respondents' Feedback on Workplace Relationships The answers collected indicate a general positive sentiment towards relationships with managers and peers. Participants mentioned communication, friendship, and supportive relationship dynamics. Six participants commented on positive relationship dynamics in their workplace. They mentioned having good and open-minded relationships with colleagues and managers, working well together, and having friendships between managers and team members; they also highlighted how these aspects positively affect the work environment and good camaraderie. Three participants commented on the crucial role of communication and tone in their workplace. Two more respondents mentioned the importance of addressing individual differences and providing support when needed. Finally, one mentions clients' behaviour and its impact on internal team dynamics and relationships.

Conversely, some participants mentioned several workplace challenges and areas for improvement. Issues such as perceived internal divisions, boring tasks, non-useful online meetings, and the mismatch between job demands and employee capabilities are notable. Additionally, one participant mentioned a need for greater transparency and acknowledgement. Further, one respondent mentioned that it is unnecessary to interact with coworkers outside of the workplace. At the same time, another commented that building bonds of trust with people who only listen to you for 15 minutes in the morning is complicated. Finally, one last answer mentioned understanding and mitigating generational clashes and challenges to separate friendship and professional relationships as factors present in their current workplace.

Regarding maintaining team cohesion and effectiveness under stress, respondents indicated that effective communication, peer support, and strategic organisation are crucial for the team to achieve their goals. While many teams have developed robust strategies to cope with pressure, some struggle with disorganisation and over-reliance on individuals. Cultivating a supportive team environment and ensuring flexible, realistic planning appear to be critical factors in sustaining team well-being and productivity during challenging periods.

Recommendations Given by Participants to Support their Well-being In their recommendations, several participants mentioned that hybrid work should be allowed. Some commented that working from home has been great for their well-being. Additionally, they also recommended flexibility in schedules.

One more participant commented on having workouts 2 -3 hours per week, walks, breaks to relax and more exercise activities. They also commented on giving complete or at least increased friskvårdsbidrag (Swedish health care allowance). Better salaries, bonuses for good work, more benefits, and considering effectiveness without putting pressure or micromanagement were also mentioned.

There were several points about managers, such as clarity in the tasks of managers and leaders, choosing qualified managers who know how to manage a team, prioritising personal coaching or mentoring over a traditional manager relationship, giving and

implementing feedback, viewing employees as humans, and improving managers' training in human aspects to transmit knowledge and skills to their employees more effectively.

Some other recommendations were creating better workspaces designed to improve focus. Note that constant firing can decrease commitment, as employees may feel insecure about their job stability. It was also recommended to focus on increasing employee interaction, having informal meetings to discuss their challenges, and listening to their basic comfort needs. However, some other participants recommended reducing the number of meetings. They mentioned that addressing migrant issues can support their well-being, too. Additionally, employees appreciate having fruit baskets and plants in the office, which can contribute to a more pleasant and motivating workspace.

Final Thoughts by Participants on the Personal Well-being and the Well-being of Software Engineers in General The final question was about anything participants wanted to add to their well-being or the software engineers' general well-being. The answers highlighted various experiences, challenges, and recommendations. Key well-being factors include maintaining a healthy work-life balance, accessing good work tools, fostering social interaction, and establishing personal routines. Respondents also valued environments that allow them to grow and feel connected to their work, and they recognised the importance of managing stress to maintain mental and physical health.

Furthermore, participants emphasised the importance of taking breaks, such as walks, to maintain well-being. Some expressed difficulty connecting with their employer and having difficulties finding motivation to perform well at work. Working from home was seen as beneficial for balancing work and family life, though it could blur the line between work and personal time. Others praised AI tools for easing their workload and improving productivity. Additionally, several respondents stressed the importance of physical exercise, proper ergonomics, and a good sleep routine. Some mentioned that software development can be lonely, and regular social interaction is necessary for well-being. Finally, they highlighted that motivation and enjoyment in work are crucial for maintaining overall satisfaction.

2.5 Discussion

Our findings reveal factors that influence software engineers' well-being across individual, team, peer, and organisational levels and indicate a varying significance of these factors.

In this section, we triangulate the qualitative and quantitative data to identify the main similarities and differences between our interviews and the survey. Triangulation was conducted by comparing each theme with patterns observed in the survey, examining where responses converged, complemented, or diverged. We then situate our findings within existing literature and frameworks, noting areas of alignment and divergence.

Table 2.5 summarises the themes and sub-themes from the qualitative analysis aligned with the survey findings. Quantitative indicators and notable frequencies are

included to support each sub-theme.

Table 2.5: Summary of Survey Results compared to Interview Themes and Sub-themes

Theme	Sub-theme	Survey Results
1. Personal Conception of Well-being	n.a.	77.6% rated their well-being as good to very high. Personal well-being activities were the main workplace well-being factor.
2. Personal and Collaborative Factors	Personal Practices	71% engage in physical health practices. 51% in mental health activities.
	Influence of Social Interactions on Well-being	91% feel respected; 79% feel heard; 83% rated communication as good–excellent.
		63% face team-related challenges; 67% rarely/never experience negative impacts from colleagues/supervisors.
3. Support and Recognition	Recognition at Work	3 respondents mentioned recognition.
	Support from Company and Peers	82% feel supported by peers and company.
	Support for Growth	64% feel supported in their professional growth.
4. Work Environment and Culture	Trust, Physical Well-being, and Compensation	78% are satisfied with their compensation.
	Company Policies and Practices	87% are satisfied with the work environment.
	Company Culture and Diversity	48% reported a strong culture impact on well-being. 51% were unsure or saw no DEI impact.
5. Challenges and Stressors	Workload and Time Constraints	40 times, high workload was cited as a challenge.
	Social Integration and Loneliness	3rd-ranked workplace well-being contributor was colleague interaction.
	Tech Tools	21 mentions of technology as a challenge.
	Personal Life Situations	47 times personal life stress was mentioned as a challenge.

2.5.1 Alignments in Quantitative and Qualitative Results

This section examines how the interview and survey data converged.

2.5.1.1 Personal Practices

Interviewees and survey respondents reported frequently exercising three to five times weekly to support their well-being. This emphasis on exercise may serve as a coping mechanism, especially in a profession characterised by long hours, sedentary work, and high mental demands.

Survey results further indicate that personal well-being activities were identified 17 times as the most important factor influencing well-being, mirroring the insights from interviews and aligning with findings by Tsatsoulis and Fountoulakis [195]. This reflects a strong sense of personal agency among software engineers, who actively engage in activities that help them decompress outside work. While personal efforts

like exercise are undeniably valuable for stress management and maintaining well-being, broader factors, such as long work hours, high cognitive demands, and a culture that often undervalues well-being, also require attention.

2.5.1.2 Support from the Company and Peers

Participants in the survey and interviews emphasised the importance of support from their company and peers. Survey results show that most respondents felt supported by company initiatives promoting a healthy work environment and employee well-being. Similarly, interviewees highlighted how peer support, specifically in work-related matters, fosters a positive and inclusive atmosphere, benefiting their well-being and mental health. These findings align with previous studies by Hirschle [196] and Russo [173], which also identified support as a key factor in mitigating the negative effects of stress and closely linked to increased productivity.

While employees report feeling supported by their company, the prevalence of stressors related to workload and time constraints indicates that these initiatives may not tackle deeper, systemic issues. The support provided seems not to extend to critical organisational changes such as improved project management, more realistic deadlines, greater workload flexibility, and enhanced support for hybrid work, issues frequently raised by participants.

2.5.1.3 Work Environment: Trust, Physical Well-being, and Compensation

In the survey, participants expressed high satisfaction levels (87% and 78%) with their compensation and work environment. Meanwhile, in the interviews, participants emphasised trust as a key factor in their work environment. Interviewees highlighted that trust, particularly from management, was critical to their sense of well-being and ability to perform effectively. Trust was linked to a positive emotional state and practical aspects, such as flexibility in their roles and decision-making; this aligned with de Guerre et al. [197], who found trust to be an enabling condition for mental health in organisations. Further, according to Syahreza et al. [198], compensation and work environment significantly impact employee satisfaction at work.

Satisfaction with financial compensation and physical work conditions is essential to maintaining a baseline level of employee contentment. However, these elements alone do not capture the full complexity of what makes a workplace genuinely supportive.

2.5.1.4 Equality, Equity, Diversity, and Inclusion (EEDI)

Several participants commented during the interview on the company culture's openness to diversity, recognising efforts toward inclusion and equal opportunities. They acknowledged progress through diversity initiatives and team composition, while others pointed out persisting challenges, such as gender imbalances and glass ceilings.

In the survey, most respondents reported that their companies promote EEDI, which aligns with the interviewees. An important point to note is that most of the participants identified themselves with the pronoun "him" (57/76). In contrast, the pronoun "her" (14/76) and other pronouns (4/76) are a minority in our popula-

tion. The minorities expressed stronger concerns and elaborated on their challenges, emphasising the need for a welcoming and inclusive culture.

There was a particular emphasis on how language barriers and lack of inclusivity often led to feelings of exclusion and missed opportunities, directly impacting their well-being. De Souza and Gama [199] obtained similar results when researching diversity and inclusion in IT companies.

While some respondents, particularly those from majority groups, were unaffected by EEDI initiatives, others reported both positive and negative impacts on their personal and professional lives. De Souza and Gama [199] argue that the active involvement of majority groups in diversity efforts is crucial for driving change. However, achieving this can be difficult if these groups do not perceive the need for such change.

2.5.1.5 Personal Life Situations

Participants mentioned that their personal life situations can significantly impact their well-being, positively and negatively, depending on the circumstances. Situations such as managing personal responsibilities, particularly family issues, added to work-related stress; however, supportive relationships and fulfilling personal activities were also highlighted as sources of positive well-being. Conditions like ADHD and challenges in balancing work and life were common themes. Our survey data confirmed this, with 47 respondents citing personal life stress as a significant factor affecting well-being. These results align with other studies identifying factors that contribute to poor mental well-being at work, such as Teevan et al.'s [200] study finding integration of work and personal life, as well as de Guerre et al.'s [197] listing interpersonal conflicts as one of them.

Many employees struggle to balance family responsibilities, personal challenges, and work demands, which likely stems from the rigidity of organisational structures. These structures typically lack the flexibility to accommodate diverse needs, such as flexible working hours or support for managing ADHD or family care responsibilities. As a result, employees are often expected to sustain high productivity while managing significant personal stressors without sufficient support. Although we did not analyse country-specific differences, it is clear that broader systems shape individual experiences differently (for instance, Sweden offers a better work-life balance) and offer a stark contrast. In organisations with rigid structures, the absence of flexible schedules, mental health resources, or accommodations for neurodivergent employees intensifies stress and diminishes employee engagement.

2.5.1.6 Workload and Time Constraints

Tied to the previous factor is the workload and time pressures. Participants reported that deadlines, customer demands, and poor allocation of responsibilities significantly impact their work experience. In the interviews, a lack of proper planning led to backlogs and increased stress, with employees feeling overwhelmed when client expectations exceeded the organisation's capacity. Survey responses also highlighted that high workload (40 respondents) and tight deadlines (35 respondents) were prominent sources of stress, which aligns with Scholarios and Marks' [201] and Teevan et al.'s [200] findings.

An overload of tasks, particularly in environments with poor planning, leads to a demotivated workforce and, without proper intervention, risks burnout and decreased long-term productivity.

2.5.1.7 Social Integration and Loneliness

Interviewees frequently discussed the difficulties of integrating socially within their teams and forming meaningful connections, especially in contexts where shyness or a lack of social support networks created barriers to inclusion. Geographic factors (the country) and migration were mentioned as amplifying these feelings of isolation. Survey results support this, with participants citing social isolation as a significant stressor and naming issues like peer pressure and hostile workplace interactions. Other studies, such as D'Oliveira and Persico [202], have reported on the effects of isolation on workplace well-being, colleague and supervisor satisfaction, job satisfaction, and organisational commitment, aligning with our results.

The challenges of socialising and the resulting loneliness reflect individual characteristics like shyness and a workplace culture that may not facilitate inclusion or collaboration. This isolation is particularly pronounced for those who may be migrants or part of minority groups, as participants commented.

2.5.1.8 Tech Tools and Their Impact on Communication and Productivity

Both survey and interview participants mentioned frustrations with tech tools. These tools, such as Zoom or Teams, were seen as sometimes creating unnecessary meetings that could be replaced with emails, hindering productivity, which aligns with findings by Nawrat [203]. Additionally, respondents complained about slow IT responses and tech tools that frequently crashed, leading to inefficiencies in communication and frustration.

2.5.2 Contrasting Quantitative and Qualitative Results

This section explores where the data from the interviews and the survey presented contrasting views.

2.5.2.1 Influence of Social Interactions on Well-being

Interview participants highlighted positive social interactions and connections as crucial for emotional support and resilience in the workplace, directly influencing their well-being. They associated these connections with a sense of belonging, emotional support, and mental health, emphasising that positive workplace interactions create a more fulfilling and supportive environment.

In contrast, the survey results did not place as much emphasis on social interactions as a key factor in well-being (see Figures 2.6 and 2.7). While participants acknowledged social aspects (such as communication, friendship, and supportive relationships), these were framed as contributing factors rather than primary concerns. Other factors, such as personal well-being activities, flexible work environments, and overall workplace support, ranked higher in terms of impact on well-being.

The survey's lower prioritisation of social interactions may stem from participants focusing on more direct and measurable aspects of their work experience, such as workload, while viewing social dynamics as secondary. In contrast, interviews gave participants more time to reflect on the broader factors affecting their well-being.

2.5.2.2 Recognition at Work

During the interviews, participants mentioned that feeling recognised and valued at work plays a significant role in their motivation and well-being, emphasising the importance of recognition. Meanwhile, in the survey, recognition was mentioned only as an "other factor," with some respondents citing the positive effects of feeling listened to by management. However, it was not highlighted as a major contributor to well-being; it was grouped with minor factors.

2.5.2.3 Professional and Personal Growth Support from Companies

Regarding companies' professional and personal growth support, there were some differences in perceptions and opinions in the interview and survey. In interviews, participants expressed mixed feelings. Some felt there was a disconnect between their personal development goals and what the company offered, while others appreciated efforts like learning platforms and goal-setting opportunities. The survey respondents briefly mentioned growth opportunities as part of the overall company culture's impact on well-being. However, it was not a prominent focus compared to other factors like work-life balance and inclusivity. This aspect needs more research to draw solid conclusions; participants acknowledge the importance of growth opportunities and the need to align with individual career paths. Companies may need to tailor their initiatives to reach their employees' expectations and goals.

2.5.2.4 Company Policies and Practices

Interviewees valued company policies and well-being initiatives like wellness allowances and physical activity opportunities, but expressed mixed feelings. Some commented to appreciate these efforts, while others felt insufficient and called for more comprehensive support. The survey highlighted broader aspects of positive workplace cultures, emphasising flexibility, support, inclusivity, and meaningful work as critical contributors to well-being. Hybrid work and work-life balance were frequently mentioned, but these were not mentioned in interviews. The difference in opinions can be due to the participants' contexts. All the interviews were done in Sweden, where hybrid work is already well established, while the survey covered different countries. Such countries may not have adapted hybrid work as Sweden has.

Table 2.6: Comparison of our Framework to Other Well-being Theories

Our Framework	**Gallup's Five Elements of Well-being**	**Seligman's Five Pillars of Well-being**	**Michaelson's Pillars**	**Added Value of Our Framework**
Personal Conception of Well-being	-	-	-	Directly addresses personal interpretations of well-being, which the other frameworks overlook
Personal and Collaborative Factors				
Personal Practices	Physical well-being	Positive emotion	Emotional well-being, vitality, resilience, and self-esteem	Combines physical and emotional factors, acknowledging a broader scope of personal well-being practices
Influence of Social Interactions on Well-being	Social well-being	Relationships	-	Incorporates formal and informal social interactions inside and outside work, which are not fully considered in other frameworks
Support and Recognition				
Support from the Company and Peers on Well-being	Community well-being	-	-	Focuses on organisational and peer support, offering a more detailed look at company-level factors
Recognition at Work	-	Accomplishment	-	Directly addresses the impact of individual recognition at work on well-being, whereas others focus more on outcomes (e.g., accomplishment)
Professional and Personal Growth Support from Companies	Career Well-being	Engagement	-	Emphasises the dual impact of personal and professional growth on well-being
Work Environment and Culture				
Work Environment: Trust, Physical Well-being, and Compensation	Financial well-being	-	Positive functioning	Expands on workplace well-being by addressing trust and compensation in addition to physical and financial aspects
Company Policies and Practices	-	-	-	Considers companies' well-being policies into the broader factors influencing well-being

Our Framework	Gallup's Five Elements of Well-being	Seligman's Five Pillars of Well-being	Michaelson's Pillars	Added Value of Our Framework
Company Culture and Diversity	-	-	-	Considers companies' culture and efforts to achieve diversity as factors that contribute to well-being
Challenges and Stressors				
Workload and Time Constraints	-	-	Positive functioning	Acknowledges the impact of workload and time pressures more explicitly than the other framework
Social Integration and Loneliness	-	-	-	Stresses the importance of a person's sense of belonging and its influence on working life
Tech Tools and Their Impact on Communication and Productivity	-	-	-	Elaborates on how technology hinders and enhances work and its impact on well-being
Personal Life Situations	-	-	-	Acknowledges the positive and negative influence of personal life situations on working life

2.5.3 Comparison to Other Theories of Well-being Factors

This section compares our framework to other authors' theories. In Table 2.6, we align our findings with Gallup's five elements of well-being, Seligman's five pillars of well-being, and Michaelson's pillars. Many of our themes align with these authors' proposals regarding the **multidimensional nature of well-being**. We placed each theory in a separate column and listed the pillars or components that align with ours. We indicated this with a '-' where there was no alignment. Consistent with these theories, our study acknowledges that well-being is shaped by various factors, including emotional, psychological, social, and economic dimensions. Additionally, following Michaelson's work [9], we advocate for integrating well-being into public policy, recognising that it reflects a broader understanding of the quality of life beyond economic growth alone.

While Wong et al. [87] study provides important insights into **internal experiences** and some **organisational factors**, we affirm that a more comprehensive approach is needed, one that considers and balances the external factors shaping well-being. Drawing on international data, our framework critically examines how workplace dynamics and external pressures interact across organisational contexts, offering a more comprehensive understanding of how to support well-being in diverse work environments.

Table 2.7: Policy Recommendations Based on Well-being Themes

Theme	Recommendation
Personal Conception of Well-being	Provide access to well-being and mental health resources, self-reflection exercises, and goal-setting programs that allow employees to understand their unique needs and preferences regarding well-being. Encourage and role model using such resources and activities to establish a caring culture. *Examples: Post a weekly reflection question in the coffee corner; let teams set a well-being goal with an indicator that helps keep an eye on it; include a movement break in longer meetings [204].*
Personal and Collaborative Factors	Create policies promoting individual well-being practices and positive interpersonal interactions at work using team-building activities and peer support networks. Encourage informal creative working spaces for ideas to flourish. *Examples: Make a community corner with creativity games and puzzles; make well-being an explicit concern to discuss in employee reviews; plan team events to enhance morale and connection [205].*
Support and Recognition	Implement support systems that acknowledge and recognise professional achievements. Similarly, strategies should be implemented to provide guidance and emotional support to ensure employees' well-being. Facilitate formal and informal mentoring and establish a visible role model culture. *Examples: feeding back directly when someone does good work; being available for regular work-related conversations; asking what support they need to help them achieve their goals [205].*
Work Environment and Culture	Develop policies that ensure a supportive and inclusive work environment, including trust, fair compensation, and diversity. Provide space and opportunity for the expression and exploration of local culture and the diversity of cultures if employees are from elsewhere. *Examples: Invite a speaker on well-being and mental health to an event as part of activities for diversity; encourage your team to adopt healthier working habits; normalise conversations about mental health [205].*
Challenges and Stressors	Create flexible work policies that address workload management, social integration, and the use of technology. Provide parental leave or other care support to allow an employee to flourish while also fulfilling family duties. *Examples: provide flexible work hours and hybrid work (supporting employees in balancing good work ethics with other life demands), offer virtual coffee breaks or social events for remote teams, provide training in different tools for workload and task management [204].*

In contrast, Wong et al. [87] primarily focus on poor mental well-being at work, addressing individual and organisational challenges, such as company culture, organisational policies, and personal coping strategies. While Wong et al.'s framework touches on external factors like organisational culture and technologies for mental well-being, it primarily focuses on internal self-reported well-being experiences and the strategies software engineers use to manage it. This inward-looking focus, while important, leaves the broader and more systemic external factors unacknowledged that influence well-being, particularly those that are not under the direct control of

individuals, such as workload demands, leadership dynamics, or cultural differences. Our approach expands on Wong et al.'s results by emphasising the role of these external pressures at every level (individual, team, and organisational). For instance, while Wong et al. acknowledge organisational challenges like company policies and culture, our research critically examines how specific external factors such as compensation, leadership practices, and structural job demands directly affect well-being.

We argue that well-being is not just about how individuals or organisations manage mental health but also about how external factors shape the experience of well-being.

Additionally, Wong et al.'s study focuses on a U.S. population, which limits the generalisability of its conclusions. Our research expands the scope to include software engineers from various countries worldwide, enhancing the generalisability of our findings across diverse organisational and societal contexts

2.5.4 Policy Recommendations

In our research on well-being over the past five years, we observed that companies are unlikely to invest in well-being interventions beyond current policies. Recognising this, we have developed policy recommendations based on our research findings to enhance future policymaking. These recommendations target software development companies, particularly those in regions where well-being practices are less established or institutionalised.

Our recommendations are grounded in a rigorous analysis of the empirical data we collected through surveys and interviews for this study. By exploring well-being factors at individual, peer, managerial, and organisational levels, we identified key patterns, challenges, and opportunities related to the well-being of software engineers, and we reflected those findings in our guidelines.

One of the clearest and most consistent interpretations across our findings is the necessity for **flexible work policies that address workload management**. Several of our participants commented on their need for flexibility in their workplaces to ensure they take care of their needs outside work. Moreover, this recommendation stems from evidence indicating that flexible schedules can reduce stress and enhance productivity.

Table 2.7 shows these evidence-based recommendations on useful guidelines to 1) offer a roadmap for companies to effectively enhance the well-being of software engineers and 2) bridge the gap between research insights and practical policy. We aim to motivate organisations to implement measures beyond their current well-being frameworks, ultimately promoting a healthier and more resilient work environment and, hence, more resilient software engineers.

Most of the policy recommendations in Table 2.7 and implementation examples are not groundbreaking. Since they are based on empirical employee data, they reflect and respond to their needs. If we were to ignore that, we would not be basing our recommendations on the empirical data we gathered. In some contexts, these measures remain aspirational due to organisational constraints or cultural resistance, which makes their continued emphasis even more important. ***Best* practice is not always *common* practice.** That means our recommendations still include flexible work hours, hybrid work options, and supportive managers who check on employees'

well-being and workload, because these practices are not yet common everywhere. They still need to be implemented daily in many companies and countries. For further best practices and detailed examples on establishing them in companies, we point the inclined reader to [9, 204, 205].

From the experience of the second and third authors, who have held management positions for several years, the policy recommendations accurately reflect needs that employees repeatedly raise in practice, in one-on-one conversations (similar to some of the interview questions) and in yearly employee surveys (similar to related survey items). While the examples in Tab. 2.7 are neither new nor surprising, they reflect practical steps still required to strengthen software engineers' well-being in everyday work settings. A simple way to see the relevance of these measures is to consider whether one consistently gives direct feedback when someone does good work or makes time for regular work-related conversations, even during periods of heavy workload. We therefore encourage readers to not only consider the policy recommendations but also renew their commitment to putting them into daily practice.

2.5.5 Validity Threats

This section outlines our study's possible threats to internal, external, and construct validity and the mitigation strategies we implemented. By identifying these threats and proposing mitigation strategies, the study aims to enhance the credibility of its conclusions about the factors influencing software engineers' well-being across different contexts.

2.5.5.1 Internal validity

To affirm our internal validity and deal with selection bias, we targeted different sectors of software companies and engineers with different backgrounds.

One more aspect we considered was the response bias. Participants may have given socially desirable answers during interviews, particularly when discussing sensitive topics like EDI, well-being or mental health. To encourage honest responses, we ensured anonymity and confidentiality during the interviews. Piloting the interview guide and survey helped refine the questions' wording and tone to encourage more authentic answers. We also asked open-ended questions and used indirect questioning techniques to reduce pressure on participants to conform to perceived social norms.

2.5.5.2 External validity

We acknowledge that the external validity can be compromised since our interviews were done only with software engineers working and living in Sweden, which may not represent the general population of software engineers. To mitigate this threat, we used purposive sampling to ensure diverse participants within Sweden (gender, ethnicity, cultural background, country of origin and company size) to capture varied perspectives. Further, we targeted a broader sample with the survey. To ensure diverse representation, we aimed to recruit globally through various channels, including professional networks, social media, and industry groups and made our survey available in three different languages.

Cultural and linguistic differences may influence perceived well-being, leading to inconsistent or incomparable results across regions. To mitigate this threat, we adapted the survey culturally in each language [206] and worked with local experts to ensure that questions made sense in each context.

2.5.5.3 Construct validity

To ensure construct validity in the interviews, we defined concepts such as well-being, diversity, equality, equity and inclusion and gave examples for the interviews to make them explicit. Meanwhile, in the survey, we added definitions to the questions to ensure consistency in understanding across participants. Further, we ensured the translations aligned with the three languages we used. Continuing with the languages, we performed a thorough back-translation of surveys and engaged local experts to ensure cultural nuances were considered. Pre-test translated surveys with small groups in each language to identify any problematic terms or misunderstandings.

Furthermore, we also tailored different scales to the questions in the survey in a way that measures each conception suitably.

2.5.6 Future Work

While our study presents several high-level policy recommendations grounded in the identified well-being themes, future research is needed to translate these into more concrete, context-sensitive interventions and guidelines. Our current methodological approach was not designed to elicit or evaluate practical solutions tailored to specific workplace scenarios. As such, although we offer illustrative examples of how the recommendations might apply in practice (see discussion of Table 2.7), these should be seen as preliminary rather than exhaustive or prescriptive.

Our plans for future work involve participatory design sessions, co-creation workshops with stakeholders, and longitudinal field studies to develop, refine, and assess targeted well-being interventions and a region or country focus. Said plans will allow for the generation of more actionable guidance and support companies in implementing and achieving high-level well-being goals in effective and sustainable ways within their specific organisational cultures and constraints.

2.6 Conclusion

To identify the main factors influencing software engineers' well-being, we conducted interviews in Sweden and ran a survey in three languages globally. We reported our main findings in this article.

Our study reports the main factors influencing well-being, such as **personal perception of well-being, personal and collaborative factors, support and recognition, work environment and culture, and challenges and stressors**. We confirmed the factors identified by research in other fields [9,77] and offered unique contributions specific to the software engineering context.

First, we strengthen the existing body of evidence by analysing these factors in a field where high cognitive demands and constant technological evolution intensify their impact. Second, our framework provides a **higher level of granularity**, identifying

distinct stressors and the emotional toll they might have. We looked at these stressors at different levels, enabling deeper insights into how these factors manifest specifically within software engineering.

Third, our findings are **tailored to the software engineering population**, addressing nuances that general workplace studies often overlook. For instance, the critical importance of recognising individual contributions in team-based environments is particularly evident in this domain. Finally, we propose **a set of policy recommendations**, including flexible work structures and peer support networks, that directly address these challenges.

These contributions enhance understanding of well-being in this high-pressure field and enable practitioners and other researchers to develop interventions and support for these topic areas.

Moreover, by systematically measuring various aspects of well-being, policymakers can make more informed decisions that improve overall quality of life, going beyond economic metrics that may not fully capture societal well-being and happiness.

Future work will include a more detailed analysis of country-specific differences. Additionally, we plan to conduct a study with managers on how they currently support software engineers' well-being and the outcomes of these efforts.

Acknowledgement

We thank Ricardo Caldas for helping with the translation to Portuguese. We thank all participants for volunteering their time and personal experiences.

Chapter 3

Paper B:

Emotional Strain and Frustration in LLM Interactions in Software Engineering

C. Martinez, R. Khojah

In the International Conference on Evaluation and Assessment in Software Engineering (EASE), 2025.

Abstract

Large Language Models (LLMs) are increasingly integrated into various daily tasks in Software Engineering, such as coding and requirement elicitation. Despite their various capabilities and constant use, some interactions can lead to unexpected challenges (e.g. hallucinations or verbose answers) and, in turn, cause emotions that develop into frustration. Frustration can negatively impact engineers' productivity and well-being if it escalates into stress and burnout. In this paper, we assess the impact of LLM interactions on software engineers' emotional responses, specifically strains, and identify common causes of frustration when interacting with LLMs at work.

Based on 62 survey responses from software engineers in industry and academia across various companies and universities, we found that a majority of our respondents experience frustrations or other related emotions regardless of the nature of their work. Additionally, our results showed that frustration mainly stemmed from issues with correctness and less critical issues, such as adaptability to context or specific format. While such issues may not cause frustration in general, artefacts that do not follow certain preferences, standards, or best practices can make the output unusable without extensive modification, causing frustration over time. In addition to the frustration triggers, our study offers guidelines to improve the software engineers' experience, aiming to minimise long-term consequences on mental health.

3.1 Introduction

Software Engineering (SE) comes with many challenges, from fixing bugs to dealing with changing requirements. Recently, Large Language Models (LLMs) and LLM-powered chatbots like ChatGPT and GitHub Copilot have been used by software engineers to assist them in performing various tasks, including code generation, and quality assurance [207, 208]. Current research focuses on understanding how practitioners aim to increase their productivity and make their work process more efficient by targeting LLMs to automate the generation of software artefacts or receive guidance on how to solve certain problems [209, 210].

Challenges and limitations of LLMs hinder their effectiveness, such as unhelpful responses, which can lead to frustration among engineers [211]. This frustration contributes to techno-stress, affecting their workflow, well-being, and productivity [42, 134]. While frustrations have gotten little attention in LLM research for software engineering, we argue that understanding the causes of frustrations when using LLMs for software-related tasks is the first step to minimising them and thus improving the productivity and well-being of practitioners in the SE industry and academia. Moreover, revealing such triggers helps the designers of LLM-powered tools (e.g., AI Chatbots) improve the user experience and evolution of such tools.

This exploratory study aims to empirically investigate the causes of frustrations in software engineers' interactions with LLMs (and LLM-powered chatbots) and propose strategies for improvement. We focus on the following research questions:
RQ1: What are the triggers or sources of frustration of software engineers when using LLMs?

Our study presents four main categories that can cause frustrating emotions for the software engineer. We found that the main cause of frustration is when the software engineers receive an unhelpful or incorrect answer, followed by misunderstanding the intention, failing to meet personal preferences, and other limitations of LLMs. We argue that these categories are specific to SE since it is a domain that requires precise, context-aware, and technically accurate responses that can be directly applied, like using LLMs for code generation.
RQ2: How can the frustrating experience impact motivation?

We report that while frustration from unmet expectations can momentarily impact motivation, software engineers typically remain engaged in tasks despite these frustrating interactions with LLMs. This suggests that such interactions are not usually disruptive enough to prevent task completion.
RQ3: How can the user experience be improved to reduce frustration for software engineers?

We provide recommendations for improvements based on software engineers' expectations and lived experiences. Engineers offer practical, grounded recommendations that reflect user needs and expectations since they are the primary stakeholders directly interacting with LLMs in real-world contexts and can guide chatbot designers in enhancing design and usability.

In addition, we suggest managers provide training and raise awareness among software engineers in order to manage and minimise frustrations.

3.2 Background and Related Work

This section presents the conceptual framework of this study, as well as previous work done on the topic and the research gap addressed.

3.2.1 Large Language Models in Software Engineering

In SE, recent research has shown that LLMs have the potential to support practitioners in a variety of tasks, including, but not limited to, implementation, testing, requirements engineering [209, 212–214]. Moreover, despite the challenges LLMs impose on academia (e.g., bias and hallucination), they provide many opportunities for researchers and educators as assistants with creating study guides and academic writing [215].

Despite this extensive use and the time saved when automating tasks, integrating LLMs into SE practices can lead to user frustration, which can be understood as the emotional state experienced by a person when they are prevented or hindered from obtaining something they have been led to expect [216].

Researchers have noted that while LLMs assist software practitioners, they may sometimes generate errors, causing them to spend additional time solving errors or seeking clarification. This can also trigger negative emotions, such as frustration [211].

Our study addresses a gap in the literature by exploring whether the above findings also apply to LLM interactions for SE tasks, with a focus on frustration.

3.2.2 Emotions Involved When Using Technology

In this study, we considered the definition by the American Psychological Association [66], Emotion is *"a complex reaction pattern, involving experiential, behavioural, and physiological elements, by which an individual attempts to deal with a personally significant matter or event"*. We acknowledge the diversity of the conceptual definitions of emotions; we selected this definition for its scope, which integrates experiential, behavioural, and physiological dimensions. The Emotions Wheel, developed by Gloria Willcox [70], is a psychological tool designed to facilitate identifying and expressing emotions. It is widely used in therapeutic settings and personal development to help individuals articulate their emotional states more precisely. The tool organises emotions into a concentric structure, with six core emotions—happy, sad, angry, scared, strong, and calm—situated at the centre of the wheel (See Figure3.1). As one moves outward from the core, these primary emotions subdivide into more specific and nuanced feelings, allowing for a more granular understanding of emotional experiences.

In this study, we used the adaptation done by [217] to explore how individuals experience their interaction with LLMs.

While tools like the Emotions Wheel help categorise and articulate emotional experiences, technology-specific contexts introduce unique emotional challenges. These include emotional strain, which refers to adverse reactions and feelings triggered by stressors [218], and technostrain and its components, particularly relevant for understanding interactions with LLMs.

Salanova [219] defines Technostrain as a negative psychological experience characterised by i) heightened anxiety and fatigue (affective dimension), ii) scepticism (atti-

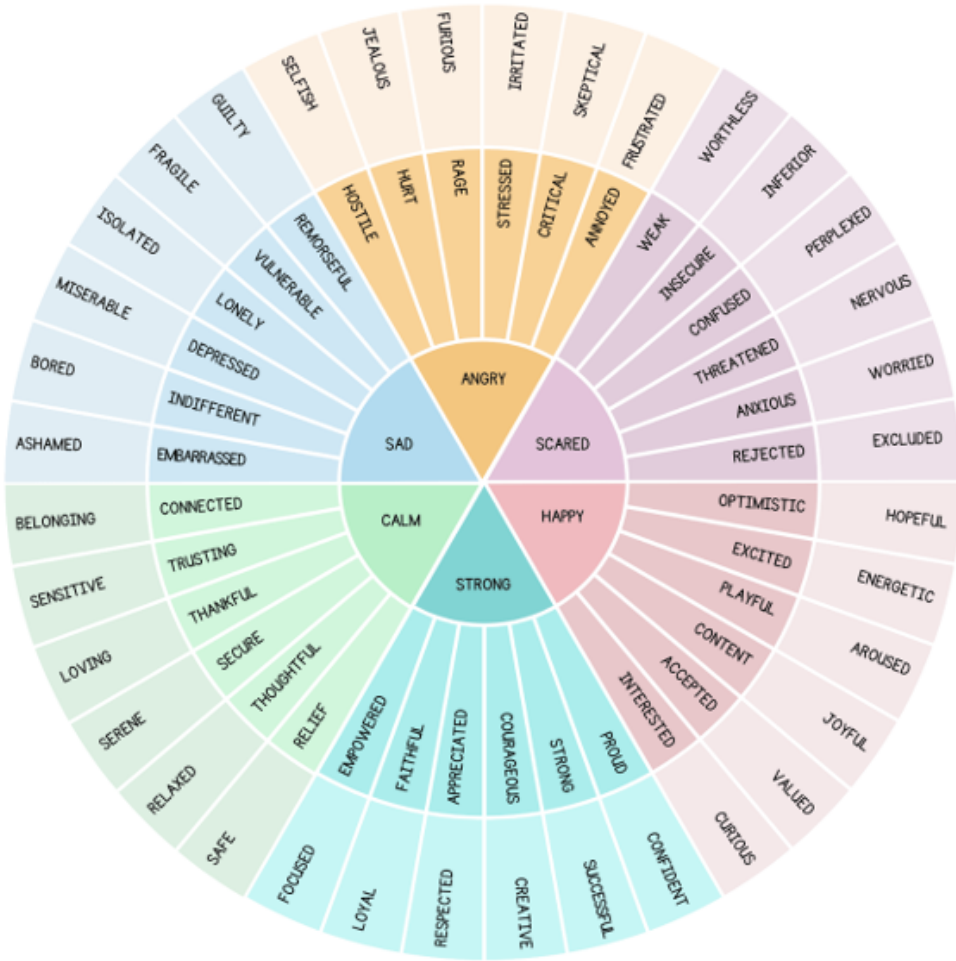


Figure 3.1: Willcox's Emotions Wheel [166]

tudinal dimension), and iii) a sense of inefficacy (cognitive dimension) associated with technology use. Aligned with this, Muller et al. [134] researched techno-frustration, a component of technostrain, which refers to the psychological strain caused by the disorganised or inefficient use of Information and Communications Technology (ICT). Techno-frustration describes the experience of feeling discouraged, uncertain, stressed, confused, and upset as a result of using ICT. Such psychological strain can lead to decreased job satisfaction or an increased risk of burnout [220]. Muller et al.'s work is the closest to our study; however, we focus on a specific interaction, LLM-user interaction, which is still underexplored despite the rapid integration of LLMs in SE.

Regarding research on how technology impacts users' emotions, studies have focused on the psychological effects of prolonged technology use, recognising that digital tools influence productivity and mental and emotional states [221, 222]. Wester et al. [223] found that incorrect outputs, which reject the user's request, can lead to frustration and greatly diminish their perception of the LLM's usefulness, appropriateness, and relevance.

This reflects an effort to explore the balance between the benefits of technology and the potential strain it places on users, particularly in high-demand environments like SE. Examples of those efforts are in the form of interventions [107], [19] or looking for the causes [87].

Furthermore, users' expectations when using technology, particularly LLMs, are crucial to their frustration. Studies have shown that prior experiences with technology can shape users' expectations, influencing their perceptions of performance and trust [224]. When technology does not meet users' expectations, whether based on past interactions or external portrayals, frustration can escalate, hindering effective task completion and satisfaction [220]. By examining these responses, we aim to inform LLM design improvements that enhance productivity and reduce technostress.

3.2.3 Emotions in Software Engineering Tasks

Emotions related to software engineering tasks have been widely researched. Several studies have identified a wide range of emotions, from anger and frustration to joy and satisfaction, as software engineers perform their tasks and communication channels within the development context [225–231]. Sánchez-Gordón's [232] literature review further analysed the diversity of emotions developers face, identifying 40 discrete emotions, the most frequent being anger, fear, disgust, sadness, joy, love, and happiness. Various situational and contextual factors shape this rich diversity of emotions. Additionally, the impact of affective states has been related to performance and code quality. For example, Graziotin et al. [233] found that positive emotions, such as happiness, are closely linked to improved performance and productivity. Conversely, negative emotions, such as frustration and unhappiness, can reduce motivation, hinder task completion, and increase the likelihood of turnover [8]. Furthermore, studies have found that developers report higher productivity when in a state of flow and often experience frustration due to being stuck, technical difficulties or unfulfilled information needs [234, 235].

Moreover, specific triggers of negative emotions have been extensively studied. For example, unhappiness, Graziotin et al. [45] found that everyday sources of unhappiness are time pressure, bad code quality, repetitive tasks, and inadequate

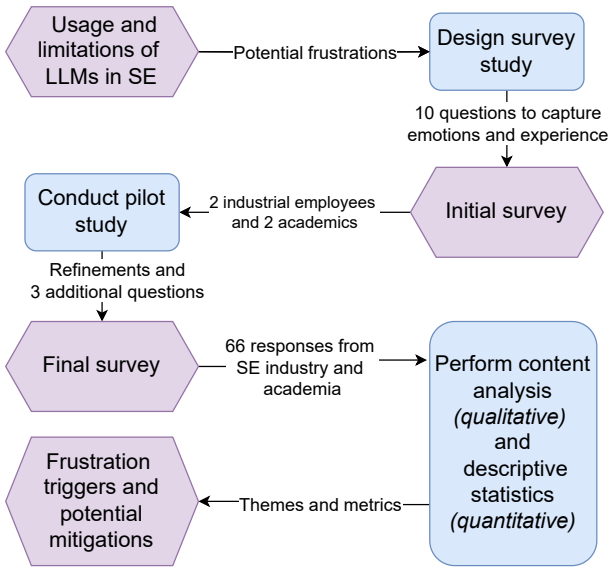


Figure 3.2: The exploratory study process that we followed to design our survey and analyse the responses qualitatively and quantitatively.

decision-making. Regarding frustration, Ford and Parnin [79] identified program comprehension challenges, poor tooling, and fear of failure as common causes.

Despite significant research on emotions associated with traditional software engineering tasks, there is still a gap in investigating the emotional impact of interactions with emerging tools like Large Language Models (LLMs). This gap leaves unexplored how emotions such as frustration evolve in response to LLM use, which has critical implications for improving their design and effectiveness in supporting software engineering tasks.

3.3 Methodology

Our study aims to understand software engineers' frustrations and emotions when interacting with LLMs, identifying causes and potential solutions. This study implements an exploratory design to gather initial insights and detect interaction patterns [236]. We used surveys following Stol and Fitzgerald guidelines [124] since they are particularly effective for exploratory studies aiming to generalise findings across a population. We collected qualitative and quantitative data through open and closed questions, covering specific and broader aspects of LLM usage. Figure 3.2 illustrates our methodology.

3.3.1 Target Population

We surveyed software engineers in academia and industry to get a broader exploration of how LLMs are used in software engineering tasks. While industry engineers provided information about practical applications such as coding, debugging, and meeting production demands, academics, on the other hand, added LLM usage in research-related tasks such as programming, testing and general research on software processes and quality. Despite differences in their environments, both groups share core practices and challenges, such as dealing with tool limitations and managing cognitive load. Including both populations ensured a comprehensive analysis of how LLMs impact software engineering workflows, making our findings broadly relevant to diverse real-world and research contexts.

3.3.2 Data Collection

We used a questionnaire as our main data collection method. We designed the survey questions based on previous research about the usage of LLMs in SE and how it impacts their experience and productivity [209] as well as general limitations of LLMs [237].

After a pilot study with two PhD students (academia) and two software engineers (industry), we added more questions and refined the survey (available in our Zenodo package [238]). The survey started asking for consent to participate and included (i) four demographic questions and queries about participants' experience and familiarity with LLMs, (ii) two open-ended and one closed-ended question about their usage of LLMs during their work and their expectations from these LLMs, (iii) four open-ended questions and one closed-ended question about the participants' emotions when certain expectations are not met and about frustrations specifically that the participants experienced when using LLMs, and (iv) five Likert-scales of the level of importance of different LLMs abilities and aspects to minimise frustrating experiences.

The survey was created on Google Forms and distributed across social platforms (LinkedIn and Facebook). We used stratified random sampling to get software engineers from academia and industry, drawing on Baltes and Ralph's [239] work as a framework for our sampling approach. We sent approximately 20 personal invitations to employees across 10 software organisations of different sizes (Startups and large companies) and domains (e.g., automotive and eLearning). We also invited researchers and academics in SE conferences (RE'24 and FSE'24) to allow our sample of software engineers to be diverse in terms of countries and domains. All data was anonymised, and we did not ask for personal data or identification. We followed our university's ethical regulations and guidelines. We believe that it is important that our sample was diverse in terms of areas and domains, particularly that it included software engineering academics. Software engineering practices often overlap between academia and industry, with some differences in priorities and contexts. This diversity enabled us to capture a broader range of perspectives and demonstrate that academics and practitioners share emotions and challenges related to using LLMs in the field.

3.3.3 Data Analysis

We used content analysis with an inductive approach following the steps by Erlingsson and Brysiewicz [240] to analyse the open questions. Both authors carried out the whole data analysis together systematically. The analysis started by reviewing each answer in detail and discussing it to ensure a shared understanding. This allowed us to identify initial patterns and codes. The codes were then categorised based on their similarities and differences [240], with the categories refined iteratively to ensure accuracy. Themes emerged organically from the data, reflecting the participants' perspectives and providing meaningful insights. We used the emotion classification and feeling wheel by Willcox [70] to categorise and identify the range of emotions expressed by participants. For example, one participant's comment, "I acknowledge that it might give incorrect answers so it is indifferent for me unless it happens often" was coded as Indiferent'. Enabling a deeper understanding of their emotional responses during interactions with LLMs. For the Likert and closed questions, we use descriptive statistics and visualisations.

3.4 Results

This section presents the results of the visualisation of the closed questions and the content analysis of the open questions. We distinguished between academics and practitioners when their results differed, such as in LLM usage. We combined the results when their patterns were similar, like emotional responses or frustration triggers.

3.4.1 Respondents Demographics

Our survey sample included software engineers in diverse roles (see Table 3.1) with a median age of 32. Participants represented organisations from seven countries across three continents, spanning aviation, automotive, game design, infotainment, eLearning, cybersecurity, telecommunications, trade, and SE research and education domains.

Most participants (58 of 62) described themselves as "familiar" or "very familiar" with LLMs. Figure 3.3 shows the range of LLMs they use at work, with ChatGPT being the most popular.

3.4.2 Usage of LLMs in Software Engineering Industry and Academia

The results show that 56 participants (94.9%) use LLMs occasionally, out of which 38 participants (66%) use them on a weekly or daily basis. Only 3 (5.1%) participants indicated that they rarely use LLMs at work. Table 3.2 shows the tasks for which academia and industry respondents apply them. We considered more than one answer per participant.

In **industry**, respondents use LLMs for **programming tasks** like code generation, debugging, and optimisation. They also employ LLMs for **creative and communication tasks**, such as drafting emails and brainstorming ideas, and for

Table 3.1: Participants' areas and roles.

Area	Roles	# Participants	Total
Academia	PhD Student	15	27
	Researcher	7	
	Professor	5	
Industry	Software Developer	10	35
	Software Engineer	6	
	Manager	5	
	AI Engineer	4	
	Researcher	3	
	Tech Lead	2	
	Software Designer	2	
	Software Tester	1	
	Application Specialist	1	
	Applied Scientist	1	

generating and improving text. Additionally, LLMs help users on **learning new technologies and research** by providing starting points, best practices, and summaries of lengthy information. Lastly, respondents view LLMs as digital assistants for **task management and problem-solving**, streamlining workflows and enhancing productivity.

In **academia**, LLMs are primarily used for **writing tasks**, including generating drafts, checking grammar, and providing content suggestions. Users find them helpful for managing busy work, such as email writing and idea generation, and for creating initial drafts for refinement. Additionally, participants view LLMs as **educational tools**, using them to understand new technologies and programming concepts or to assist in teaching. For **programming tasks**, LLMs help write simple code, debug, and learn new coding concepts, offering initial code snippets and quick insights into technologies.

LLMs are also employed for **research-related tasks** such as summarising academic papers, generating ideas, and finding references. They assist in translating data, cleaning datasets, and extracting key information from research.

3.4.3 Emotions During LLM Interaction

Due to the complexity of the emotional responses to unexpected LLM interactions, participants often described multiple, layered feelings in their experiences. We mapped

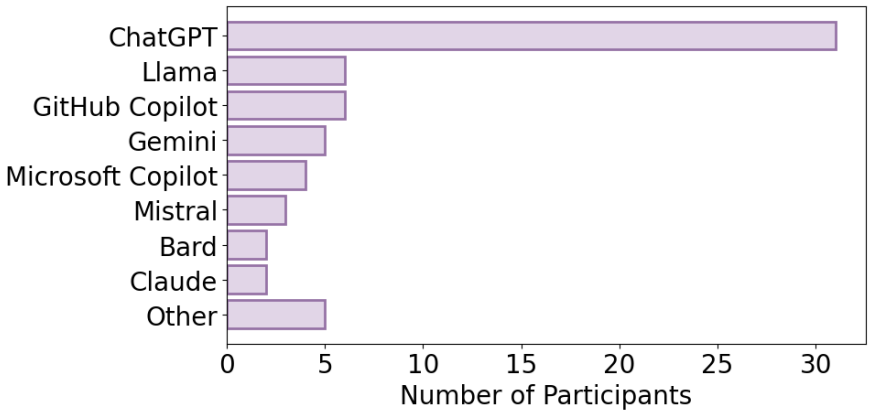


Figure 3.3: LLMs used by our participants.

Table 3.2: LLM Usage Across Industrial and Academic Tasks

Usage Group	**Area**	**Total Count**
Writing Tasks	Academia	13
Programming Tasks	Academia	8
Educational Tasks	Academia	8
Research Related tasks	Academia	5
Non-Critical Tasks	Academia	5
Programming Tasks	Industry	18
Written Communication Tasks	Industry	10
Learning New Concepts	Industry	8
Task Management	Industry	5

these feelings described to Willcox's emotions wheel (see Figure 3.1) to assess on a more fine-grained level how evolved the emotions were. Following this framework, we could trace how initial feelings, such as anger (first-level emotion), might progress into more specific emotions, like annoyance (second level) and then frustration (third level), and quantify how often these emotions occurred at each stage.

In Figure 3.4, we show the different emotions that were expressed by our participants for each category that maps to Willcox's emotions wheel. For instance, we found that the most common emotions (54.6%) are frustration or emotions that can develop into frustration, such as annoyance and anger. This poses potential challenges to well-being and smooth workflows, with a risk of cumulative emotional strain over time.

Many respondents (27.8%) have also reported sadness-related emotions such as disappointment, indifference, or even guilt. In contrast to anger-related emotions, where respondents primarily blamed the LLM for its limitations, those who felt disappointed, sad, or guilty often turned the blame inward and criticised themselves for not being able to write the right prompt or meet their own expectations. When expectations were lower, the disappointment turned into indifference.

> Participant: 29 "Knowing how LLM[s] work, I typically have lower expectations. So I [don't] feel as frustrated or disappointed, particularly if I know that the task I asked is not trivial."

Such expectations come from building knowledge about the LLM and understanding its capabilities and limitations based on previous interactions.

Less frequent reactions included positive emotions like calmness, thoughtfulness, playfulness, and curiosity, as well as negative emotions such as confusion and fear. This shows the varied and sometimes unexpected emotional responses that emerged. These emotions suggest that interacting with LLMs is not just a functional exchange but an exploratory experience for some software engineers. Curiosity, for example, has attached an investigative mindset often aligned with a trial-and-error approach. Playfulness, meanwhile, shows a willingness to engage with the LLM on a more open-ended basis. Fear introduces a new angle, suggesting that some engineers may feel a sense of responsibility if the interaction does not go as expected.

3.4.4 Expectations When Interacting with LLMs

As shown in Table 3.3, participants' expectations for LLMs extend beyond functionality to quality, usability, and versatility, which are important factors in developing effective and trustworthy products in software development and design. Regarding *quality* and *performance*, engineers expect LLMs to consistently deliver correct and unhallucinated information without error or delay, as inconsistency can erode trust in the development workflow.

> Participant: 19 "I rely on the LLM to provide accurate, relevant information that I can trust for both coding tasks and daily life management. It's like having an expert who gets it right the first time!"

In terms of *understanding*, engineers value that LLMs understand context and intent, preferring models that ask clarification questions when necessary. They expect the LLM to understand the context without the need for a detailed context description in the initial prompt. This expectation is important since software research and development happen in dynamic and complex environments that require a lot of context, such as best practices, policies, and relevant software artefacts. With the lack of consideration for such contexts, the outcome can become unusable and hard to integrate into the solution.

> Participant: 19 "I need an LLM that can seamlessly adapt to different contexts. Whether it's helping me with technical jargon, understanding project management lingo, [...], the LLM should be versatile enough to handle it all."

Furthermore, since software engineering is a broad field with researchers from various sub-domains and specialities, researchers need to organise their texts and

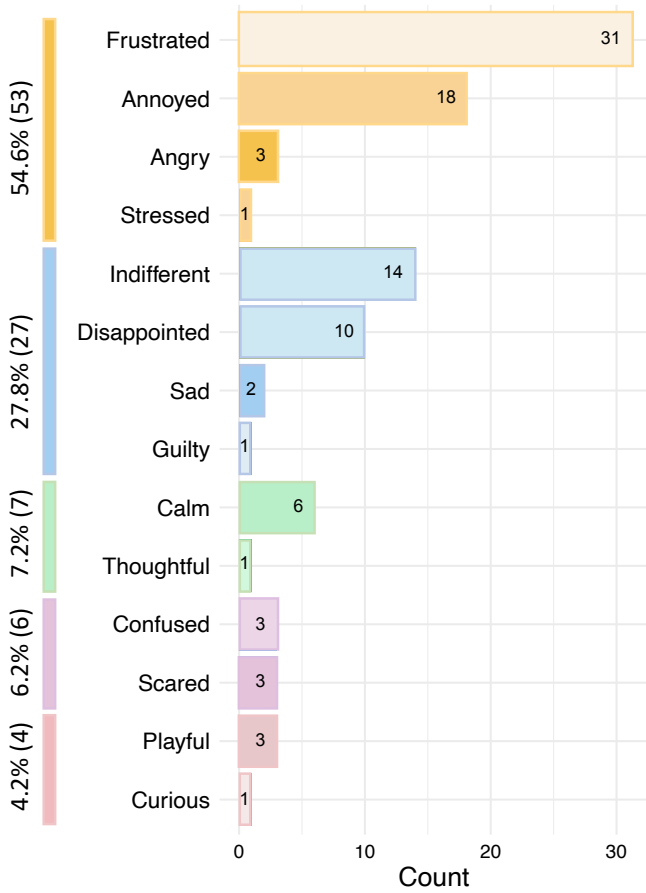


Figure 3.4: Emotional responses when receiving an incorrect answer. The colours map to Willcox's emotions wheel in Figure 3.1.

use language that fits the targeted audience. Therefore, software researchers both in academic and industrial organisations emphasised the importance of clarity and organisation of LLM-generated text, and that tailoring the answer to the *structure* that the task needs is crucial. For code-related tasks, engineers prefer responses starting with code snippets followed by explanations. When using LLMs to learn new concepts or explain artefacts, participants preferred elaborative answers.

Additionally, they expect LLMs to be transparent by providing the source of the information and confidence estimation of the output accuracy to ensure *information integrity* and support decision-making throughout the development lifecycle. This should also come with a need to protect the shared information in the chat.

Finally, in terms of *versatility*, engineers in industry increasingly expect LLMs to integrate with other tools and adapt to diverse workflows, reflecting the growing need

for flexibility in software engineering environments. This also requires a *usable* LLM, ideally with a user-friendly interface that enables intuitive interactions and seamless integration of the LLM in software-related tasks.

Table 3.3: Users' Expectations when Using LLMs

Themes	**Categories**
Quality	Accurate and Correct (39)
	Reliability and Consistency (13)
	No Hallucination (4)
Answer Structure	Conciseness vs. thoroughness (14)
	Complete with Examples (7)
	Organised and Good Grammar (6)
Performance	Efficiency (response time) (20)
Information Integrity	Transparency (10)
	Up-to-Date Information (4)
	Data Security and Confidentiality (5)
Understanding	Intent Understanding (10)
	Domain/Contextual Understanding (2)
Usability	Ease of Use (5)
Versatility	Integration with Other Tools (1)
	Adaptability to Workflow (1)
	Adaptability to Communication (2)

3.4.5 Frustration Triggers in Software Engineering

After exploring emotions in general related to receiving unexpected answers from LLMs, we focus on frustration-related emotions. We asked participants to describe specific situations where they felt frustrated when interacting with LLMs. From this, we identified patterns of triggers that cause LLMs to fail to meet engineers' expectations, leading to strains such as frustration. We outline the frustration triggers below.

Repeated inaccuracies and hallucinations: One of the most common causes of frustration was receiving repeated incorrect or hallucinated responses from the LLM.

The definition of incorrectness varied among participants. Some examples of incorrect answers were uncompilable or buggy code, incorrect explanations of error messages, or incorrect factual information that was verified using other sources (e.g., an expert or a search engine). Hallucination was described as nonsense explanations, references that do not exist, made-up packages, and invalid syntax in a programming language.

Participants explicitly described that frustration arose when these issues persisted despite attempts to rephrase prompts or correct the LLM. For instance, participants referred to such situations as "annoying" or "disappointing" initially. However, they noted that *repeated failures* led to frustration, describing the LLM as "stubborn" and "insisting on an incorrect or hallucinated answer".

> Participant: 3 "After several corrections, and repeating the prompt in different manners, it decided to reiterate the same wrong response."

This pattern of repeated failures can disrupt workflows in software engineering, where development cycles are often fast-paced and agile, which requires more reliable and stable tools. For example, when the LLM provides a code that imports hallucinated libraries, it renders the code unusable, which leads to wasted time fixing, debugging, or rewriting the entire implementation.

> Participant: 38 "During a coding problem, I was looking for the usage of a specific function in a library. I was frustrated when it provided a different [function] (which did not work or even exist)."

Intent not understood: Frustrations (or related emotions) are also caused when the software engineers feel that the LLM did not understand their prompt. Not understanding can be reflected in a response that is irrelevant to the initial question. Intent understanding was a common frustration trigger among our participants from industry and academia, since in practice, engineers' queries are often highly technical and domain-specific and deal with complex software artefacts. Similarly, researchers and academics deal with novel techniques and niche problems that may cause the LLM to misunderstand the intent. Note that such misunderstandings are more common in general-purpose LLMs than in fine-tuned and specialised LLMs.

> Participant: 29 "I was trying to ask [LLM] to fill one specific cell in a notebook based on the others but it kept returning the same generic code for two cells instead of one. I had to talk to [LLM] like a child and say don't do that and do this, and only this."

On another note, some participants pointed out their perceived usefulness of prompt programming and carefully constructing a prompt that would minimise such misunderstandings that can be caused by poor phrasing or the lack of context in the prompt.

> Participant: 5 "I provide short and maybe unclear prompts [then] I usually get irrelevant responses. The better the prompt, the better response."

While prompt programming has shown a high potential in enhancing the LLM outcome, it remains unclear whether it is effective in software engineering-related

tasks.

Personal preferences unmet: Many of our participants pointed out that they get frustrated with reasons related to their personal preferences. Some aspects of certain LLMs (e.g. answer structure) can be annoying to some engineers, and when the frequency of interactions with the LLM increases, the annoyance turns into frustration. For example, two participants pointed out that an LLM apologising every time they tried to correct the LLM was a source of frustration since it can disrupt the workflow, especially during a refactoring or debugging process with the LLM, which results in a long conversation with many follow-up prompts. Others were frustrated with how the LLM they use only provides answers that are long or only in bullet points. Forcing the LLM to structure an answer that aligns with their preferences required specifying many requirements and constraints in the prompt.

> Participant: 35 "[I get frustrated] when [LLM] gives too long answers. I quite often ask things that can be answered with a short sentence, but still I get half a page of ramblings back."

Such preferences depend on the task that the software engineer is solving. For example, important details when debugging might be hidden in long responses, while overly concise bullet points might omit crucial information needed to understand a system's architecture.

Limitations of the LLM: LLM limitations (e.g., inability to perform specific tasks) or configuration constraints (e.g., context window size) are frustration triggers for software engineers. When an engineer attempts to force the LLM to overcome these limitations by prompting, it often leads to more frustration. For example, trying to generate a large application code in one prompt can result in missing lines and errors when the LLM hits its maximum token limit [241]. The participants highlighted that there are some of the many tasks in software engineering research and practice that LLMs are just "not good at".

> Participant: 56 "[I got frustrated when] formatting of a table in latex, [had to] move to another LLM"

Finally, as a verification question, we asked the participants to rate aspects of LLMs (correctness, lack of hallucinations, understanding, performance, and ability to answer) on a 5-point Likert scale based on how important they are in an LLM in order to ensure a better user experience. The results in Figure 3.5 resonate with the previous results, where correctness, lack of hallucination, and correctness are the most important. In comparison, performance (e.g., response time) and ability to answer were seen as less critical.

3.4.6 Unmet Expectations' Impact on Motivation

When the LLM failed to assist our participants, the participants' motivation to complete their task was influenced mainly in three ways.

21.3% (13 out of 62) of our respondents reported that their **motivation decreased** when LLMs did not give the correct answer. Responses expressed frustration, stress, or disappointment, impacting participants' willingness to continue. Other participants commented that they eventually gave up. These respondents mentioned that after

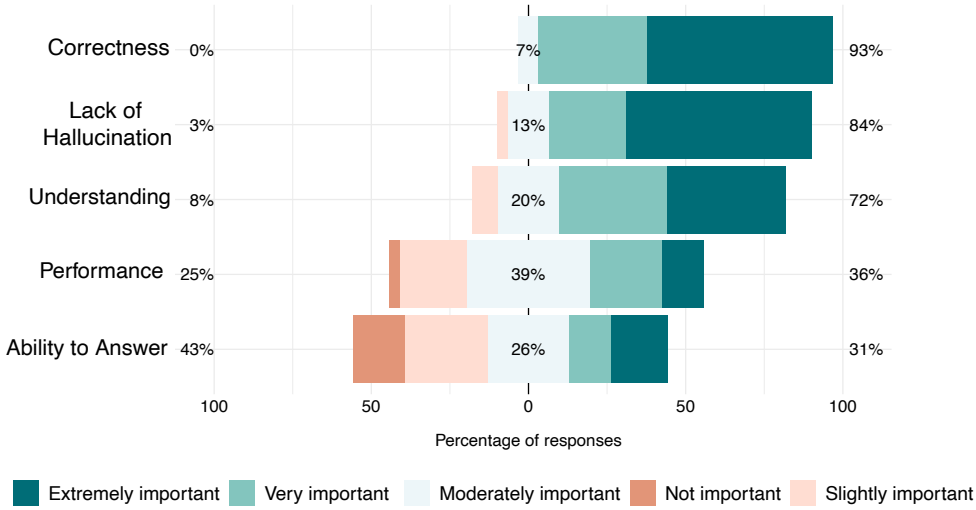


Figure 3.5: Likert-scale results of the importance level of different aspects of the LLMs that can impact the user experience. The scale ranges from Not Important (left - red) to Extremely Important (right - green).

some effort, they decided to abandon the LLM and move on to other methods or stop altogether.

> Participant: 29 "Eventually I give up [on the task], or report negative experimental results."

Another group was formed of 77% (48 out of 62) of the participants who were unfazed by the LLM's failure, treating it as a non-critical tool and continuing with the task with their **motivation not being affected**.

> Participant 8: "My motivation is not affected, I just realise that the task will take longer."

In an interesting case, one participant mentioned that their **motivation increased** which was due to perceiving the interaction as a learning opportunity.

> Participant 16: "I usually understand the problem a lot more so I want to complete the task more"

3.4.7 User Actions for Improving LLM Interactions

When asked what actions participants typically take after receiving an unexpected answer from an LLM, the majority (41 out of 62 participants) said they changed the prompt to try again. A smaller group (12 participants) reported providing feedback to the LLM, while a few (2 participants) said they did nothing (see Figure 3.6). In the

"Other" category, participants described various strategies. Some combine multiple actions, such as changing the prompt, switching to another LLM, or even reverting to traditional search methods like Google or Wikipedia. A few participants mentioned disengaging from the LLM entirely or constructing the solution themselves.

> Participant: 39 "Sometimes the LLM hallucinates and puts me in loops. When I realise this, I resolve it myself using my human knowledge and [X] years of experience in the development field"

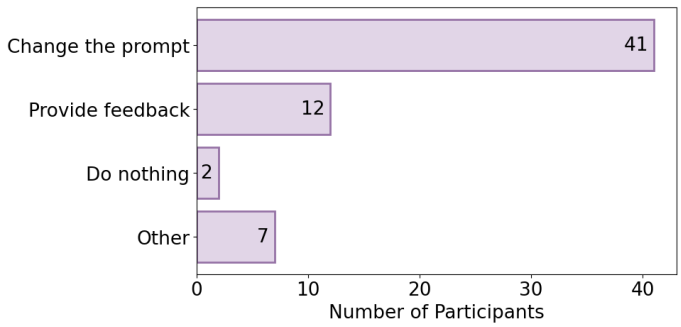


Figure 3.6: Actions done after receiving an unexpected answer from LLMs

We asked our participants about the improvements they would like to see that would enhance their experience when using LLMs. They identified several areas that we grouped and sorted in Table 3.4 based on the number of times they were mentioned. The suggestions mainly align with the engineers' expectations (See Table 3.3), though with a variation in the emphasis and distribution.

Regarding the *quality* of the outcome, the participants emphasised the importance of reducing hallucinations and improving response accuracy. They also commented on the need for enhanced analytical capacity to tackle more complex problems. However, the participants highlighted that while they expect an LLM to be accurate, transparency is more crucial for a less frustrating user experience. The transparency was described by the participants concerned **information integrity**, particularly specifying the data sources and the confidence level of the LLM's responses. This also includes stating any assumptions and reasoning by the LLM before providing an answer. This helps practitioners decide whether to rely on the LLM's suggestions, trust their own judgment, or seek assistance from a colleague instead. Establishing trust in this context is crucial, as it determines the main flow of the interaction, and hence how the LLM output will be eventually used by the software engineer [242].

Regarding the *answer structure*, participants preferred concise and short answers (unless they were prompted otherwise). While in terms of *understanding*, prompt comprehension was considered essential, with users expecting LLMs to ask clarifying questions when necessary. However, some participants mentioned that this could be a *human-related* improvement where training that is designed specifically to learn to "talk" with the LLM is important to make communication more effective.

Finally, *versatility* was another improvement that was consistently mentioned among our participants; they indicated the importance of handling more complex tasks

and adapting to different user needs. The increased functionality to support a broader range of tasks was also crucial for improving overall user satisfaction. This is especially critical as software engineers, researchers, and practitioners often work with a variety of tasks that range from learning new concepts and tasks to implementing products and using tools. Furthermore, practitioners particularly stressed the importance of LLM *performance* in terms of increased memory and learning capabilities, allowing LLMs to retain relevant information and build on past interactions with a faster processing time.

Overall, these requirements guide chatbot designers to understand practitioners' and researchers' needs and priorities in software engineering. For instance, the next-generation LLMs may need to direct efforts toward transparency rather than performance when aiming for a better user experience.

Table 3.4: Improvements for Better User Experience

Themes	**Categories**
Information Integrity	Transparency (source, confidence) (16)
	Trust and Data Security (4)
	Relevant information (Up-to-date) (2)
	State reasoning and assumptions (2)
Versatility	Adaptability in communication (11)
	Integration in project environment (2)
	Extended Functionality (2)
Quality	Improved Response Accuracy (6)
	Reduced Hallucinations (4)
	Less creative (3)
Understanding	Context understanding (5)
	Clarification questions (3)
	Intent Understanding (2)
Answer Structure	Elaborative answers on-demand (4)
	Consistency in Responses (3)
Performance	Higher memory utilisation (4)
	Efficiency (processing time) (1)
Human-related	Training for engineers (2)

3.5 Discussion

In this section, we answer our research questions, highlighting the main takeaway per question.

3.5.1 Frustrations in the Context of LLM Interaction

We identified four main frustration triggers for software engineers using LLMs, accuracy issues, hallucinations, misunderstandings, unmet preferences, and LLM limitations, pointing to significant challenges with potential long-term repercussions. For example, repeatedly dealing with inaccurate, buggy or non-standard code disrupts workflow, forcing engineers to spend extra time debugging [243] or reworking tasks that an efficient tool should minimise. This is particularly important if the code produced by the LLM is less maintainable [244] and can be more frustrating than writing the code with no LLM assistance [211]. This extra workload impacts project timelines and creates a deeper frustration that could enforce frustration and emotional strain [245, 246] as engineers may feel that they have spent longer than planned to reach their goal (e.g., completing a task). Additionally, participants' frustration resulted from other emotions, including anger and annoyance (see Figure 4).

On Wilcox's emotion wheel, these frustrations can be linked to various emotions within the frustration spectrum, including anger, annoyance, and confusion. For instance, frustration over inaccurate or faulty code can easily evolve into anger when engineers feel a lack of control over the situation, especially if the tool is supposed to improve efficiency. Similarly, when an LLM produces outputs that deviate from expectations, it can lead to another emotion within the spectrum, for example, annoyance, particularly when the tool fails to meet personal preferences or the engineer's standards. In addition, misunderstandings or incorrect outputs might lead to confusion as engineers try to reconcile the LLM's output with their original intent.

Studies on GitHub Copilot [247, 247] showed similar emotions, especially around data privacy concerns, intrusive code suggestions and usability, particularly unnecessary large code suggestions. Eshraghian et al. [248] explained that frustration and anger can come from feeling a threat without being able to control it. In the previous example, the threat of leaking confidential data with very little control over it (i.e., to use the LLM, they need to accept the policy) was the trigger for frustration.

Unlike other domains where frustration often stems from performance issues (e.g., system crashes) [246], our participants did not report such frustrations, likely due to recent LLMs being stable and fast (Table 3.3). Other frustrations in the medical domain arise from the emotionally exhausting work environment along with their dependence on the technology (e.g., to document patient data) [249, 250]. In contrast, software engineers can still rely on their expertise or alternative tools, as LLMs are not essential for task completion (Section 3.4.6).

Takeaway: Frustration triggers studied in software engineer literature (including ours) come from spending extra time refining output, unlike in existing studies in other domains, where it is due to performance and usability issues.

3.5.2 Impact of Frustrating Experiences on Motivation

While most of our participants expressed frustration (or similar emotions), their motivation to complete the task was not necessarily affected. This can suggest that although frustration may reside in the emotion wheel's 'anger' or 'irritation' sections, the engineers' resilience and coping mechanisms allowed them to manage these emotions without diminishing motivation.

Our results showed that participants often felt demotivated when an LLM failed to meet their expectations, such as when it did not assist them as intended, frequently leading to frustration. However, most of those who felt frustrated reported that their motivation remained intact, likely due to perceiving the LLM as merely one tool among others in achieving their goal. When an LLM could not provide the necessary assistance, participants commonly resorted to alternative solutions, such as using search engines like Google or relying on their expertise to complete the task independently. These observations align with findings by Franca et al. [251], who explored the connection between motivation and the satisfaction and happiness of software engineers, revealing that happiness slightly overlaps with but does not correlate with motivation.

These findings suggest that, despite facing emotional challenges, software engineers maintain their motivation to persist with demanding tasks. Although frustration may not directly impact motivation, it remains a critical factor to consider due to its known connection with burnout. Sustained frustration is still significant as it contributes to emotional strain [250], a known precursor to burnout in high-demand professions like SE. Thus, understanding and managing frustration, even when the motivation appears unaffected, is essential in supporting the long-term well-being of software engineers.

Takeaway: Although frustration occurs when LLMs fail to meet expectations, it generally does not diminish the motivation to complete tasks.

3.5.3 Towards a Less-Frustrating User Experience

When designing chatbots and LLMs, it is essential to prioritise not only high accuracy and performance but also emotional intelligence, such as recognising user frustration. Wilcox's emotion wheel provides a nuanced view of emotional states, categorising emotions into primary and secondary feelings. Recognising and responding to these emotions in real-time is key.

The emotional intelligence of LLMs has been explored by Wang et al. [252] where they saw that newer-generation LLMs such as GPT-4 (at the time of the study) show a better ability to understand user emotions that compared to humans' emotional intelligence. However, recognising emotions is insufficient as the LLM should also know how to act accordingly. In Section 3.4.5, we saw that even minor LLM behaviours such as apologising after receiving feedback were making the engineers' experience more frustrating. This was also described by Erlenhov et al [253] about ideal development bots adapting communication to different individuals. While several studies focus on enhancing LLMs, we recognise that these systems will always have room for improvement. Therefore, our focus in this study shifts to the human

element. We propose enabling users with the knowledge and skills to navigate LLMs effectively to reduce frustration and improve overall user experience. The goal is to address frustration triggers that arise during their use. Specifically, triggers such as misunderstanding of the intent, unsatisfying personal preferences, or even getting incorrect answers can be minimised by prompt engineering the query before sending it to the LLM. Prompt engineering can involve incorporating prompt techniques (such as Few-shot learning), relevant contextual information (e.g., system description), or constraints about the output (e.g., the output structure). Other frustration triggers, such as hallucinations and limitations of LLMs, can be minimised if the engineers use the LLM according to its capabilities and limitations. Since unmet expectations are among the primary frustration triggers [254] (see Section 3.4.4), providing software engineers with training on effective usage and a clear understanding of LLM capabilities to set realistic expectations can help reduce disappointment and enhance user satisfaction.

Additionally, raising awareness about potential frustration triggers is important; engineers can manage their reactions accordingly if they recognise the likelihood of frustration in certain situations. For instance, using coping strategies rather than repeatedly attempting to elicit a perfect answer from the LLM. Therefore, we suggest that software engineers need training on how to use LLMs safely. This idea was also discussed by Barman et al. [255] where they propose providing guidelines for LLM users to know how to interact with different LLMs, for instance, if it is appropriate to generate artefacts or only to get some guidance. Our participants commented that they were familiar with LLMs; however, most reported frustration, raising questions about whether they truly understand how to leverage LLMs effectively. Familiarity does not necessarily equate to proficiency [256], stressing the need for improved training and guidance on optimal usage strategies. A complete understanding of LLM capabilities and limitations can help users to manage their expectations.

Takeaway: A less frustrating experience arises from combining "emotionally intelligent" LLMs with engineers' awareness to manage their expectations and reactions to frustration triggers.

3.5.4 Threats to Validity

In this section, we explain the strategies to address this study's threats to validity.

Internal Validity: To ensure internal validity, we considered several biases and employed mitigations accordingly. Self-selection bias: Participation in the survey was voluntary; hence, individuals with particularly strong positive or negative experiences with LLMs might be overrepresented, skewing the data. To mitigate this, we tried to recruit diverse participants across different experience levels, regions, and fields. To address social desirability bias, we collected anonymous data by including a statement at the beginning of the survey and avoiding personal questions. This approach aimed to prevent participants from feeling pressured to align their responses with what they perceived as socially or professionally acceptable. This could lead to underreporting frustration to appear more competent with new technologies.

External Validity: Our sample size of software practitioners and academics can limit to extent to which our findings can be generalized to the broader population of

software engineers, to minimize this threat while avoiding overrepresenting certain groups or regions, we targeted respondents from different countries and several fields within SE. Similarly, we employed stratified sampling to ensure a balanced representation across demographics, skill levels, and industries.

Construct Validity: We operationalised key concepts like frustration and hallucination to guarantee construct validity, adding their definitions. Additionally, we provided examples throughout the survey to clarify the scenarios we were exploring. Finally, we explained the Likert scale by adding information on how to measure each level. Further, we piloted the survey to ensure clarity. We used the feedback to fix ambiguous questions, clarify terms, and minimise confusion.

3.6 Conclusion

This study focused on the emotional strains, particularly frustration, experienced by software engineers when interacting with LLMs and not being assisted as intended. By identifying main triggers, such as the correctness and reliability of responses and issues related to personalisation, we emphasise that understanding the emotional impacts of LLM use in SE is important. This study's insights bring attention to the potential risks to productivity and mental health if emotional responses go unaddressed. Future research should further explore the psychological implications of LLM use, focusing on sustainable strategies to support the well-being of software engineers and optimise their user experience with AI tools.

Chapter 4

Paper C:

Take a deep breath: Benefits of neuroplasticity practices for software developers and computer workers in a family of experiments

B. Penzenstadler, R. Torkar, C. Martinez

International Journal Empirical Software Engineering (EMSE), 2022.

Abstract

Context. Computer workers in general, and software developers specifically, are under a high amount of stress due to continuous deadlines and, often, over-commitment.

Objective. This study investigates the effects of a neuroplasticity practice, a specific breathing practice, on the attention awareness, well-being, perceived productivity, and self-efficacy of computer workers.

Method. The intervention was a 12-week program with a weekly live session that included a talk on a well-being topic and a facilitated group breathing session. During the intervention period, we solicited one daily journal note and one weekly well-being rating. We created a questionnaire mainly from existing, validated scales as entry and exit survey for data points for comparison before and after the intervention. We replicated the intervention in a similarly structured 8-week program. The data was analyzed using Bayesian multi-level models for the quantitative part and thematic analysis for the qualitative part.

Results. The intervention showed improvements in participants' experienced inner states despite an ongoing pandemic and intense outer circumstances for most. Over the course of the study, we found an improvement in the participants' ratings of how often they found themselves in good spirits as well as in a calm and relaxed state. We also aggregate a large number of deep inner reflections and growth processes that may not have surfaced for the participants without deliberate engagement in such a program.

Conclusion. The data indicates usefulness and effectiveness of an intervention for computer workers in terms of increasing well-being and resilience. Everyone needs a way to deliberately relax, unplug, and recover. A breathing practice is a simple way to do so, and the results call for establishing a larger body of work to make this common practice.

4.1 Introduction

Physical, mental, and emotional resilience is necessary for taking good decisions under pressure, staying healthy, and experiencing a good quality of life [257]. Resilience is specifically relevant for software engineers in comparison to other knowledge workers as 1) they develop some of the most complex systems in the world where they have to combine computational thinking (e.g., intellectually taxing divide and conquer) [258] with systems thinking (e.g., context dynamics and side effects) [259], 2) they tend to work either in isolation or in intense team environments [260], often globally distributed [261], and 3) they need empathy and relational skills to communicate effectively with colleagues and clients[1], and to put themselves into the shoes of future users and build shared meaning [263]. Empathy for connecting with others [264] requires self-awareness (which has been linked to productivity [265]), and may be at risk due to increased connection via technology [266].

Yet, software developers (and other computer workers) tend to live high-paced work lives and this comes with long-term consequences for their health [143] and happiness [8]. Sleep deprivation is often worn as a badge of honor [44]. In addition, the pandemic has taken a toll on well-being and productivity [185]. For example, people tend to experience lower motivation, productivity and commitment while working from home in a disaster situation [267].

Our survey on 'Healthy Habits in Software Engineering' attached to an IEEE Blog article [268] showed that the number one method named by respondents to counteract perceived stress was physical activity, whether as workout or team-sport or recreational activity. While movement is a valuable way to decrease perceived stress and to relax the body, there are two limitations to it: first, not everybody may be able to incorporate physical exercise into their routine due to bodily limitations, and second, restrictions due to recent developments prevented most team sports for extended periods of time. Consequently, alternative or additional practices to take care of mental and emotional well-being promise to significantly enhance the overall perception of well-being.

In this article, we explore the use of neuroplasticity practices, more specifically, the use of a breathing practice, in terms of its benefits for software developers and computer workers. When looking at the general list of practices to enhance well-being and resilience and to relax and release stress from the body, these can be categorized into (i) movement practices (e.g. Yoga Asana, Tai Chi, Qi Gong), (ii) mental practices (e.g. Meditation, Contemplation), and (iii) breathing practices (e.g. Wim Hof, Pranayama, Rebirthing Breathwork, Holotropic Breathwork). The listed movement practices are fairly accessible in terms of physical capabilities compared to other more athletic forms of movement. They may still require four functioning limbs, the confidence to practice them with a group, or an environment that ensures sufficient feedback mechanisms to make sure the exercises are carried out correctly. The referenced mental practices offer simple beginner methods and techniques — however, simple does not equate to easy in this case. Consequently, people who already struggle with stress and/or anxiety and worries may be overwhelmed with the request for determination to sit down quietly and calm their mind. Therefore, we chose a breathing practice as mode of intervention for our investigation. Specifically,

[1] as shown in the medical field [262]

one that was not physically or mentally challenging for the aforementioned reasons.

To this end, we designed an intervention with a weekly live group practice that was held online and followed up by the reflective practice of journaling. In addition, we used surveys for collecting quantitative data.

Research Questions: We answer three main research questions on what changes are observable over the course of the breathwork program in the participants' (1) mindfulness attention awareness and daily perceptions of life, (2) well-being, and (3) productivity and self-efficacy.

Contribution: We provide the first empirical study of the "Pranayama Vyana Vayu" breathing practice, which is also the first empirical study on breathwork within engineering. The results indicate usefulness and effectiveness as intervention for computer workers to increase well-being and resilience.

In Section 4.2, we explain the background of the study and related work. Section 4.3 provides details about the intervention, research questions, applied method and research design. Section 4.4 describes the analysis. Section 4.5 presents the results. Section 4.6 explores these in discussion and Section 4.7 concludes with a summary and an outlook. Section 4.8 points to the open data archive where the scripts and data have been made available. The appendices include the survey instruments (App. A.2), model designs (App. A.3), and detailed findings (App. A.4).

4.2 Background

In this section, we present the context of stress in software development, the background on breathing practices, the theory of mindfulness and the related work on well-being and resilience in software development and IT work.
As we introduce a number of concepts relevant to this study, we provide an overview of the most important terms summarized in a table at the end of this section.

4.2.1 Context: Stress in Software Development and IT Work

Stress factors have a negative influence on cognitive task performance [269], and lead to burnout [270, 271]. Contributing to that, over-scheduling and double-booking have been signs of progress and belonging for two decades. Progress equals fast, and fast equals success, which is a recipe for addiction [272]. In addition to the known effects of stress on software quality as evaluated by Akula and Cusick [273], Amin et al. [274] found that occupational stress also negatively affects knowledge sharing, which leads to long-term detrimental effects for software systems development, particularly in global development settings.

Fucci et al. looked into the effects of all-nighters for software developers, as they are often willing to work late for project deadlines because "forgoing sleep appears to be a badge of honor in the programmers and start-up communities" [44]. Sleep deprivation and disrupted circadian rhythms may lead to adverse metabolic consequences [275], all the way up to increasing the risk for developing cancer [276]. The effects of sleep deprivation are clearly negative, and stressed software engineers report a decreased quality of sleep [273], which in turn negatively impacts health [277]. This also leads to economic losses, recognized in the US, but also United Kingdom, Japan, Germany,

and Canada [277], estimated to between $280 billion and $411 billion for the US in 2020, depending on the scenario, and between $88 billion and $138 billion for Japan. Consequently, the potential benefits of the practice evaluated in the study at hand could help improve quality of sleep and decrease the respective physical and mental health consequences. Lavallée and Robillard [278] found in a ten month study that many decisions made under the pressure of certain organizational factors negatively affected software quality, which further motivates our goal to increase stress resilience for people in this line of work.

Ostberg et al. [143] show a methodology of how to physiologically evaluate the stress that software developers are under, so that interventions against the stress and its long-term consequences for health can be empirically measured. We are using their self-efficacy instrument in the study at hand.

4.2.2 Background: Breathing practices

Origins: Yogic Pranayama. The origins of traditional breathing practices, also known as pranayama, are found in the Vedic scriptures that date back to [279]. The "Vedas" are regarded as the world's oldest piece of literature. They are the basis of Ayurveda (Science of life) and Yoga (Union of body, mind and spirit). Yoga made its way into the West during the 20th century, with a huge increase in popularity first in the sixties and seventies and then over the last two decades. One of the ways yoga is practiced is pranayama, composed by the two words 'prana' (life force) and 'ayama' (that which animates). So what gets somewhat lazily translated as breathing practices are energy practices that serve to increase and adapt the energy flow in a practicing individual.

Breathwork. The term breathwork has been used as a synonym in the West by a wide variety of teachers, so we briefly introduce it at this point. Specifically the adaptation of energy enhancement and balancing via means of breathing practices has been popularized outside of its yogic origins in the West since the 1960s, by researchers and teachers like Stanislav Grof and Leonard Orr. Stanislav Grof discovered that a specific breathing pattern that he named Holotropic Breathwork produced similar effects as the consumption of LSD (after he had accidentally discovered LSD while testing drugs in the lab for a pharma company, and the substance was proclaimed illegal later on). Leonard Orr coined the Rebirthing Breath after going through an awakening experience in a sensory deprivation floating tank. Both patterns work with circular breathing, which means there are no breaks in between in-breath and out-breath, which can lead to a strong energetic stimulation of the body that can trigger cognitive and emotional experiences. There are many other forms and other accomplished and experienced facilitators like David Elliot [280] and Dan Brule [281] who have studied in depth, guided thousands of participants, and pass the knowledge on globally. In the article at hand, we build on the lineage passed down by the Breath Center [2] where the first author become a certified practitioner in 2019.

[2] `http://www.thebreathcenter.com`

The Neuroscience and Empirical Benefits of Breathing Practices. The vagus nerve is the largest cranial nerve in our body and a vital player for the parasympathetic nervous system [282]. Our nervous system is in a fight-or-flight response during stress (of any kind, be it physical, mental, or emotional) and to recover more quickly from stress it is vital to tone the vagus nerve [32]. When we experience stress, the kidneys release adrenaline, which gets transported to the brain via the vagus nerve, where it gets compared and stored to memory. The breathing practice used in the intervention of the article at hand gently resets the nervous system and thereby provides the grounds for responding to life from a resourced place as opposed to fight-or-flight.

Several studies have been carried out related to the breathing practice of Sudarshan Kriya trained by The Art of Living[3], e.g., Seppala et al. [283] address the decline in mental health on U.S. university campuses by examining the effects of three interventions: Sudarshan Kriya breathing ("SKY"; N = 29), Foundations of Emotional Intelligence ("EI"; N = 21) or Mindfulness-Based Stress Reduction ("MBSR"; N = 34), with SKY showing the greatest impact, benefiting six outcomes: depression, stress, mental health, mindfulness, positive affect and social connectedness. Sharma et al. [284] explored the topic of the same SKY breathing practice and found positive immunological, biochemical, and physiological effects on health (N = 42). Walker and Pacik [285] showed a reduction of Post-Traumatic Stress Disorder in Military Veterans (three cases). Brown et al. [286] report it to be used successfully in the treatment of stress, anxiety, and depression.

The family of experiments at hand provides a first empirical evaluation in a related technique and thereby serves as comparative data point as well as confirmatory research for the benefits of breathing practices in general.

4.2.3 Theory: Mindfulness and Mindfulness Attention Awareness

William James (1911/1924), who studied consciousness, was not sanguine about the usual state of consciousness of the average person, stating, "Compared to what we ought to be, we are only half awake" [287, p. 237]). Based on this and according to [140], mindfulness is inherently a state of consciousness. A direct route through which mindfulness may enhance well-being is its association with higher quality or optimal moment-to-moment experiences [140].

Mindfulness-based practices have been studied and evaluated in research since 1979 by Kabat-Zinn [50], who developed a clinical 8-week Mindfulness-based Stress Reduction (MBSR) program that has been successfully replicated all over the world, including in correctional facilities [288]. Kabat-Zinn clarifies that meditation is a direct and very convenient way to cultivate greater intimacy with your own life unfolding and with your innate capacity to be aware [289]. The objective was to offer an environment with methods for facing, exploring, and relieving suffering at the levels of both body and mind, and understanding the potential power inherent in the mind-body connection itself in doing so [50]. Clinically proven results include positive affect with regard to emotionally stressful situations as well as increased immune system response [50].

[3] `https://www.artofliving.org`

Westen [290] defined that consciousness encompasses both awareness and attention. *Awareness* is considered the background "radar" of consciousness, continually monitoring the inner and outer environment, and one can be aware of stimuli without them being at the center of attention. *Attention* is defined as a process of focusing conscious awareness, providing heightened sensitivity to a limited range of experience.

Thus, *mindfulness attention awareness*, defined as a concept by Brown et al. as "present-centered attention–awareness" [140, p. 824] - which can be explained as being conscious of being mindful - plays a broad and important role in self-regulation and emotional experience, which also impacts work and productivity. Therefore, we included the standard (validated) instrument of the Mindfulness Attention Awareness Scale [140] into the study at hand. The MAAS was developed to examine empirical links between mindfulness and well-being, and is focused on the presence or absence of attention to and awareness of what is occurring in the present rather than on attributes such as acceptance, trust, empathy, or gratitude [140].

By evaluating the change in mindfulness attention awareness is one of the variables looked at, we investigate whether the used breathing practices contribute to a potential improvement.

4.2.4 Related work: Well-being and Resilience for Engineers, Software Developers and IT Workers

Bernardez et al. [291,292] performed experiments showing that the practice of mindfulness significantly improves conceptual modeling efficiency and improves effectiveness. The authors pointed out that specifically introverts may benefit, and the software field is dominated by introverts [293].

Graziotin et al. [8] investigate the happiness of developers and found consequences of unhappiness that are detrimental for developers' mental well-being, the software development process, and the produced artifacts. They use the SPANE instrument [141] to measure differences in the perception of positive and negative affect in experiential episodes, which is also used in the study at hand.

Rieken et al. [294] explore the relationship between mindfulness, divergent thinking, and innovation, specifically among engineering students and recent engineering graduates in two studies. In the first, they looked at the impact of a 15-minute mindfulness meditation on divergent thinking performance among 92 engineering students at Stanford University. Previous studies have shown that a single meditation can improve idea generation in general student populations. Engineering students who reported higher baseline mindfulness performed better on the divergent thinking tasks. The impact of a single 15-minute mindfulness session on divergent thinking performance was to improve the originality of ideas in the idea generation task, but not to impact the number of ideas students came up with in the idea generation task or the engineering design task.

In the second study, they look at the relationship between mindfulness and innovation in survey results from 1400 engineering students and recent graduates across the U.S. from the longitudinal Engineering Majors Survey [295], to measure baseline mindfulness and confidence in one's ability to be innovative. Baseline mindfulness predicted innovation self-efficacy across the engineering sample, where a mindful attitude was the strongest predictor of innovation self-efficacy. This suggests

that the more essential component is the attitude with which you pay attention – or whether you have an open, curious, and kind attitude, often referred to as "beginner's mind".

The only other work we were able to identify up to now that targets breathing practices in IT is den Heijer et al. [90] who performed a controlled experiment with agile teams that practiced three minutes of a breathing technique for a month at the beginning of every Daily Scrum meeting. The participants perceived the practice as useful, and statistically significant improvement was reported on some of the dimensions in the groups performing an exercise that included listening, decision-making, meeting effectiveness, interaction, and emotional responses. In contrast, our study works with individuals instead of teams, and a breathing practice designed to support long-term well-being as opposed to short-term situational interventions.

The aim of the study at hand is to further contribute to the body of knowledge of how to decrease stress and increase well-being and resilience for software developers and IT workers.

Table 4.1 summarizes the most important concepts introduced in this section along with their differentiation.

Concept	**Definition**
Mindfulness	a state of consciousness, the practice of purposely bringing one's attention in the present moment without evaluation [289]
Attention	the behavioral and cognitive process of selectively concentrating on a discrete aspect of information [296]
Attention capacity	Amount of cognitive resources available within a person [297]
Awareness	the quality or state of being aware, knowledge and understanding that something is happening or exists [298]
Mindfulness Attention Awareness	present-centered attention–awareness [140, p. 824]
Stress	non-specific response of the body to a demand [299]
Well-being	what is non-instrumentally or ultimately good for a person [300]
Resilience	positive adaptation, or the ability to maintain or regain mental health, despite experiencing adversity [257]
Self efficacy	People's beliefs about their capabilities to produce effects. [301]

Table 4.1: Summary of the most important terms relevant for the study at hand

4.3 Research Design

Why use a breathing practice as a central technique to increase well-being and resilience? Research shows that *meditation* is great for enhancing emotional resilience and a healthy stress response [302, 303]. However, people have been restless at home, so for many the idea of meditating can trigger additional anxiety or restlessness, which counters the intention for the exploration of this study. Research shows if we can engage in *deliberate movement* where the mind gets focused on a task, this has calming effects, for example in yoga or running, but also extreme sports [304]. Explosive and/or high-intensity intervals can be great for exerting, then recovery, where the recovery part is crucial for the benefit of the overall activity, physically and mentally. Consistent activity and/or moderate intensity is sometimes better for lowering stress because if the body is already under a high level of stress hormones, the additional input can burden the body further, especially if the individual is not used to increased levels of physical activity. It depends on personality and physiology whether we relax better after exertion, or by means of moderate activity [304]. However, most importantly for the study at hand, it requires certain physical abilities by the participants whereas we wanted to make this *intervention accessible for everyone*. And everyone has to breathe.

Our brains and bodies learn through repetition and establishing habits, and "Your actions today become your brain's predictions for tomorrow, and those predictions automatically drive your future actions" [305, p. 82]. Consequently we chose for the intervention to last long enough to establish a new habit.

The program is unique to the best of our knowledge in adapting breathwork practice and framing topics specifically for a computer worker and software engineering background. The closest program we came across is the mindfulness program by Google [306], which is based on meditation.

4.3.1 The intervention: Rise 2 Flow program

We created a program to help build mental and emotional resilience and increase well-being.[4] This program is built around a specific yogic breathing practice, a so-called Pranayama (see Sec. 4.2.2) that affects Vyana Vayu (the 'wind of the nerves') or nervous system. It is a three-part breath through the mouth that is practiced laying down. The first part is an inhale in the belly, the second one an inhale into the chest, and the third part is a complete exhale. This specific pattern triggers a release in the parasympathetic nervous system and thereby helps to deeply relax. The practice is very gentle and therefore easy to use for people who are new to this type of modality. The first author is a certified facilitator of this particular technique.

The main component of the intervention program was the breathing practice. We framed the breathing practice with a weekly topic in the area of self-development. The weekly topics were used as framing for two reasons: First, to increase the interest in the study and give the participants an additional buy-in to learn applicable tools (to speak more to the left brain hemisphere oriented thinkers that we tend to have as a majority in software engineering), and to give them these tools such that they could make use of and harness their increased attention awareness and use their energy in

[4] `https://www.twinkleflip.com/rise-2-flow/`

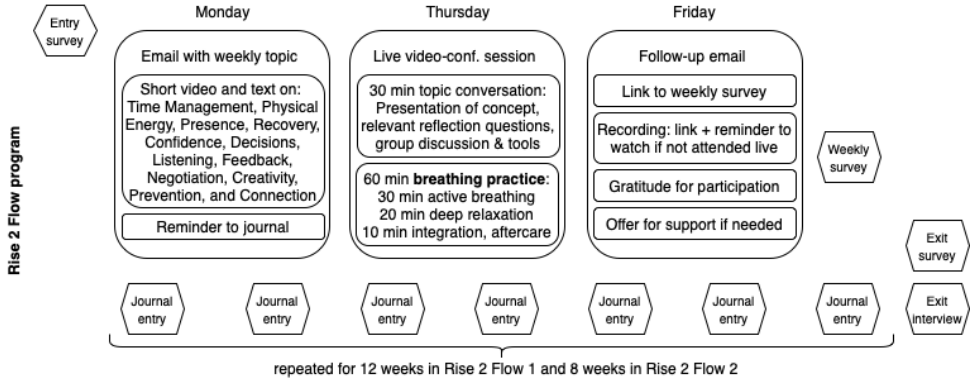


Figure 4.1: An overview of the Rise 2 Flow intervention

more effective and efficient ways. In the 12-week version (September to December 2020), the topics were Time Management, Physical Energy, Presence, Recovery, Confidence, Decisions, Listening, Feedback, Negotiation, Creativity, Prevention, and Connection. In the 8-week version (January to March 2021), we used the first eight of that list. The choice of topics is based on extensive reading in self development over the past decade and a selection of topics that to us appeared most relevant to the target population in relation to their well-being and resilience at work. Each of the topics gets primed at the beginning of the week by an email that offers a few questions to reflect upon over a few days until the group breathing practice session. A diagrammatic overview is provided in Fig. 4.1. In the practice session, conducted via an online conferencing tool, we start with a conversation around these questions (e.g., "How much sleep do you get on average?") to give the participants a bit of time to wind down as they usually come out of an intense workday. Then we give the instructions for the breathing practice, which is performed for three rounds of seven minutes each with brief relaxation pauses in between, followed by a 20-min relaxation. After the session, we finalize with aftercare suggestions (e.g., to hydrate well) and are available for questions.

There are supplementary materials that the participants can choose to make use of, namely a brief video primer introducing the weekly topic, a guided meditation on the weekly topic, the presentation slides that are used during the live session, and a workbook derived from those slides with a few more reflection prompts to journal on if desired. For participants who could not make it to the live sessions, we recorded them and made them available for home practice later during the week.[5]

We scheduled and administered data collection by means of surveys, in an entry and exit survey (about 20 min), a short weekly survey (5 rating items and a comment), and a daily journal entry (1 rating and a comment). We offered exit interviews, but these are omitted from the result analysis in this article for reasons of space.
The invitation to a brief daily journal entry and short weekly surveys were necessary

[5] `https://www.twinkleflip.com/rise2flow/` with recording links removed for privacy of the participants.

to collect data that would help us understand that comparison points from entry and exit survey and to see a development over time. We do not see them as part of the intervention per se, as such a reflective practice would not necessarily change the wellbeing of participants much by itself (unless specifically prompted, for example, as gratitude practice [307]) but rather helps to notice changes that occur (in our case, prompted possibly by the intervention).

We conducted two pilots, which we include here for completeness and also to report a logistic failure in the first one[6]. The first attempt enrolled 30+ participants in Spring but then led to only 10 submissions of the entry survey and 0 submissions of exit survey. The logistic failure was that the first author enrolled students from a course she was teaching and the ethical research guidelines required that she did not know who in the course was participating in the study (to mitigate the risk of coercion) and therefore she could not follow up with participants after sending the initial instructions. In addition, there were no live components, only recordings, so that may have significantly lowered engagement.
The second pilot came into existence upon reporting on the failed pilot at a conference (ICT for Sustainability 2020), where researchers in the audience volunteered for a second round. So we set up a second round, this time with direction communication, and an option to participate in weekly live sessions or to listen to recordings. However, the timing was sub-optimal as many participants dropped off over summer (30+ sign-ups, 15 submissions of the entry survey, 3 submissions of the exit survey).

In the two instances from the family of experiments (Rise 2 Flow 1 and Rise 2 Flow 2), direct interaction online was an important component, and the timing was just after the start of the academic year. For Rise 2 Flow 1, 34 completed the study exit survey, and for Rise 2 Flow 2, 33 completed the exit survey (both times out of 100+ sign-ups that we had advertised for globally).

4.3.2 Research Questions

For reporting on the implementation of the study, we are interested in how participants engage with the program. Specifically, how many sessions they attended or listened to the recording of, what their results of the practices were, and which practices participants applied in their daily lives.

The research questions for the study at hand are as follows:

[a] How did participants' mindfulness attention awareness and daily perceptions of their experience of life change over the course of the breathwork program?

  (a) Does the intervention bring about change in the participants Mindfulness Attention Awareness?

  (b) How did the daily perceptions of life experience progress over time?

[b] What is the observable change in participants' reported well-being over the course of the breathwork program?

  (a) Is there change in the participants' perceptions of positive and negative experiences? If so, how are their experiences affected?

[6] Because we report rarely on failure in our discipline, we miss out on collective learning from these failures.

(b) Is there change in their psychological well-being? If so, how is it affected?

(c) Is there change with regard to their positive thinking? If so, how is it affected?

(d) How does the well-being fluctuate and vary over the course of the course of the intervention?

[c] What are the observable changes in participants' perceived productivity and self-efficacy over the course of the breathwork program?

(a) Does the intervention lead to change in the participants perceived productivity? If so, how is it affected?

(b) Does the intervention lead to change in the participants' self-efficacy? If so, how is it affected?

The overall expectation is that well-being and resilience increase over the course of the study, evidenced in quantitative and qualitative data as collected in the survey instruments and interviews.[7]

4.3.3 Population and Inclusion Criteria

The target population are people who work in IT and software development. The inclusion criterion is that they spend at least 70 percent of their work time in front of a screen.[8] We expected a sample that would include software developers, IT practitioners, IT researchers, IT consultants, faculty, and students. We included the latter as options because we find it highly relevant to address the educational aspect of offering practices early on in career development, not only when all routines have been set in place. The participants for both runs of the experiment were recruited across a range of personal and online networks, including the global personal network of the first author, university networks, mailing lists, online spaces, and social media channels. We reached out to several hundred colleagues around the globe per email to ask them to promote the study in their courses. We pitched the study live to six courses given by colleagues in Sweden and California. We also reached out to our alumni network in several countries per email, consisting of several hundred members. Furthermore, we posted invitations on research community mailing lists (ICT for Sustainability, LIMITS), and advertised in a series of posts on Twitter, LinkedIn, Facebook, Instagram, and several large Slack spaces. There was no compensation for the study, so the only incentive was to learn the breathing technique and practice in a facilitated group. The video pitch for the study is available here: `https://youtu.be/ifdo4-ZCoFM`

While it is not a classical convenience sample because of the number of channels used for broadcasting, it can be seen as an extension thereof [309].

[7] We decided to not work with hypotheses for the statistical part because the American Association for Statistics recommends to not use dichotomous hypothesis testing because very seldom when we study complex systems we can derive to such an easy T/F answer [308]. In addition, our study uses a combination of qualitative and quantitative data and therefore provides richer answers than T/F.

[8] We advertised the study as such and asked participants in the sign-up form, and we have to rely on their self-assessment of that criterion.

This study was carried out in accordance with the recommendation for experimental guidelines of Chalmers University of Technology with informed consent from all subjects. Because of the informed consent and the non-intrusive nature of the study, no formal ethics committee was required to review the study as per the university's guidelines and national regulations.

4.3.4 Instrument Design

The instruments for the Rise 2 Flow study comprise an entry survey and an exit survey, to be taken before the first practice session and after the last practice session. It encompasses items on attention awareness, positive and negative experiences, psychological well-being, positive thinking, perceived productivity, and self-efficacy, (Sect. 4.3.4.1).

In addition, there is a weekly well-being check-in survey, where participants rate their well-being using five items (Sect. 4.3.4.3), as well as a daily journal entry (Sect. 4.3.4.2). The instruments are included in App. A.2.

Answering several calls in the field [310–313], this work adopts validated measurement instruments that come from psychology.

4.3.4.1 Entry/Exit Survey

Our entry survey is composed of several validated instruments in related work, the Mindfulness Attention Awareness Scale (MAAS), the Scale of Positive and Negative Experience (SPANE), the Psychological Well-Being scale (PWB), the Positive Thinking Scale (PTS), a Perceived Productivity instrument (HPQ), and a Self-Efficacy instrument—all of which are introduced and explained in this section. The exit survey had those same instruments, in order to have a comparison point.

Mindfulness Attention Awareness Scale (MAAS, for RQ1a) Brown and Ryan [140] presented their scale to validate the benefits of being present by demonstrating the role of mindfulness in psychological well-being under the name 'Mindfulness Attention Awareness Scale' (MAAS). The instrument assesses individual differences in the frequency of mindful states over time. Its development began with a pool of 184 items that was subsequently reduced to 55 and then 24 items.

After exploratory factor analysis, the final version included 15 items. The items are distributed across cognitive, emotional, physical, interpersonal, and general domains. MAAS respondents indicate how frequently they have the experience described in each statement using a 6-point Likert scale from 1 (almost always) to 6 (almost never), where high scores reflect more mindfulness. To control for socially desirable responding, respondents are asked to answer according to what "really reflects" their experience rather than what they think their experience should be [140].

It has been widely used in clinical psychology, e.g., [314, 315], behavior assessment, e.g., [316], cognitive therapy, e.g., [317], and psychosomatics, e.g., [318].

Most later research has verified its validity, for example Baer et al. [315] combined five questionnaire instruments and confirmed MAAS' good psychometric properties. Barajas and Garras confirmed its validity in a large Spanish sample [319] and Deng et al. [320] in China. MacKillop and Anderson [316] performed a confirmatory factor

analysis that supported the unidimensional factor structure of the MAAS in their overall sample.

MAAS has been critized for only one aspect, which is using negative statements in their rating [321], which could affect construct validity. Hofling et al. [321] propose that MAAS can be assessed by both positively and negatively worded items if trait-method models are applied. Their 10-item version MAAS-Short uses five positively and five negatively worded items and is superior to the MAAS with regard to internal consistency, but content validity might be restricted with fewer items.

Consequently, we chose to use the original version of the instrument.

Scale of Positive and Negative Experience (SPANE, for RQ2a) Diener et al. [141] proposed a set of related instruments in 'New measures of well-being' that includes the Scale of Positive And Negative Experience (SPANE), the scale of Psychological Well-being (PWB), and the Positive Thinking Scale (PTS). In his meta analysis of studies applying Diener et al.'s instruments, Busseri [322] examines the structure of subjective well-being, and confirms the associations among positive affect, negative affect, and life satisfaction. To measure these different aspects, all three instruments are part of the entry and exit survey in the study at hand.

The Scale of Positive and Negative Experience (SPANE) elicits a score for positive experience and feelings (using six items), a score for negative experience and feelings (six items), and the two are combined to create an experience balance score. The respondent selects on a Likert scale how often they have experienced the specific feeling over the past month. The scale assesses a broad range of negative and positive experiences and feelings with only twelve items.

Each item is scored on a scale ranging from 1 to 5, where 1 represents "very rarely or never" and 5 represents "very often or always." The summed positive score can range from 6 to 30, and the negative scale has the same range. The two scores are combined by subtracting the negative score from the positive score, and the resulting scores can range from -24 to 24. The SPANE is based on the duration during which people experience the feelings, which is beneficial because this aspect of feelings predicts long-term well-being, and it can be better calibrated across respondents. Furthermore, the SPANE is based on feelings that occurred during the previous four weeks, and thus reflects a balance between memory accuracy and experience sampling [141].

The instrument has been mostly supported by later studies, including for example by Jovanovic [323] wo demonstrated that SPANE is a useful measure of affective well-being. It performs better than the earlier Positive and Negative Affect Schedule (PANAS) by Watson, Clark, and Tellegen [324], in predicting well-being among young adults and adolescents.

Psychological Well-Being (PWB, for RQ2b) Diener et al.'s [141] scale of Psychological Well-being (PWB) is a broad measure of a number of aspects of psychological well-being. It assesses meaning, positive social relationships (including helping others and one's community), self-esteem, and competence and mastery. The PWB provides a good assessment of overall self-reported psychological well-being. While, for the objective of brevity, it does not assess the individual components of

psychological well-being described in some theories, it proved to have high internal and temporal reliabilities and high convergence with other similar scales [141].

The Psychological Well-Being scale (PWB) consists of eight items describing important aspects of human functioning ranging from positive relationships, to feelings of competence, to having meaning and purpose in life. Each item is answered on a 1–7 scale that ranges from Strong Disagreement to Strong Agreement. All items are phrased in a positive direction. Scores can range from 8 (Strong Disagreement with all items) to 56 (Strong Agreement with all items). High scores signify that respondents view themselves in very positive terms in diverse areas of functioning.

Positive Thinking (PTS, for RQ2c) People's habits of positive thinking are not the sole determinant of happiness as circumstances can influence well-being as well. However, the propensity to positive or negative thinking can influence a person's feelings of well-being, while controlling for environmental circumstances. Thus, Diener et al. [141] developed a scale of Positive Thinking (PTS) as a measure of the propensity to view things in positive versus negative terms.

The Positive Thinking Scale (PTS) is composed of 22 items, where 11 items represent positive thoughts and perceptions and 11 items represent low negative thinking. The 22 items are answered on a yes/no format. Negative items are reverse scored with a 'no' response counting as a '1'; and for positive items a 'yes' response counts as a '1'. After reversing the negative items, the 22 items are added, thus yielding scores that range from 0 to 22 [141].

The authors point out that currently the focus is on attention and interpretation, while taking into account both rumination and savoring would require greater sampling of memories. In addition, a desirable future extension of the scale would be to include thoughts about nonsocial aspects of the world [141]. For the study at hand, our reason for including the scale was that low PST scores might contribute to explaining variance of well-being scores in otherwise similar contexts.

Perceived Productivity (HPQ, for RQ3a) To assess perceived productivity we used items from the WHO's Health and Work Performance Questionnaire (HPQ) [142], a self-report instrument designed to estimate the workplace costs of health problems in terms of reduced job performance and sickness absence. It was developed because untreated (and under-treated) health problems demand substantial personal costs from the individuals who experience them as well as from their families, employers, and communities [142] and was later validated further as an adequate instrument [325].

The HPQ[9] measures perceived productivity in two ways: First, it uses an eight-item scale (summative, multiple reversed indicators), that assesses overall and relative performance, and second, it uses an eleven-point list of general ratings of participants' own performance as well as typical performance of similar workers.

Self-efficacy (for RQ3b) We used the same Self-efficacy instrument used by Ostberg et al. [143] in their work on psycho-biological assessment of stress. The instrument was developed by Jerusalem et al. [326] and based on Bandura et al.'s [327] self-efficacy model. It is used to assess the individual stress resilience of the participants

[9] http://www.hcp.med.harvard.edu/hpq

and encompasses ten items that offer a positively phrased statement on change, challenges or unexpected circumstances which the participant has to rate as "Not true", "Hardly true", "Rather true" or "Exactly true".

The study at hand, serves to gain further insights into and potentially confirm the correlation of well-being, positive thinking, and stress resilience.

Personal Data We collected information on the participants' country of residence, gender, living situation, occupation, and age. The options we offered were: (1) Gender: Man (including cis-man and trans-man), Woman (including cis-woman and trans-woman), Non-binary, and 'prefer not to say'. (2) Living situation: by themselves, with a partner, with their family, in shared housing. (3) Occupation: Student, Faculty, Researcher, Developer, Administrator, IT Services, Manager, Digital Artist, Analyst, Consultant, Retired, Currently not working, and Other.

4.3.4.2 Daily journal (for RQ1b)

The daily journal entry contained three items (plus the participant's alias and the date, for reference):

- Which well-being practice did you do today (if any)? Select all that apply: Breathing practice, Yoga postures, Meditation, Nature time, Other.
- How was your day? Select a rating from 'really bad' (1) to 'absolutely great' (10).
- Please write about 100 words: What stood out to you today? What caught your attention? What makes you reflect?

We intended to have a daily plot over time, and to check the correlation with the practices that were carried out. We were aware that there is a bias in rating as people may not do a well-being practice or even remember to write a journal entry on days where they felt particularly high or low. The free text gave room for individual reflection and was deliberately prompted in a very open manner.

4.3.4.3 Weekly survey: WHO-5 (for RQ2d)

The 5-item World Health Organization Well-Being Index (WHO-5) is a short and generic global rating scale measuring subjective well-being. Because the WHO considers positive well-being to be another term for mental health [328], the WHO-5 only contains positively phrased items, and its use is recommended by [329]. The items are: (1) 'I have felt cheerful and in good spirits', (2) 'I have felt calm and relaxed', (3) 'I have felt active and vigorous', (4) 'I woke up feeling fresh and rested', and (5) 'My daily life has been filled with things that interest me'. The respondent is asked to rate how well each of the 5 statements applies to him/her/them when considering the last 14 days. Each of the 5 items is scored from 5 (all of the time) to 0 (none of the time).

Topp et al. [330] performed a recent systematic review on the WHO-5, which included 213 articles from the PubMed and PsycINFO databases. They concluded the WHO-5 has high validity, can be used as an outcome measure balancing the

wanted and unwanted effects of treatments, and is a sensitive and specific screening tool. Furthermore, its applicability across study fields is very high. Consequently, it is a valid choice for the purpose at hand.

We included the weekly survey to be able to observe a development over time. We modified the instrument slightly as we asked to consider the last week, because we administered the survey every week within the 12 weeks of the study. The instrument is included in the replication package [331].

This weekly measurement is relevant for several reasons: 1) To have a longitudinal study that shows the trends over time as opposed to only entry and exit data points, 2) To correlate and validate the insights from the difference in well-being ratings from the entry and exit survey, 3) To investigate the correlation with the development of the pandemic (would mood go down with restrictions increasing?), 4) To check for variance and fluctuations in the well-being as additional indicator for long-term psychological well-being [332].

4.4 Analysis Procedure

The previous section covered the research design of the interventions, the study, and the instruments that were administered. This section provides the demographics and the two types of analyses that were performed on the data. First, a three-way quantitative analysis of the administered instruments. Second, a thematic analysis of answers to the open survey questions. After the analyses, we provide a summary of the results in Sect. 4.5.

4.4.1 Demographics

For Rise 2 Flow 1, of 137 sign-ups, 87 converted to the entry survey, and 34 completed the study by submitting their exit survey. We collected 1040 individual journal entries. For Rise 2 Flow 2, of 169 sign-ups, 101 completed the entry survey, and 33 the exit survey. Participants submitted 616 individual journal entries.

The high drop out rates have a number of reasons, which are partially known to the first author because of emails kindly sent to her by participants who cared to explain their personal reason for not completing the study. The reasons included the live session being in the middle of get-kids-to-bed time, conflicting commitments for some of the days, and general life stress, which participants did note was a good reason to actually do the sessions but they felt they just could not at the time.

The entry surveys showed an age range from 19 to 58, with a majority of participants in their 20s and 30s.

The occupation was a selection where participants could check all that apply, so the following numbers add up to more than the total number of participants: Student: 33 (run 1) + 45 (run 2) = 78, Developer: 6 + 11 = 17, Researcher: 19 + 42 = 61, Faculty: 10 + 19 = 29, Other (Consultant, Analyst, Manager, etc.): 27 + 37 = 64;

Participants joined from all over the world: Argentina, Austria, Bangladesh, Brazil, Canada, Costa Rica, Denmark, Ecuador, Finland, France, Germany, India, Iran, Ireland, Italy, Netherlands, Mexico, Poland, Portugal, Saudi Arabia, Spain, Sweden, Switzerland, UK, US, and Venezuela.

4.4.2 Statistical Analysis of Instruments

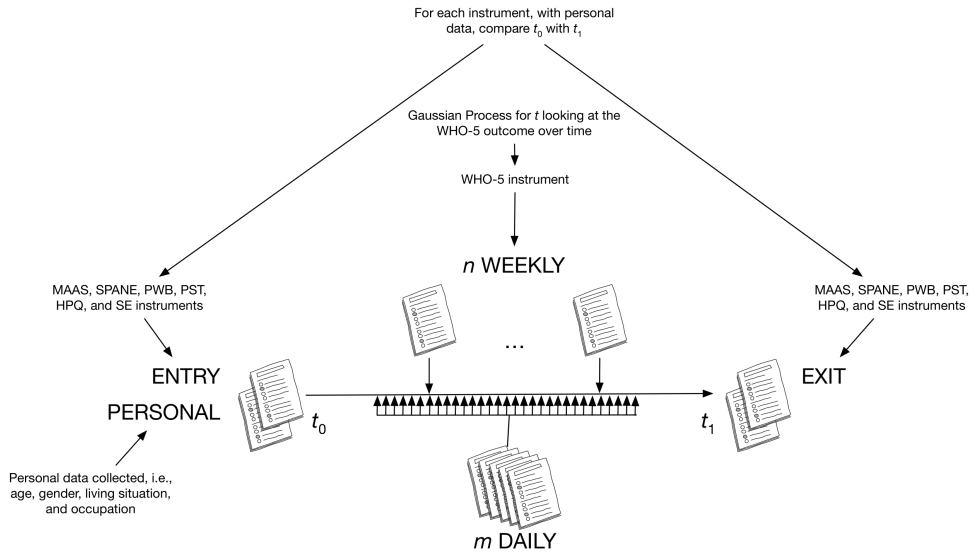


Figure 4.2: An overview of instruments used in the study. Note that the above was executed twice, once for each run of the experiment.

As presented in Sect. 4.3.4, the study used several instruments for the two experiments. Figure 6.2 provides an overview of the instruments and at which time they were administered. A subject was administered an entry survey consisting of six instruments (as presented in the previous section) plus personal data, e.g., current living situation. For each week the subject was administered a weekly survey (the five questions of WHO5) and for each day they answered a daily survey (one question). Finally, upon exiting the study, the subject was administered an exit survey, which was identical to the entry survey.
The purpose of the quantitative analysis is to look at how responses change over time (a temporal analysis).[10] This will be done in three ways:

[a] Temporal analysis for each instrument at t_0 vs. t_1, i.e., entry vs. exit.

[b] Temporal analysis of daily trends.

[c] Temporal analysis of weekly trends.

In the last case, we will use dummy variable regression estimators (DVRE). The DVRE approach dummy encodes the time variable t and sets an index 0/1, where $t_0 = 0$ and $t_1 = 1$. In short, each subject (ID) will have two rows where one row are the entry instruments at t_0, and one row are the exit instruments at t_1. The main

[10] A replication package can be found at `https://github.com/torkar/rise2flow` DOI:10.5281/zenodo.5082388

reasons to use this approach is: 1. We will see if there is a difference in responses between t_0 and t_1. 2. If such a difference exists, which predictors, if any, are the main drivers for that difference, i.e., is there a difference in the β estimators for each predictor, in each question?[11]

For the first two cases (weekly and daily trends), we will model these using a Gaussian Process ($\mathcal{GP}$). The distribution of a $\mathcal{GP}$ is the joint distribution of all random variables. In short, it is a distribution over functions with a continuous domain, i.e., time in our case. In Bayesian and frequentist multilevel models it is common to model varying intercepts (random effects), which are categorical. However, for continuous values (such as time or space) one needs to use a different strategy. A $\mathcal{GP}$ is such a strategy, i.e., it is a varying intercept approach, but for continuous values.[12]

Before introducing the statistical model design, the next section will present descriptive statistics, dependent and independent variables (and their encoding), and the sample sizes involved in each of our three analyses mentioned above.

4.4.2.1 The Data and Data Cleaning

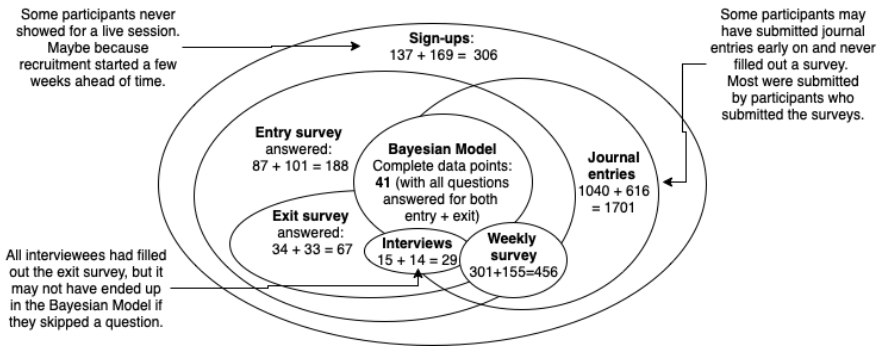


Figure 4.3: Break-down of collected data and subsets for analysis

The entry and exit surveys consisted of six instruments (in parenthesis the number of questions): MAAS (15), SPANE (12), PWB (8), PST (22), SE (10), and HPQ (11). The questions were coded as ordered categorical, i.e., Likert scale, or binary (Yes/No answers).

Concerning the weekly and daily surveys there were five and one question(s), respectively, and the questions were coded as ordered categorical. For the weekly data 98 subjects answered the survey at least once and some as much as 12 times, $\tilde{x} = 3$ ($N = 456$). Concerning the daily data, 111 subjects answered the survey at

[11]The same approach was used to encode data for the two experiments that were conducted, i.e., Experiment 1 was coded as 0 and Experiment 2 was coded as 1. The main reason for why we chose to encode experiments separately for each subject was to analyze the estimates and their uncertainty; ultimately we needed to ascertain that they did not vary considerably between the two experiments, which could indicate that, e.g., the experiments were not executed in a similar fashion.

[12]Generalized additive models is another approach one could use, however, that approach uses various types of smooth functions instead.

Table 4.2: Independent variables (IV) used as predictors.

IV	Type	Levels
ID	factor	Unique for each subject
age	continuous	n/a
gender	dichotomous	male ‖ woman
occupation	dichotomous	student ‖ non-students
living condition	categorical	I live by myself ‖ I live in a shared housing ‖ I live with a partner ‖ I live with my family

least once. The maximum value was 84, while the median was $\tilde{x} = 16$ ($N = 1646$).[13]

The above are the outcomes.[14] The outcomes will be predicted given a number of predictors. An overview of the predictors used in this analysis can be found in Table 4.2.

The data cleaning included correcting IDs that the subjects had spelled differently. In some cases we had subjects that only answered one of them; 105 subjects filled out only the entry survey, while 41 filled out every single answer for both entry and exit survey. The Bayesian model required 'complete case analysis' [333], i.e. all questions answered, as missing data analysis would have required a causal model for why a participant didn't answer a particular question, which was not feasible in our case due to the number of questions. While we were curious to investigate where the gender 'non-binary' could be a predictor, the number of answers with that value was too low to result in a valid model. An overview of how the data breaks down for analysis is shown in Fig. 4.3.

4.4.2.2 Temporal Analysis of Entry vs. Exit

In Appendix A.3.2 a complete specification of the model is listed. Here follows a brief summary of modeling choices; details can be found in the replication package.

A Cumulative likelihood was assumed for the Likert scale questions. In one case (where the outcome consisted of 'Yes'/'No' answers) the maximum entropy distribution was used (i.e., Bernoulli). Variance between questions in an instrument was modeled using a covariance matrix. Additionally, subject variability (ID) was modeled with adaptive priors to employ partial pooling (more information can be found in Appendix A.3.2 and the replication package).

The question posed for these models was: Given a number of predictors (age, gender, occupation, living condition) is there a difference between t_0 (entry) and t_1 (exit), when accounting for subject variability (ID). Prior predictive checks (prior sensitivity analysis) and posterior predictive checks were conducted.[15]

[13]The 1646 differ from the 1701 in Fig. 4.3 because the rating of the day was optional and 1646/1701 entries were rated.

[14]Throughout this text we will from now on use the terms outcome and predictor, concerning dependent and independent variables.

[15]Additionally, all diagnostics ($\widehat{R}$, ESS, traceplots, E-BFMI, divergences, and treedepth) indicated

To check whether the observed effects correlated with actual participation in sessions instead of only time passing, we also ran models for the predictors of total-number-of-sessions-attended. To observe differences in between participation in live and recorded sessions, we also ran models for the predictors of sessions-attended-live and sessions-attended-recorded.

4.4.2.3 Temporal Analysis of Weekly and Daily Trends

In Appendix A.3.1 a complete specification of the model is listed. Here follows a brief summary of modeling choices; details can be found in the replication package.

Before designing a model, assumptions concerning the data generation process need to be considered. In this study, an information theoretical comparison of possible data generation processes, i.e., Cumulative, Continuation ratio, Stopping ratio, and Adjacent-category, was conducted. The analysis showed that the differences in standard error, between the likelihoods, was fairly large, in comparison to the relative difference in expected log point-wise predictive density. In short, no likelihood showed significantly better out of sample prediction capabilities, when compared to the other likelihoods. Hence, a Cumulative likelihood was assumed for Likert-type questions, while for the questions that were dichotomous the maximum entropy distribution was selected, i.e., Bernoulli.[16]

The statistical model was designed (see Appendix A.3.1) with three things in mind. First, the covariance between questions for each subject was modeled employing a covariance matrix. The idea here is that the variability among questions, for each subject, should be captured. Second, the temporal variable (weeks or days) was modeled with a Gaussian Process. Gaussian Process has not been applied in software engineering, as far as we know, but is not uncommon in other disciplines and the concept is perceived as particularly suitable for longitudinal data [334] and variable selection [335] in a Bayesian context [336]. Finally, when modelling the between-subject variability, partial pooling was used (i.e., self-regularizing priors) to avoid overfitting.

Concerning the later design choice, due to us employing a multilevel approach with partial pooling (using a varying intercept for each subject), subjects with a large sample size (answered many times) will inform subjects with a small sample size (answered once or a few times), i.e., the uncertainty will propagate through the model depending on the sample size and we will avoid learning too much from the data (to avoid overfitting). Figure 4.4 shows the challenges researchers face when dealing with response rates in longitudinal studies.[17]

In order to handle said threat, and due to the study starting at different time points for each subject, each subject's answer was coded with a time, i.e., $1, \ldots, n$, where n is the last answer they provided. This way a time point in the study, e.g., Week 3, was the same for all subjects who were still participating in the study at Week 3. To summarize, the logic was that the intervention, that is, participating in the study, was exchangeable from a statistical point of view, i.e., Week 3 was the same no matter whether the subject joined the study for the first or the second run. For

that the Markov Chains had converged to a stationary posterior distribution.

[16] Please see the Appendix A in the replication package.

[17] A survival analysis might be a different method to use in future work.

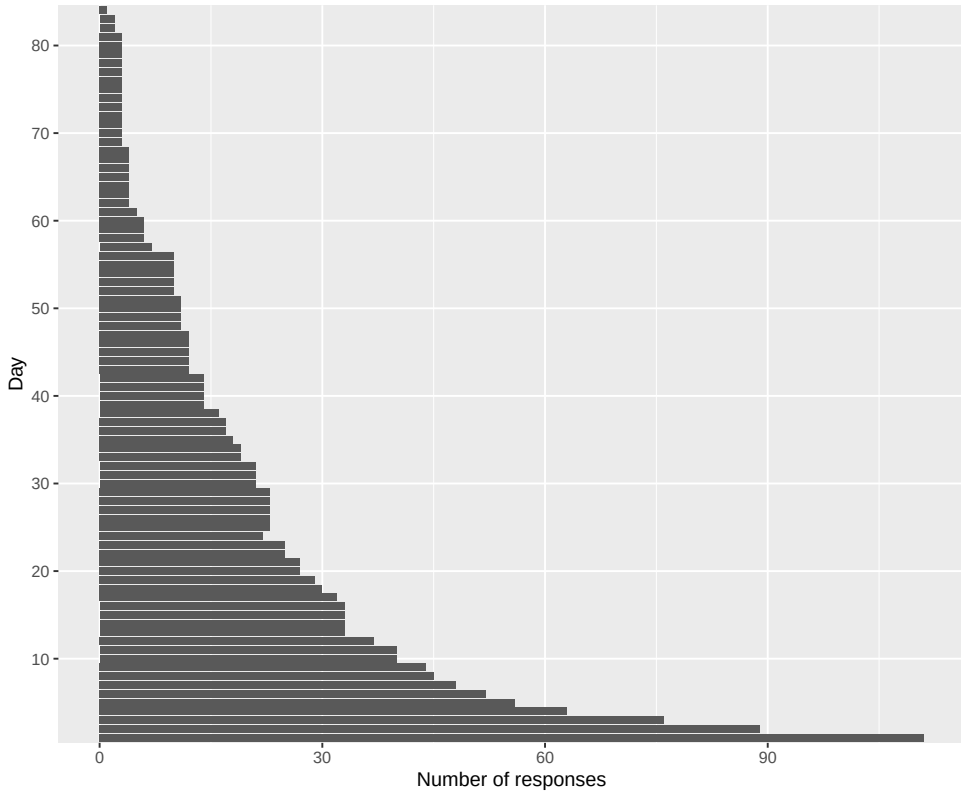


Figure 4.4: Response rate for each day ($N = 1646$). Respondents replied enthusiastically during the first week (111 respondents), while on Day 84 only 1 respondent replied.

details on the validity and latency, see App A.4, Fig. A.1. Prior sensitivity analysis was conducted.[18]

As mentioned previously, the model took into account that we had two experiments in this study. By contrasting the underlying latent scale of Experiment 1 and 2 (for each question) we could see that they had overlaying medians and similar shapes.[19] Hence, there were strong indications that the two experiments had been executed in a very similar fashion.

4.4.3 Qualitative Analysis of Instruments: Thematic analysis

Thematic analysis extracts themes from text [120]. Coding qualitative research to find common themes and concepts is part of thematic analysis, which is part of qualitative data analysis. We present an analysis for the daily journal as well as the entry and exit survey. The coding was performed independently by the first and third author and cross-reviewed. The weekly survey did not include any open questions. Table 4.3 shows an example of the coding process.

Table 4.3: Example of the coding process

MEANING UNIT	CODE	THEME	SUB-THEME
"All together, all the awareness of the program has helped me to be more focused and present, enjoying what I am doing, what I don't enjoy still finalize it without falling in the temptation of getting distracted by the first thing that pops in my mind (still happens, but is getting better)."	Focus	Perceived changes in participants	At work

For the process of coding, the third author initially coded the data until meaning saturation was reached ("we learned everything from the data that we could)" to the best of guidance in the state of the art [337]. The first author audited the coding.

Figure 4.5 shows the complete code map of the thematic analysis along with the times that each code was assigned during the coding process.

4.5 Results

In summary (Tab. 4.4), we found a number of quantitative results that were statistically significant, plus a plethora of revelations in the qualitative data that explained some statistical results that had made us wonder, and gave insights into the deep processes of change and growth that some of the participants experienced.

[18]The priors were uniform on the outcome space (i.e., medians were distributed evenly with large uncertainty). After sampling, posterior predictive checks indicated that each model had learned from the data and washed out the effect of the priors.

[19]See Sects. 2.1.1–2.1.5 in the replication package.

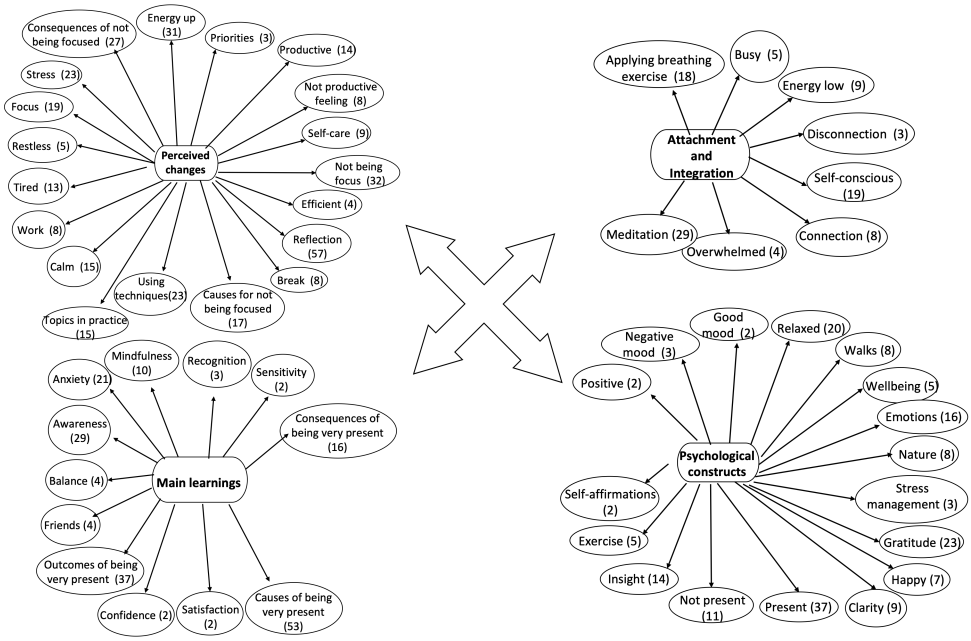


Figure 4.5: Final Code Map of the Thematic Analysis

Where there are quantitative results, we present the estimates that were significant at the arbitrary 95% threshold. For each instrument, we will first analyze if there was a difference between questions at t_0 vs. t_1 (i.e., entry vs. exit surveys). After that significant effects for the other parameters (for the predictors age, etc.) are presented. Regarding the qualitative findings to answer the research questions, four themes and four sub-themes were identified presented in Table 4.5 below.

4.5.1 Overall Engagement

Before we answer the individual research questions, we present qualitative findings around the **overall engagement of participants** with the program to give an insight into the context.

Attachment and Integration of techniques in daily life—Actions. The engagement was variable in each participant. Some people were really committed and put much effort into doing the practices every day, and some other participants could not manage due to their schedules. For example, the quote below shows a strong commitment from this participant based on their need.

> *"I will definitely continue with more of these activities (meditation), as they already change my day, mood, and approach towards daily life, routine, and also future plans." (participant 14, run 2, journal, Mar 30 2021)*

RQ	Phrasing	Quantitative	Qualitat.
RQ1	**How did mindfulness attention awareness and daily perceptions of experience of life change?**		
RQ1a	Does the intervention bring about change in the participants Mindfulness Attention Awareness?	negative (MAAS)	positive
RQ1b	How did the daily perceptions of life experience progress over time?	inconclusive (daily)	positive
RQ2	**What is the observable change in participants' reported well-being?**		
RQ2a	Is there change in the participants' perceptions of positive and negative experiences? If so, how are they affected?	supported (SPANE)	positive
RQ2b	Is there change in their psychological well-being? If so, how is it affected?	supported (PWB)	positive
RQ2c	Is there change with regard to positive thinking? If so, how is it affected?	supported (PTS)	positive
RQ2d	How does the well-being fluctuate and vary over the course of the intervention?	inconclusive (weekly)	positive
RQ3	**What are the observable changes in perceived productivity and self-efficacy?**		
RQ3a	Does the intervention lead to change in the participants perceived productivity? If so, how is it affected?	inconclusive	positive
RQ3b	Does the intervention lead to change in the participants' self-efficacy? If so, how is it affected?	supported	positive

Table 4.4: Overview of the evidence for answering the research questions

On the other hand, there were cases when participants could not find the opportunity during the day to do the practices. However, they found time to write in the journal, as the following quote shows.

> *"Unfortunately, my day was so full, from rising to bedtime, that I didn't have time for any of the practices that remind me to breathe." (participant 3, run 1, journal, Nov 26 2020)*

The commitment of the participants can be seen in the different activities they carried out. Not everyone carried out all the activities at the same time. There were

Table 4.5: Themes and sub-themes in response to the RQs

Themes	**Sub-themes**	**RQ**
Main learnings (changes) identified by the participants		1
Attachment and Integration of techniques in daily life	* Results	1
	* Actions	1
Psychological constructs modified during the course		2
Perceived changes in participants	* At work	3
	* In overall performance	3

those who attended the live sessions but did not write in the diary. Some participants wrote every day or almost every day and who wrote constantly but not daily. Similarly, there were also a significant number of drop offs.
The results of the practices reported by the participants are varied and interesting, as shown in the following quotes.

> *"I felt pretty bad so I decided to do the breathing practice. It is interesting to notice that I think I have a new tool to calm after pretty bad days, a tool that does not involve heavy use of alcohol." (participant 21, run 1, journal, Sept 28 2020)*

This quote is one of the ones that stands out the most, since the participant found in the breathing practice an alternative to calm down instead of the use of alcohol.

> *"But I did relieve a growing panic attack with breathing exercises, which felt nice. And of the waking hours, I did feel like I had spent my time more wisely than usual." (participant 46, run 2, journal, Jan 30 2021)*

This participant received a diagnosis of initial depression and burnout; they explain how breathing was the tool that helped them release emotionally.
The quote above is also an essential example of the benefits that participants obtained by practicing breathing exercises. This participant wrote that they could control a panic attack and expressed a better use of time.
The previous paragraphs describe complex situations in which these participants, through breathing, could find tools that they needed according to their problems.
Other results expressed by participants are being more relaxed, in general, and in situations where they previously would not have been; the feeling of being more present at crucial moments of the day; fewer negative thoughts; and ease of letting go. Similarly, some other results were greater focus and better function, more energy, and more centered, a clear mind, calm, peace, gratitude, reflection, or better analysis of situations.

Overall, we observe diversity. In terms of *Attendance*, participants came from around the globe, had to deal with time zone differences and work and family situations,

and did their best to show up when they could. Some attended only one session live, others almost all of them; some attended only or mainly live, others mainly or only recorded sessions. Calculated from the survey responses, notes from the sessions and views on the online platform, 75% of the participants who filled out the exit survey had participated in at least 75% of the sessions. [20]

The *Results of their Practice* varied as well. In the exit survey, we asked "What did you get out of the breathing sessions?" and the most frequent answers were: It is an easy practice to follow (17), I was able to deeply relax during the breathing and in the relaxation period after (22), I remained relaxed after, and felt well rested and recharged the next day (15), I shifted my perception of the world and have interesting insights (9), I feel more present in my body (16), and I have the desire to return to the breathing practice (23).

In terms of *Daily Applied Practices* and records thereof, we received 1032 journal entry submissions. Of the daily well-being practices reported in the journal entries, nature time was selected 261 times (32.6%), followed by meditation (240/30%), breathing practice (227/28.4%) and yoga poses (82/10.3%). Other practices respondents listed include dancing, offline time, reading, swimming, massage, Qi Gong, art, family/friend time, and exercise. In the article at hand, we limit ourselves to the quantitative analysis and a more high-level analysis of the qualitative survey data and the journal entries. More details on the daily applied practices reported on in those are presented in another publication for reasons of space.

4.5.2 RQ1: Changes in Mindfulness Attention Awareness and Daily Perceptions

Research question 1: "How did participants mindfulness attention awareness and daily perceptions of their experience of life change?" was split into three sub-questions which we answer in the following.

For the qualitative analysis, Van Dam's suggestion was taken into consideration to analyze the concepts on their theory-based conception [338] to better capitalize the benefits of Mindfulness. Given that the concept of Mindfulness per se is multifaceted [339], in order to optimally analyze the effects and experiences in the participants, it was necessary to handle them separately. Hence the authors decided to divide the concept of "Mindfulness Attention Awareness" into the concepts of "Mindfulness", "Attention, and "Awareness". This division allows to better explain the experience of the individuals for the encompassing concept. At the same time, this separation made it possible to implement one more of Van Dam's suggestions, which is to consider the contribution of traditional Buddhist conceptualizations and psychological implications [338]. This was enabled by the long-term meditation study background of the first author and the psychology expertise of the third author.

[20] Due to the varied attendance, we also ran models with attendance as predictor to make sure the effects we saw over time correlate with attendance and not just time.

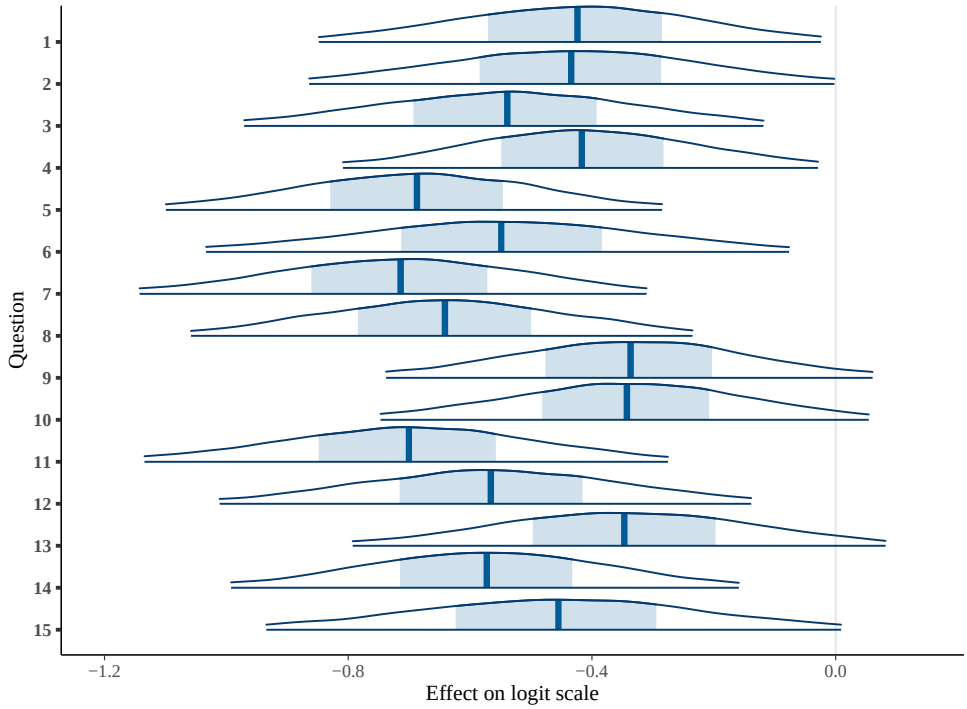


Figure 4.6: Density plots computed from posterior draws for MAAS. The densities are cut off at 95%, the blue vertical line is the calculated value of the model for this item, and the shaded area is the 50% uncertainty interval. We can see a number of questions crossing zero (no effect observed). Most effects are negative (to the left of zero), which means that participants rated more negatively at t_1 (exit survey) than at t_0 (entry survey), so they were under the perception that their mindfulness attention awareness had decreased.

4.5.2.1 Does the intervention bring about change in the participants Mindfulness Attention Awareness? (RQ1a)

The MAAS instrument (App. A.2.1) consisted of 15 statements to agree or disagree with. In the **quantitative findings**, eleven of the ratings indicated a significant difference at t_0 vs. t_1: Q1–8, 11–12, and 14. In all the above cases the effect was negative, i.e., the responses were higher at t_0 than at t_1, as visible in Fig. 4.6. For the other predictors, age and gender did not have a significant effect, while occupation was significant (negative) for Q2, i.e., "I break or spill things because of carelessness, not paying attention, or thinking of something else."

Additionally, the predictor living condition was significant (negative) in Q1–3, 8, and 12 (items listed in App. A.2.1). Figure 4.7 provides an overview of what this implies on the outcome scale, the Likert scale for the five questions, where living condition was significant. This result could indicate that people who live with their

family may be more occupied with the well-being of the ones around them that they feel responsible for, or that they tend to be more preoccupied because they do not find sufficient time and space for themselves to unplug and recharge.

In summary, a number of significant effects were found. Considering the temporal variable, five questions indicated a difference between t_0 and t_1 (generally speaking subjects answered with higher values at t_1). We also ran the models for the predictors of number of sessions attended as well as number of sessions attended live and number of sessions attended recorded and we see the same overall effect, see App. A.4.7.

To make sense of the negative shift in the quantitative results state that the participants rated themselves worse than at the start of the study, we found a large amount of evidence in the qualitative data that shows quite the opposite - that participants have become way more aware. The reason for the more critical self-assessment may well be a consequence of increased awareness, see Fletcher and Bailey [340] for details on issues with self-awareness assessment.

We move on the **qualitative results** for this question, answered by the responses coded under theme "main learnings (changes) identified by the participants" (see Tab. 4.5).

Theme: Main learnings (changes) identified by the participants. Participants described how they experienced the changes in their perception during the course, mainly in awareness, mindfulness, and attention. They commented how these changes influenced their relationship with themselves, with others, and with their environment.

Awareness. Participants reported significant changes on this construct. The participants mentioned enhanced awareness about themselves, identifying their breathing pattern during the day, for example.

> *"I noticed that I am more aware of my full breathing during the day and not only during yoga/meditation." (participant 31, run 2, journal, Feb 4 2021)*

The participants' reports focus mainly on the awareness of the body, as the previous quote mentioned and as it can be read in the section below.

> *"I will so much like to really feel all my sensations and be aware of my body, for instance, I still have a lot to learn, but I can see the progress over all these weeks, though." (participant 14, run 2, journal, Mar 30 2021)*

In this case, the participant commented that they had noticed progress in the process of awareness of sensations and their own body during the weeks that go from the breathing course. In addition to noticing changes in their perception of sensations, breathing, and the body, another participant wrote about the changes in their general needs and also actions taken towards those them.

> *"I think this experiment is making me more aware of my needs and doing what is nice for my body and soul." (participant 31, run 2, journal, Feb 6 2021)*

The changes in the daily perceptions of the participants were not limited to the body and the self; instead, they expressed that the changes during the course

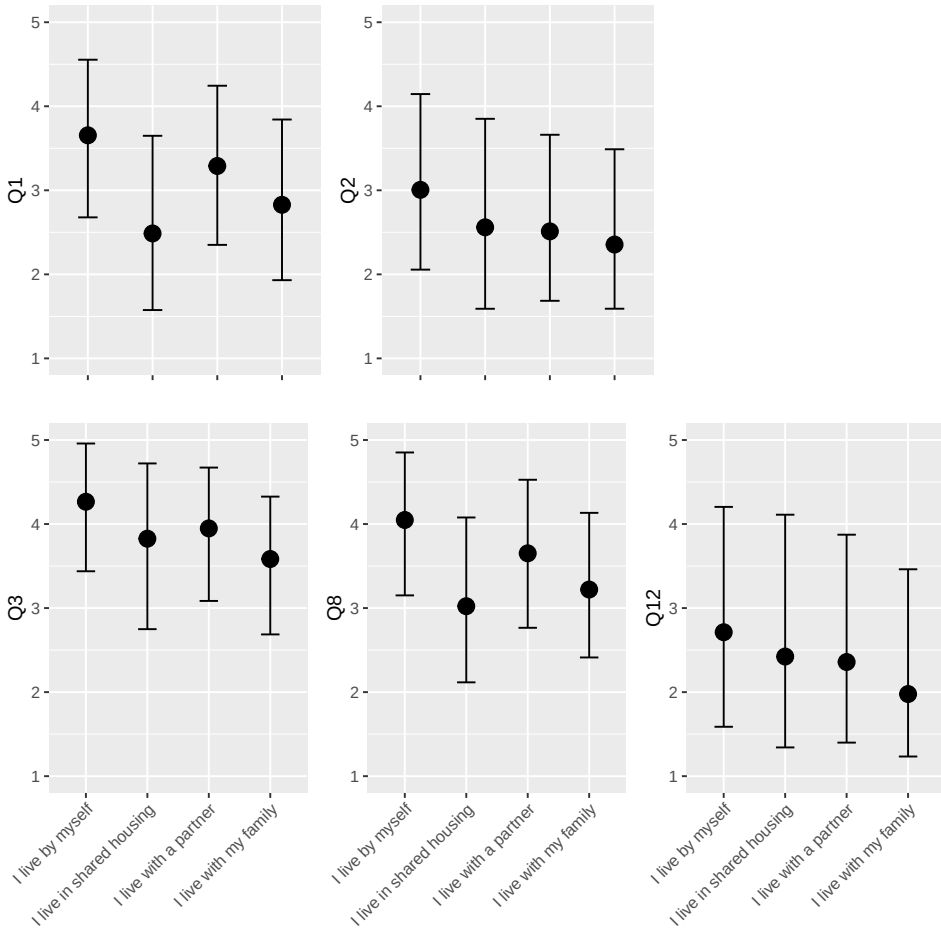


Figure 4.7: Conditional effects concerning the predictor living condition and its categories for the five MAAS ratings that were significant. By fixing all predictors at their mean or reference levels, a view of each category's effect for our predictor living condition is obtained. The bars correspond to the 95% credible interval, while the dot indicates the mean. One or more categories were significant, i.e., from left to right on the horizontal axis in MAAS Q1 the 2nd and 3rd, in Q2 the 4th, in Q3 the 4th, in Q8 the 2nd and 4th, and in Q12 the 4th category were significant and, hence, the main drivers for the predictor to be significant. The vertical axis shows the Likert scale values.

motivated them to put a higher score for their day-to-day life when writing the journal. The participant below describes in a very clear way the motivation to raise their daily score.

> *"I remember that I started this voting above with a 7, which was something like 'yes, the day was ok, I feel ok, nothing unusual happened' [...] I thought that a 5 or 6 is too low for that. But know I raised that up to an 8, because I'm aware of a change that happened in the last weeks." (participant 21, run 2, journal, Feb 16 2021)*

The tools that mainly led the participants to be more aware of their emotions and to being able to control them were analysis and reflection. The above quotes express the variety of changes in awareness that the participants experienced during the weeks of the course at different levels and areas.

For *consequences of being present*, the qualitative responses included more intentionality, better listening, better connection with people, better self-care, more relaxed, happier, less obsessive-compulsive, more alertness, higher effectiveness, calmness, softness, more creative, more productive, being relaxed and energized, more joyful, and feeling inspired.

Mindfulness. Participants also identified changes in mindfulness. Some were able to recognize them and even link them to other thought processes such as focus and the feeling of happiness.

> *"Meditation has helped me greatly, I have been able to focus more on the present moment, have more focus and feel genuine joy more often, and not only thinking that I am enjoying but having a divided heart and mind." (participant 14, run 2, journal, Mar 16 2021)*

In this quote, it is clear that the tool used is meditation; at the same time, breathing exercises were also mentioned as the way to work with mindfulness. In the same way, concentration and focus were vital for several participants in their mindfulness process.

> *"There are things I have no control on. When these things happen it is hard to concentrate on what I am doing. but reminding what is important at this moment and breathing is helping me recently to get my thought together." (participant 53, run 1, journal, Oct 28 2020)*

The previous paragraph illustrates how a participant, focused on mindfulness added to the breathing practice, manages to focus and organize their thoughts. The participant can also identify what makes it difficult for them to manage their concentration, which implies a process of awareness.

> *"In reflection, lots of similar experiences in the past were less pleasant, due to my lack of self awareness and poor mindfulness which allowed me to defuse disruptive behaviours triggered by stress (e.g., loosing focus, feeling insecure, cognitive overload, etc.)". (participant 82, run 1, journal, Nov 22 2020)*

This participant even compared how they lived their past experiences in a different state of awareness and mindfulness and concluded that stress was the reason for those behaviors. They also identified their emotions and behaviors. Likewise, they mention that it was thanks to the reflection that they managed to have this insight after participating in the breathing course.

Attention. The analysis also showed how the intervention influenced the attention of the participants. They explained how they experienced these changes in the quotes below.

> *"I did tonight the breathwork practice, and I feel emptier and lighter. My mind clearer, fewer thoughts, and more directed. Higher sensuality with my body, higher sensitivity to music." (participant 18, run 1, journal, Sept 24 2020)*

This participant explains the results of using the breathing techniques. They describe how their senses came into focus and clarified. They also express a feeling of lightness and increased sensitivity and connection.

> *"I will say, the breathing exercise guide you hosted really helped with clearing my mind from thinking far ahead and behind. I stayed present all throughout the day. I felt more at peace with myself. [...] Coming back into it with your guidance reminded me again of why it's so important. I am relaxed and willing to take on whatever task comes my way." (participant 32, run 1, journal, Sept 25 2020)*

This quote describes how the breathing exercises focused the participant's attention on the present, as they commented that they cleared thoughts of the future and the past. In the same way, it served as a tool for enhanced mindfulness and relaxation, motivating the participant to focus their attention on future tasks as mentioned.

The excerpts of the daily journal show that the participants carry out more than one thought process at the same time. Sometimes they are not aware of the combination of these processes. However, when describing their experiences, the combination of attention and, in some cases, also mindfulness is clear. Similarly, most of the participants' insights happened through reflection, which was motivated by writing every day. Keeping a journal pushed participants to rate their day and reflect on what was most important to them. That brings us to RQ1.2, the rating over participants experience over time.

4.5.2.2 How did the perceptions of life experience progress over time? (RQ1b)

Participants rated their day in the journal entries from 1 (Really bad) – 10 (Absolutely great). A positive trend is present for two thirds of the intervention, then dips back down towards the end, Figure 4.8 provides a visual overview. Given the uncertainty over time, we cannot claim a significant trend. While still showing an absolute improvement from beginning to end, the reversed tendency was significant enough to look into. We attribute the observed slight decline towards the end to two effects: 1) The newness of the intervention is wearing off and the end of the study is in sight. 2) There is a plateau effect after practicing for a while that shows up as a less positive rating of items. Later conversations with participants confirmed these hypotheses.

From the quantitative analysis we see a positive trend that is indicative but not conclusive. However, the qualitative data from the journal entries around the theme "Integration of techniques in daily life" support the positive trend as follows.

Attachment and Integration of techniques in daily life—Reflections. Participants described feelings, situations, and emotions that, through reflection, they noticed that they lived differently.

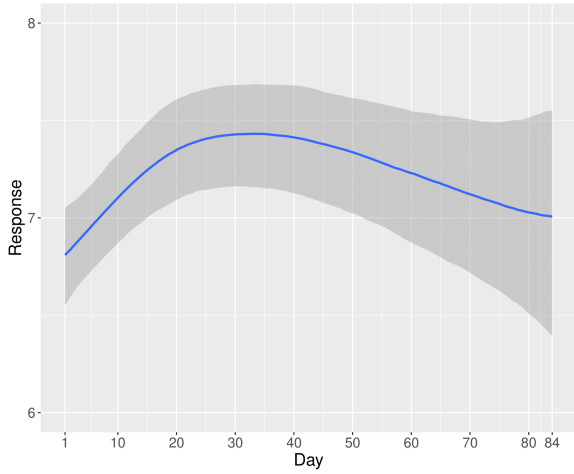


Figure 4.8: Trend for daily survey ($N = 1646$). The blue line indicates the median while the band signifies the 95% credible interval. On the vertical axis we have the response on Likert scale (1–10), while the horizontal axis indicates the day. Even though an initial positive trend is visible, due to the uncertainty (especially in the later part of the study) one cannot draw any conclusions.

> *"This daily journal exercise is making me feel really good about little things in the day that I might've otherwise forgotten... Rise 2 Flow is making me reflect!" (participant 10, run 1, journal, Sept 29 2020)*

This participant wrote how, through reflection, they were more aware of things going on during the day that, otherwise, they would forget. They also mentioned how to write every day about their day creates a feeling of good.

> *"I feel a lot of gratitud to the people and situations I am encountering. I feel more compassionate about other souls and I connect myself quicker than three months ago." (participant 18, run 1, journal, Dec 12 2020)*

This quote describes how the participant felt more compassion, a lot of gratitude, and enhanced their ability to connect to other people. They are aware of when this change started with the course.

> *"There are behavioral patterns of mine that I am trying to observe and see how I can make any changes in them. Previously I was so unconscious about them and I used to notice them long after my actions. I did one of them on Thursday and I recognized it right after." (participant 53, run 1, journal, Oct 1 2020)*

These participants wrote about the wish to observe and later change specific behavioral patterns. They mentioned that it was difficult to identify these behaviors when they took place; instead, they realized them later on after participating in the course. This participant expressed that they were able to locate said pattern right after it happened; this is a notable improvement in awareness that can be used in daily life to spot and improve behavior.

The recurring topics in the daily journal were work, stress, and family and friends. Regarding work, the participants identified it as one of the core causes of stress and, on some other occasions, as a measure of productivity. Stress was present in the descriptions mainly as a result of work or illness. Family and friendly relationships were the principal support that participants used to cope with stress.

Throughout the course, the participants explained how the way they handled stress and work changed, improving control and the feelings linked to them. As for family and friends, the changes focused on enjoying and valuing these interactions more.

Summary RQ1: We answer *"How did participants mindfulness attention awareness and daily perceptions of their experience of life change?"* with **yes, indicating an improvement**. While the survey instrument shows a negative trend, the qualitative data and the experiences described in the free-text answers of the survey and the journal entries paint a different pictures.

The qualitative analysis shows how participants focused their attention towards appreciation and gratitude, and became more reflective in reporting on situations that did not go so well. We observe a 'growing up' tendency in taking responsibility for their own experience of life and in choosing their focus. One participant concluded:

> *"Awareness and consciousness. Time seems to expand as I feel more effective in processing information and seeing connections. I can pinpoint parts of my body I hadn't realise were sending me signal, and my mind becomes more responsive to information and connections. It somehow becomes easy to notice subtleties and details in images, sensations, text, dialogues, etc." (participant 82, run 1, exit survey)*

Or, by another participant, stated more informally:

> *"Generally, things are good. Took a while to get here, but adulting has finally paid off." (participant 29, run 1, exit survey)*

4.5.3 RQ2: Changes in Well-being

Research Question 2 *"Does the intervention lead to change in the participants well-being?"* was composed by several subquestions listed in the following. Before we dive into answering each one of them, we present the overall qualitative findings on this topic.

Constructs modified during the course. The changes expressed by the participants span areas of well-being. Participants explain how various aspects of their lives changed throughout the course. One of these aspects is the way they lived their experiences, which were influenced positively. Some participants expressed perceiving the differences compared to before applying the techniques.

> *"I had a new presence experience today. Waking home today without any headphones I started to actually listen. It was a cool experience, and I really felt as though I could shake off the day, reload and come back with new energy." (participant 73, run 1, journal, Oct 13 2020)*

Similarly, this participant talks about how they perceived different an activity as simple as getting up in the morning. They also mentioned how a small change, not wearing headphones, causing a feeling of well-being and sensation of renewal of energy.

Regarding daily activities, the participants commented on how they organized their routine to carry out meditation and breathing practices. Likewise, they expressed the modification of habits, greater reflection, and mindfulness.

> *"I am very proud of that daily meditation/breathing practice in the morning before turning on the computer. I also changed my breakfast habits and was more present and reflective more. And besides, I'm motivated to take it to 9 someday. :)" (participant 21, run 2, journal, Feb 16 2021)*

The above quote shows the motivation and commitment of the participants to work to improve their well-being. Participants also wrote that they feel proud of their actions and set the goal of rating their day with a 9 in the future.

On the other hand, participants wrote about their emotions, as the quote below explains.

> *"By being present, I was able to overcome the dark thoughts that would have otherwise consumed me." (participant 52, run 2, journal, Feb 27 2021)*

This participant mentioned having better control of their emotions after practicing breathing. It is a recurrent comment among other participants. At the same time, it is linked to better awareness about feelings.

> *"I'm getting better in taking time for some well-being practice." (participant 21, run 2, journal, Mar 15 2021)*

An important point is that participants commit to integrating wellness practices into their routines. This commitment served as the basis for the changes they later expressed regarding well-being and other areas.

In addition to the previous changes, participants talked about how they perceive their thinking. They expressed they feel more positive.

> *"I did again the morning meditation, feel really happy to had a more fresh and positive perspective on the week and the days to come." (participant 87, run 1, journal, Oct 23 2020)*

The quote above shows how this participant feels after meditation. Happiness, fresh and positive perspective are the results they describe.

Gathering what was mentioned by the participants, it can be seen how they gradually perceive the changes during the weeks that the course lasted. They reflected on how their perceptions were modified and explain small experiences that they perceived differently. They better identify emotions and feelings and can deal with them in an improved way. They also mention how their moods changed. They feel calmer, at peace, and more in control of their emotions.

Similarly, the participants commented they realized that there are many ways to well-being. One participant wrote that they never tried dancing, for example, because they were "consistently failing on the things that were supposed to work", but now they are more open to trying different things. The content of the slides also influenced the changes in well-being. An example is "pick three things that are

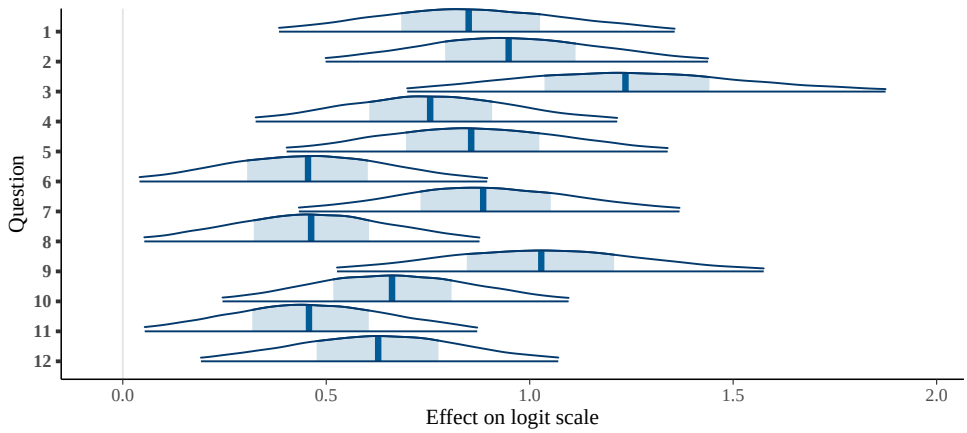


Figure 4.9: The effects of t for the SPANE instrument. The densities are cut off at 95% probability mass, the dark blue line indicates the median, and the light blue area is the 50% probability mass. The temporal variable t clearly has an effect (positive) in all questions, i.e., responses were generally speaking higher at t_1 (exit instrument).

important to you" a participant mentioned that by using this practice, they were able to realize the care they put in others but not on themselves. Finally, other reported improvements were to sleep more, less emotionally burdened overall, resting, and better stress management.

4.5.3.1 Does the intervention lead to change in the participants' perceptions of positive and negative experiences? If so, how are their experiences affected? (RQ2a)

The Scale of Positive and Negative Experience (SPANE) instrument consisted of 12 questions (see App. A.2.2), that asks one general question:

> *Please think about what you have been doing and experiencing during the past four weeks. Then report how much you experienced each of the following feelings, using the scale below.*

Responses are on a 5-level Likert scale and there are six categories of questions, each category having two contrasting question (i.e., positive/negative, good/bad, pleasant/unpleasant, happy/sad, afraid/joyful, and angry/contented).

All items were **significant (positive)**, i.e., higher responses at t_1, as can be seen in Fig. 4.9.

The same effects show for attendance of sessions as predictors, i.e. the more sessions a participant attended, the more increase in their SPANE score, see App. A.4.7. Significant effects of the other predictors were the higher the age, the higher the response in Q9. Concerning gender, males answered with higher values in Q3, Q6, and Q7.

Both the quantitative data (all rated items) and the qualitative data from the surveys and the journal entries shows a more immediate perception of positive and negative experiences as well as a general tendency towards a more positive perception of participants' daily lives. One participant summed it up as follows:

> *"It's crazy how much some of these modalities/ tools/ phrases/meditations/breathing exercises can really change a very negative mind! I'm so grateful that these gifts have come to me in such a timely manner, and that I can hold them to me for the rest of my life. Now to keep them present and alive, and in everyday use! [...] I have practiced the breathing portions quite often and have a real appreciation for the effect it has on my physical being." (participant 23)*

4.5.3.2 Does the intervention lead to change in the participants' psychological well-being? If so, how is it affected? (RQ2b)

The Psychological Well-Being (PWB) instrument consisted of eight questions (Likert 1–7, see App. A.2.2). All t parameters are **significant (positive)**, i.e., higher values at t1, except for Q3, The details are visible in App. A.4, Fig. A.3 for visualization, correlating with the number of attended sessions (see App. A.4.7. We found significant effects of the predictors for Age, Gender, Occupation, and Living conditions, see also App. A.4.

From the quantitative analysis we see that all except one item were rated higher by the end of the intervention, and the qualitative survey data shows participants have had good learning experiences around well-being. For example, one participant reports:

> *"I am happy that I am aware of my strengths and weaknesses, and that I am able and have learned in my life that we all are meaningful, have a purpose—although that might not be clear or visible most of the time. I am content with my place in life and I have grown to love how my being me makes other people seek my help, presence or comfort." (participant 45, run 1, exit survey)*

4.5.3.3 Does the intervention lead to change with regard to their positive thinking? If so, how is it affected? (RQ2c)

The Positive Thinking Scale (PTS) (App. A.2.2) consisted of 22 questions (Yes/No answers) and contained some reverse scored items to ensure instrument validity (as noted below). Questions 4, 9, 12, 15, and 17–18 showed a significant difference between t_0 and t_1 (Q9 and Q18 were positive). The details are provided in App. A.4, Fig. A.4.

From the quantitative analysis we see that participants did not necessarily change their mind about some parts of their lives or experiences they labeled bad, but there was a shift towards more positivity in a number of items. Overall, participants think more positively at the end of the intervention period. The qualitative data confirms this, for example:

> *Rumination/focus on past mistakes is a particular problem for me because I suffer with OCD, but I am trying to get better with dealing with it, e.g., letting thoughts simply pass through (as suggested through Rise2Flow). (participant 10, run 1, exit survey)*

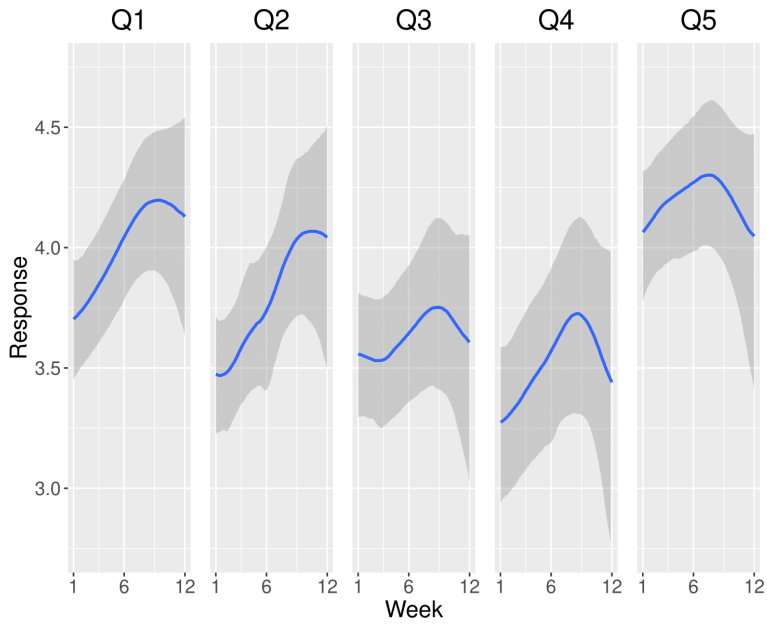


Figure 4.10: Trends per question for weekly survey ($N = 456$). The blue line indicates the median while the band signifies the 95% credible interval. On the vertical axis we have the response on Likert scale (1–6), while the horizontal axis indicates the week. In particular for Q1 and Q2 one can see a positive trend. Due to the uncertainty no conclusions can be drawn. For Q3–Q5 this is even more so the case.

> *I'm a positive person, and I try to embrace this at all levels. I find joy in finding silver linings and in life-long learning. (participant 84, run 1, exit survey)*

> *How much I've grown and my mindset has brightened since the start of the survey. (participant 42, run 1, exit survey)*

4.5.3.4 How does the well-being fluctuate and vary over the course of the intervention? (RQ2d)

For the weekly survey, the instrument is shown in App. A.2.5. Figure 4.10 provides a visual overview of the daily trends. One can see a positive trend for Q1, Q2 and Q4; however, the uncertainty makes it difficult to make any convincing claims. For Q3 and Q5, this is even more so.

In Fig. 4.8, one could see the same pattern for the daily rating as in Fig. 4.10 for the weekly trends for Q1, Q2 and Q4.

From the quantitative analysis, there is an improvement in the ratings of all items of the weekly survey. However, there are trends to be interpreted that raise curiosity. Houben et al. [332] confirmed that overall, low psychological well-being co-occurs with more variable, unstable, but also more inert emotions. "Not only how good or bad people feel on average, but also how their feelings fluctuate across time is

crucial for psychological health." [332] Therefore, it was important to us to look at the development of the WHO well-being scores over the period of the 12 weeks to see the extent and variance of fluctuations. In that development, we saw a similarly shaped curve in all five items of the weekly survey (see Fig. 4.10), with an increase over the first two thirds, then a peak, and then a decrease. Overall, the ratings were still better at the end than at the beginning, yet not necessarily significant. We attribute the decline towards the end to two effects: 1) The newness of the intervention is wearing off and the end of the study is in sight. 2) There is a plateau effect after practicing for a while that shows up as a less positive rating of items.

We observe that there is improvement that shows a positive trend with a specific pattern of fluctuation. The qualitative data shows progress in some individuals' development of well-being.

Summary RQ2: The research question around participants' well-being is answered with **yes, an increased well-being** was observable through both quantitative survey items and qualitative data.

4.5.4 RQ3: Changes in Perceived Productivity and Self Efficacy

Research Question 3 *What are the observable changes in perceived productivity and self-efficacy?* is answered by two subquestions in the following.

4.5.4.1 Does the intervention lead to change in the participants perceived productivity? If so, how is it affected? (RQ3a)

The HPQ instrument for Perceived Productivity consisted of eleven questions (with Likert scales varying, going up to 5, 7, or 10, depending on the question, see App. A.2.4).

Only Q1 (*How often was your performance higher than most workers on your job?*) shows a significant difference when moving from t_0 to t_1 (lower responses at t_1), which indicates that participants scored themselves lower in performance (which could be due to the continued pandemic working conditions). The details are visible in App. A.4, Fig. A.6. This effect confirms the finding indicated by the scores of the MAAS, where participants became more self-aware in general, also noticing when they are off.

Qualitatively,

at work was a recurring theme in the journal entries, especially as a cause of stress in the participants. The perceptions about productivity are what participants used to measure their job performance. The following lines explain how they perceived changes in the area, as mentioned earlier.

> *"I have insanely increasing workload at work. But am proud that I am able to stay calm today and focus not only on work but also on other positive things." (participant 92, run 2, journal, Feb 2 2021)*

The quote above, although brief, describes the positive changes in this participant, better focus, and calmness beyond work.

> *"The breathing practice helped me to get a nice state in my mind and reflects in things like today. More awareness made me feel like days are longer but productive." (participant 68, run 1, journal, Oct 29 2020)*

This participant linked awareness and the feeling of being productive even though he feels the days have been longer.

> *"All together, all the awareness of the program has helped me to be more focused and present, enjoying what I am doing, what I don't enjoy, and still finalize it without falling in the temptation of getting distracted by the first thing that pops in my mind (still happens, but is getting better)." (participant 87, run 1, journal, Oct 23 2020)*

The quote above explains how this participant has managed better the distractions around and became more focused and present. Besides the change in productivity, they mentioned that they enjoyed more the activities they are doing.

In conclusion, and based on participants' comments, a better awareness resulted in better productivity. The participants seemed to be more relaxed also, better focused, and function better.

4.5.4.2 Does the intervention lead to change in the participants' self-efficacy? If so, how is it affected? (RQ3b)

The Self-Efficacy (SE) instrument (App. A.2.3) consisted of ten questions (Likert 1–4). Questions 6, 7, and 9 showed a significant effect (positive), i.e., higher responses at t_1.

Q6 I can easily face difficulties because I can always trust my abilities.

Q7 Whatever happens, I'll be fine.

Q9 When a new thing comes to me, I know how to handle it.

The details are visible in App. A.4, Fig. A.5. Concerning the other predictors, no significant effects were present, i.e., it is not clear which predictors drove the significant difference between t_0 and t_1.

Qualitatively, self-efficacy was another area that changed in participants during the course. The following quotes show the perceptions of the participants in their performance in work and daily activities.

> *"I did the most important tasks (rocks) I had set for the day which made me feel great, so much better than the overwhelming feeling I often have when looking through my six page to do list." (participant 73, run 2, journal, Feb 2 2021)*

This participant is using one of the organization techniques presented during the course. They express the positive results of having implemented the rock and sand technique and how this translates into a feeling of well-being compared to similar situations in the past.

> *"Spending a whole day in online meetings but remained quiet and calm, not loosing patience while knowing that I was doing rather duties than something being passionate about." (participant 46, run 1, journal, Sept 23 2020)*

The quote above shows how this participant managed to remain calm and perform their duties even though they expressed they were not passionate about it. It is essential to mention that the activities performed by this participant are mainly online. This scenario may add more stress than usual on this participant; even so, they achieved to keep calm.

> *"I've become so much better at looking at what I actually accomplished rather than what I did not prioritize in order to achieve that." (participant 73, run 1, journal, Oct 5 2020)*

This participant has changed the way they address their activities. They decided to focus on what is achieved instead of anything else. With this, the feeling of accomplishment improves, and so the self-efficacy.

> *"I tried again the technique of taking five minutes and visualising who I wanted to be in that conference. I focused on trying to convey my excitement about my work, my good humour in connecting with my peers, my excitement to learn new things. I felt very successful in doing so. I noticed how my entire attitude changed and the day did not seem so exhaustive anymore." (participant 82, run 1, journal, Oct 25 2020)*

Learnings from the presentations previous to the breathing session also played a role in the changes in self-efficacy. This participant explains in the quote above how a visualization technique helped to perform better at a conference. They also noticed changes in their attitude that had an impact during the whole day.

Summary RQ3. We answer RQ3a, perceived productivity, for now with **inconclusive**. The quantitative analysis showed only one item with a significant change over time, which indicated slightly less productivity. However, the analysis of the qualitative data showed a lot of examples of how participants had improved their ways of working, scheduling and completing tasks.

We answer RQ3b, self-efficacy, with **yes, positively**. From the quantitative analysis, we observe three significant effects, all positive, towards more self efficacy, which is confirmed by the qualitative data in the journal entries.

4.6 Discussion

In this section, we discuss the findings, relate them to other work in the field, and then to a bigger societal picture. We point out limitations as well as various aspects to be taken into account for future work.

4.6.1 Significance

In terms of significance, the following main questions arise:

[a] How do the results compare to other modalities and the state of the art?

[b] What is the impact of an online setting versus in-person?

[c] What is the magnitude of the impact compared to other contextual factors?

State of the Art Comparison? The most well-known program to increase well-being and resilience for IT people is the mindfulness program "Search Inside Yourself" by Google [306]. It was started in 2007 by Chade Meng Tan. On their advisory board was, amongst others, Jon Kabat-Zinn (work detailed in Sec. 4.2.3). A spin-off leadership institute now teaches the program and certifies instructors. Like Rise 2 Flow, they combined a traditional introspective practice (meditation, pranayama) with scientific foundations and self-development topics. This program is what Bernardez et al.'s work [292] is based upon and evaluated with a software engineering population. Half of their student population practiced mindfulness during 6 weeks, and then the complete population participated in an experiment evaluating the efficiency and effectiveness in conceptual UML modeling. The effectiveness of the program in comparison to ours cannot be established as we did not carry out a UML modeling experiment, but evaluated broader concepts on established psychological scales.
As the breathwork technique of our study has not been evaluated empirically before, we do not have data to contrast software engineers against a more general population. We are aware of only one study that has compared a similar type of breathwork to other modalities [283], and it showed the greatest impact benefitting six outcomes (depression, stress, mental health, mindfulness, positive affect, and social connectedness). We see a similar outcome in our study in a decrease in stress and an increase in positive affect across the data, and an increase in mindfulness in the qualitative data. No study has yet compared the breathwork technique that we used in our study, so a more detailed comparison might bring insight in form of first data points for the usefulness of breathwork in general in comparison to other de-stress modalities.

Limited Impact in Online Setting. Interestingly, it was the lock-downs in many countries and subsequent mental and emotional challenges experienced by students, colleagues, family and friends that motivated us to carry out this study in the first place. The online setting was chosen such that we were able to offer a relieving intervention during the restrictions. A traditional setting would be a designated physical location where participants meet once a week in a safe space that is specifically prepared for an undisturbed session without distracting technology or disruptions from outside. Such a setting allows for an immersive experience on a different level and, often, much deeper transformation and restoration. During Rise 2 Flow, participants dealt with network service outages, software updates, and usually practiced in their living room turned make-shift office turned trying-to-be attention-restorative environment. On one hand, this certainly limited the benefits that could be received through this modality and, on the other hand, it opened the intervention to a much wider group of participants from around the world.

Magnitude of Impact Software developers create the most complex systems in the world, and need a high attention capacity for intellectually taxing tasks in often distributed team constellations. Furthermore, they need empathy for collaborators and clients alike and usually work under time pressure. This study looks into the probably most accessible mechanism to regulate and restore the nervous system - breathing. While the study focuses on a specific framing as intervention (the whole **program** with the breathing plus reflection practice for specific topics), the **breathing technique by itself** can also be used individually for a few minutes of

reset in any situation. Participants are trained on a technique that helps both to increase resilience by building long-term capacity (see results for well-being Sec. 4.5.3 and self-efficacy Sec. 4.5.4.2) as well as short-term recovery (see quotes from journal entries, e.g. p. 129).
An influence of season and weather may be present, but cannot be analyzed to a meaningful degree with the collected data as we had participants from all around the globe.
The restrictions in our daily lives due to the Covid may have led to a higher stress level for many participants, which was reflected in many survey answers relating to working from home with implications of either loneliness or taxing family situations. Despite this effect, it also showed the intervention was timely and useful, as mentioned by these participants:

> *This rating makes me think about where I'm at in life and how I view myself within my surroundings and social community. Reflecting on social relationships gave me pause because they are supportive and kind, but at times I feel so alone. This is largely because of the pandemic and stay-at-home living. I am still grateful for them and can accept the tragedies as well as the beauties of these new living circumstances. I am also realizing that I can be more engaged in the activities that mean a lot to me; it's easy for me to detach when I feel overwhelmed. I'm reminded of the importance of prioritizing rather than letting everything go.* (participant 42, run 1, exit survey)

> *It's interesting to do this exercise during a global pandemic in California... We are living in such difficult times in so many ways right now. Interesting to reflect on whether the past was 'good' or 'bad', considering the current situation. Like, in comparison, the past seems like it should have been so much joy all the time—but of course we took being close together in groups and hugs for granted back then. This makes me sad. I do find myself savoring and really appreciating time with friends and family more now than in the past, even though times are hard right now.* (participant 67, run 1, exit survey)

4.6.2 Observations and Implications

Policy. There are implications for policy and resulting applications of well-being indicators, as society becomes more aware of the importance of mental health. For in-depth discussion, see Pavot and Diener [341] who call for a national well-being index to inform policy makers specifically in the area of aging, as a first step towards a happiness index like already established in other countries. While advertising for the study, we observed a general acknowledgement of the importance of supporting well-being. At the same time, it seemed that often the support did not go beyond the acknowledgement. Mental and emotional health can only improve if individuals, organizations and institutions alike take responsibility.

Psychographics versus Demographics. Before the study, we had been wondering whether personality traits would show a difference in how participants benefit from the practice and how their awareness shifted. Therefore, in the exit survey, we added a Mini IPIP personality test[21], so we could control for personality in the results. However, IPIP did not reveal anything in the analysis. This does not mean that

[21]While we are aware that IPIP is by now considered controversial in terms of its statistical validity, to this date it is still the most widely and commonly used personality test.

personality has no influence; it only means we did not see conclusive evidence for a particular personality trait as observable via Mini IPIP.

Role models. As senior academics, we are role models—whether we want it or not—simply because we speak in front of students, we teach and supervise. If we do not model taking care of our nervous systems, including physical, mental and emotional health, we not only neglect ourselves, but also fail to provide our students with guidance. We are not advocating for every senior academic to give lectures on the topic. Instead, we advocate for every single person prioritizing their well-being over external demands so we can operate from a well-resourced place and thereby deliver better service to the world. Living into that as a role model is a more effective way of teaching it than postulating the theory. Our participants commented on this as well:

> *"It was cool to see that profs were participating in that study as well."*
> (informal conversation with participants 77 and 78 after Rise 2 Flow 2)

4.6.3 Limitations and Threats to Validity

Sampling Bias. The participants for both runs of the experiment were recruited across a range of personal and online networks, including the global personal network of the first author, university networks, mailing lists, online spaces, and social media channels. While it is not a classical convenience sample because of the number of channels used for broadcasting, it can be seen as an extension thereof [309]. However, all these networks are initially based on connections to the first author, which introduces a potential sampling bias. We mitigated this threat to the best of our ability by requesting re-posts and further distribution of the call for participation in the several hundred emails and posts the first author put out for recruitment across the disclosed variety of channels. We follow the reporting guidelines proposed in [309].

Self-selection Bias. It is possible that people who are drawn to participate in a study like ours are not a representative sub-population of the overall study population. By repeating the experiment and learning more about the participants through the surveys and follow-up interviews we aim to learn more about that aspect.

Response Bias. We used standard validated instruments in our survey that prevent response bias to the degree possible, following recommendations by Dillman et al. [342] by, for example, putting content questions before demographics. A remaining response bias is due to the fact that the participants got to know the first author as instructor, which may have introduced a bias in their free-text responses. We mitigated this by letting participants know, during sign-up, that their data was going to be anonymized before analysis.

Construct Validity. We used validated scales as referenced. We focus on the breathing element of the intervention, but there were other modalities offered for the days in between if participants wanted more, namely guided meditations and journal reflection prompts.

Measurement Validity. There might be a threat to measurement validity by participants getting tired while answering the 78 items on the entry/exit survey. Our pretests showed that the survey could be completed within ten minutes.

Internal Validity. There is a threat to internal validity as we had sessions that combined reflective conversation in the larger group with the breathing practice. As it is important for continuance of a course to build a relationship with the participants, there is no way to differentiate how much the community aspect may have contributed to the positive effects of the breathing exercises. In statistical language, given data, it is hard to control for a 'community effect'.
There is a threat to validity in terms of the effects of the breathwork in comparison to the effects of the topic conversations. The awareness raising is happening on a neurological/unconscious level by breathwork, and on a mental/rational level by topic presentations. Most participants liked both, and a few preferred only one of the components - for those participants there could be a stronger influence from one of the program components over the other. To the best of our knowledge (from being an instructor for years as well as from the qualitative data) we see that the main changes are coming from the breathwork for most people, and that the topics do have an impact that is minor in comparison.
A few participants reported later on to have performed breathwork in additional practice, which may have influenced their overall results. To the best of our knowledge, there were only few of them that did additional practice (e.g. for a few minutes before falling asleep after a stressful day), so the threat to validity is minimal.
We did not have a control group for several reasons. We tried establishing a control group in the first pilot (see Sec. 4.3), and got zero responses. It can be done with a waitlist approach where future participants serve as comparison group (as done by Bernardez et al. [292]), but that introduces a number of biases as well, for example that respondents are self-selected and in favor of trying the approach as opposed to a random control group, and that they are primed for the surveys by the time they participate in the intervention. Consequently, it is controversial (also in medical studies) whether this is a good approach. If we were to select a random control group, there would be no 'placebo' to mask whether participants receive the actual intervention under research or a different one. Bernardez et al. [291] did that in a one-session experiment to test whether cognition and concentration increase after a meditation session compared to a session about how to give good presentations, but this approach is less feasible in a 12 week intervention. A non-equivalent control group post-test-only design [343] was not feasible either. Given the drawbacks of the ways of how to work with control groups in this case, we do not think this would strongly increase the confidence in the results.

Conclusion Validity. The threat to conclusion validity brought about by potential researcher bias is mitigated by correlating insights from quantitative and qualitative data. External validity and generalizability are limited to the demographics of the participant population. Reliability is provided by using validated instruments and standard methods as well as a replication package.

The qualitative data we report on reads positive. That raises the question of whether the thematic analysis was carried out in a balanced manner. We do not have

any reports on negative or unwanted effects. Most likely, people who did not have the desired effects stopped reporting - and we cannot conclude on causality either way. We cannot report on evidence we do not have, therefore we are open about analysing what was there and that people who did not feel desired benefits yet dropped out along with the people who had urgent other matters come up (as quoted in Sec. 4.4.1).

4.6.4 Relating Back to Theory in Psychology

This study aimed to research the effects of breathing practice on the mindfulness attention awareness, well-being, self-efficacy and perceived productivity of computer workers. The results reported positive changes in these areas. Participants expressed after the breathing workshop to feel more relaxed, calmer, and more in contact with their emotions and, in some cases, with other people. Some of them also mentioned noticing improvement in areas such as creativity and their performance at work.

According to Beck's theory [344], the influence of breathing exercises on the participants results in a change in mood and, therefore, in a behavior change. This change was manifested by themselves when commenting that they can better manage their emotions, identify their thoughts and be more present at the moment (Sec. 4.5.2/4.5.3).

Fisher [345] mentions that the state of well-being is achieved by integrating seven basic skills into daily life.[22]

He suggests that a uniform balance between them creates the optimal conditions for well-being, and breathing practices support all of these. This theory emphasizes that the exercise of the skills that form well-being occurs when the individual interacts with their environment in social relations. Stress plays an essential role in these interactions; individuals will bring their moods and feelings to these interactions, qualifying their own experiences and those of others.

Considering what both theories mention and the results observed in the participants, breathing techniques contribute to the development and maintenance of personal and partially social well-being.

4.6.5 Future Work

We are planning several follow-up studies, namely a larger cohort in a company and a simplified version with less weekly time investment.

[22]

[a] Engage in sustained, constructive, self-controlled goal-directed activity within complex social environments;

[b] Respond constructively to social challenges;

[c] Engage in self-controlled, creative, goal-directed activity;

[d] Engage in and enjoy positive, reciprocal social relationships;

[e] Engage in present-focused activities of a sensory, meditative, creative, playful or aesthetic nature;

[f] Achieve a balance between the demands of socially engaged, goal-directed activity and other kinds of activity; and finally

[g] Understand the nature of wellbeing and the social and environmental conditions required to attain it [345].

Larger Cohort. We are currently looking for an industrial collaboration partner who can contribute with a larger sample from within one company, so we can see the distribution of effects in a similar work environment. Providing the study in a closed program advertised for in the company and supported by the company would increase the likelihood of employees sticking with the program instead of dropping off. For that study, instead of the original PWB, we consider using Dagenais' [346] version of the instrument in a study with larger samples from a single company or similar companies as opposed to the wide range of work contexts of the subjects in the study at hand, as they suggest it may be useful to adapt the instrument to make it more specific to evaluate a well-being in a work setting [346]. Their PWB seems to have a strong eudaimonic connotation from the survey participants' point of view.

Simplified Version. There was feedback on the level of involvedness required for the study, e.g., the 90 minute live session plus several surveys, so we are considering a reduced version. The questions to solve here are (1) how much can we slim it down with it still being a meaningful intervention, and (2) how much can we simplify the data collection but still get meaningful data. Most clinical studies get their participants from therapy interventions, and therefore have large numbers and wait-listed control groups because those people are in sufficient mental and/or emotional pain to act on it. However, if we offer to intervene before that pain becomes too dire, the intrinsic motivation may also be lower. If the reader ever dropped off a well-being practice after things got a little better, they can relate.

Further Instruments We are additionally interested in exploring the scales for mystical experiences used by John Hopkins hospital [347], as several participants indicated mystical experiences during the sessions.

Furthermore, as self-connection is the main foundation for relation to others and effective information flow (specifically important for software engineers), especially during Covid [348], we are also interested in working with the self-connection scale by [347].

4.7 Conclusion

In this article, we presented the results of an intervention with live group breathing practice to deepen the participants' connection to themselves, framed with a weekly self-development topic. Awareness raising is happening on a neurological/unconscious level by breathwork, and on a mental/rational level by the topic presentations and reflecting upon them in group conversation as well as in personal practice with proposed tools. The quantitative and qualitative results indicate that this intervention may be helpful in improving participants' mindfulness attention awareness, well-being, and self efficacy.

There is a wide selection of wellness classes available outside of work for the person looking, while at work there may be a few generic offerings that work on a content level, but often not on a neurophysiological or embodied level.

Software engineers have a strong background in rational thinking and work with empirical evidence, so there is a need for programs with adequate language such that

software engineers who feel overwhelmed are attracted - science-based and in a safe space, brought to them by someone who can relate to their specific work experiences. This may help sway hesitant software engineers to try out a relaxation and recovery technique, benefitting their personal resilience and well-being and, in turn, their work performance and job satisfaction (important for retention). Consequently, we see three ways of potential impact by our study: 1) to inform and raise awareness in the research community as well as in practice, 2) to train further cohorts of software engineers and software engineering researchers and educators in restorative practices, 3) to develop tailored programs for companies and higher education that teach these techniques and frame them science-based while still focusing on the embodiment component to increase self-connection.

The main challenge that remains is that the pace of work life is artificially high because of a perceived need for constant competition (e.g. time to market, to offer better service, to increase our skills, etc.) as remarked by several participants in our study to the point where they felt they didn't "have time" for restorative practices. The speeding-up of life we have been witnessing over the past decades has consequences for health. In a certain pattern, physical stress is healthy and makes sure that we get certain things done - and those phases of stress needs to be taking turns with phases of recovery (beyond sleeping 6 hours per night). When recovery is not sufficiently given, stress wears on our physical (adrenal fatigue), mental (burn-out), and emotional health (depression and anxiety). Restorative practices can help us recover more quickly and become more resilient - they do not change the underlying systemic misalignments.

Our vision is that restorative and contemplative practices can support us in recovering a stronger connection to self, such that we have the mental and emotional capacity to reflect on our values and how we live into them. We get to decide every day how we want to continue, and the constraints can be shifted, some immediately, some over time. There are systems with unhealthy dynamics in place, yes, and we can change them – because we humans are the ones that created them. If we don't like the constant stress and time pressure, let's change the systems and societal structures that create them. Part of that is acknowledging the tendency of the human mind to always want more (and we see how it plays out in our economy), and developing our own practice to stay present with that [349]. The first step towards that from the perspective of our research is: **Let's normalise taking care of our nervous systems as much as brushing our teeth, and thereby improve our physical, mental, and emotional health.** There could be a start of every meeting with a deep breath to become present, someone teaching peers an emergency breathing technique to relax and focus before a presentation, there a well-being course that teaches breathing practices (or other restorative techniques) twice a year at a company, a weekly meditation group that provides community support in addition to daily personal practice (when it comes to personal practice, 5 minutes is always better than nothing). The options are many, the prioritization is an individual choice.

We leave you with a quote from a journal entry that sums up results reflected for a number of participants and that seem worth acknowledging:

> *Today was the last day of the 12 weeks. I took away a whole new world, that I am still trying to reconcile with. (...) Anyway, learnings: be conscious where you put your attention, and hence your energy, what the wonder precious moment is, that I*

am not different—I am unique, to put intentions to things, how meditation with a intention/visualization can change your day, that breathing can "make you float" and have psychedelic experiences, the forgotten joy of dancing, the power of gratefulness and that I am grateful for the bad stuff that happened to me (!), how important it is to love and be kind to oneself, to surrender to feelings rather than pushing them away, the power of small routines (as well as the difficulty of keeping them), that I am not my thoughts or my emotions (what the f+@?!?!), (...) What else can I say, really? THANK YOU!!!* - participant 75, run 1, journal, Dec 10 2020

4.8 Data Availability

To support open science, the replication package including the raw quantitative data is available on Zenodo `https://zenodo.org/record/5082388`, which links to a Github repository `https://github.com/torkar/rise2flow`.

The qualitative responses are not available as many of them reveal very personal experiences, deep emotions, and individual life circumstances that might involuntarily disclose identifiable information.

4.9 acknowledgements

We thank the participants of Rise 2 Flow 1 and 2 for their trust in us to support them in cultivating a personal practice for increased well-being, for their dedication, and for their generous feedback. The first author thanks Robert Feldt for a helpful discussion of available survey instruments during the design phase of this study, and Sabine and Fritz Penzenstadler for helpful input in conversation and action. We thank Francisco Gomes de Oliveira Neto and Leticia Duboc for thoughtful feedback on earlier versions of this manuscript. We thank the anonymous reviewers who gave very thorough and thoughtful feedback on an earlier version (shout-out to especially Reviewer 1). We appreciate you.

The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973. Part of this research is financed by the Area of Advance ICT at Chalmers University of Technology under no. C-2019-0299.

Conflict of interest

The authors declare that they have no conflict of interest.

Chapter 5

Paper D:

Evaluating the Impact of a Yoga-Based Intervention on Software Engineers' Well-Being

C. Martinez, B. Penzenstadler

In the International Conference on Evaluation and Assessment in Software Engineering (EASE), 2025

Abstract

Software engineering tasks are high-stress and cognitively demanding. Additionally, there is a latent risk of software engineers presenting burnout, depression and anxiety. Established interventions in other fields centred around attention awareness have shown positive results in mental well-being.

We aim to test how effective a yoga intervention is in improving general well-being in the workplace. For that, we designed, implemented and evaluated an eight-week yoga programme in a software development company. We used a mixed-methods data collection, using a survey of six psychometric scales, pre- and post-intervention, and a weekly well-being scale during the programme. For method triangulation, we conducted a focus group with the organisers to obtain qualitative data. The quantitative results did not show any statistically significant improvement after the intervention. Meanwhile, the qualitative results illustrated that participants felt better and liked the intervention.

We conclude that yoga has a positive impact, which, however, can easily get overlaid by contextual factors, especially with only a once-per-week intervention.

5.1 Introduction

Stress is an increasing concern in modern society, with work-related stress being particularly prevalent in high-demand environments. This is especially true in fields like software engineering, where mental workload, strict deadlines, and extended periods of sedentary behaviour contribute to burnout [270], and reduced well-being [87].

Work-related stress is detrimental to workers' psychological health and costly to society. A broad analysis by the American Institute of Stress [**?**], factoring in absenteeism, turnover, reduced productivity, and higher medical and legal expenses, raised the estimate to $300 billion annually. Regarding Europe, Shaholli et al. [**?**] reviewed international studies and organisational reports to estimate the economic impact of occupational stress. Their findings reveal estimates ranging from €54 million to €280 billion, depending on the country.

Mindfulness practices have proven beneficial in demanding, high-stress work settings that require intense focus. Leading tech companies, including Intel, Goldman Sachs, Google, and SAP, have widely embraced it to promote employee well-being [350–352].

Mindfulness-based programmes have been implemented in different contexts, generally getting positive results. Penzenstadler et al. [19] did an online intervention using breathwork to improve general well-being and reduce stress in participants. Their results were positive, with an increase in attention and positive thinking. Further, Montes et al. [107] elaborate from a qualitative perspective on a similar intervention, sharing participants' positive perceptions of their course. Few interventions are done in the context of software engineering workplaces; for example, Bernardez [91] studied the effect of mindfulness practice, meditation, on a sample of 56 helpdesk employees working for a consulting and information technology company. Their participants significantly improved attention awareness. Heijer et al. [90] studied the impact of mindfulness on agile software teams in over two months of stand-up meetings with 61 participants from eight companies. The findings showed improved perceived effectiveness, decision-making, and listening.

This study is among the first ones carried out in a workplace setting with a software engineer population and using standardised scales to measure the effects of yoga as a mindfulness practice.

In this study, we aim to answer the following research question:

How does a workplace yoga intervention impact the general well-being of software engineers?

We approach this through a quasi-experiment mixed-methods design, using psychometric instruments to measure pre- and post-intervention well-being. We use six psychometric scales complemented by qualitative data from focus groups to provide deeper insights into the participants' experiences.

The paper is organised as follows: Section II reviews the related work on the benefits of yoga, specifically Hatha yoga, software engineers' well-being and existing interventions. Section III outlines the study's methodology, including participant recruitment, intervention design, and data collection and analysis procedures. Section IV presents the findings, and Section V discusses the results, limitations, and implications for practice. Finally, Section VI concludes the study and offers directions for future research.

5.2 Related Work

Yoga is an ancient Indian practice designed to "still the fluctuations of the mind" and facilitate meditative absorption, a psychological state marked by feelings of self-transcendence and unceasing happiness [353]. A regular yoga practice can improve strength, flexibility, and balance; reduce stress; and provide many therapeutic benefits [354]. The most common style of yoga practised in Western countries is **Hatha yoga**, which includes synchronised movements through postures with breath, meditation, breathing exercises, and supine rest to conclude [355]. Hatha Yoga is classified as a mind-body exercise (along with Tai Chi, Qi Gong, Pilates, and others) and a type of complementary and alternative medicine that has become a popular and effective form of exercise because of the numerous health and fitness benefits associated with a regular practice.[1]

5.2.1 The Effectiveness of Yoga in General

A number of meta studies have collected evidence on the positive effects of yoga practice:

Ross et al. [60] used the keyword "yoga," on PubMed and yielded 81 studies that met inclusion criteria. These studies subsequently were classified as uncontrolled (n=30), wait list controlled (n=16), or comparison (n=35). The most common comparison intervention (n=10) involved exercise. In the studies reviewed, yoga interventions appeared to be equal or superior to exercise in nearly every outcome measured except those involving physical fitness. The studies comparing the effects of yoga and exercise seem to indicate that, in both healthy and diseased populations, yoga may be as effective as or better than exercise at improving a variety of health-related outcome measures [60]. This empirical evidence is important to show the feasibility and likelihood of success of using yoga as mode of intervention in the study at hand.

Cramer et al. [356] searched Medline/PubMed, Scopus, the Cochrane Library, PsycINFO, and IndMED for randomised controlled trials (RCTs) of yoga for patients with depressive disorders and individuals with elevated levels of depression were included. Twelve RCTs with 619 participants were included. Despite methodological drawbacks of the included studies, yoga could be considered an ancillary treatment option for patients with depressive disorders and individuals with elevated levels of depression [356]. A similar study was conducted by the team of authors on yoga for anxiety. Eight RCTs with 319 participants (mean age: 30.0–38.5 years) were included. They conclude yoga might be an effective and safe intervention for individuals with elevated levels of anxiety [357]. Since software engineers have a comparatively high likelihood to develop anxiety and/or depression disorders over the course of their career [143], this evidence is of much interest for the study at hand.

To compare different yoga styles of practice, Cowen et al. [**?**] had twenty-six healthy adults aged 20–58 (Mean 31.8) participate in six weeks of either Astanga yoga or Hatha yoga class. Significant improvements at follow-up were noted for all participants in diastolic blood pressure, upper body and trunk dynamic muscular strength and endurance, flexibility, perceived stress, and health perception [**?**]. The

[1] `https://www.nccih.nih.gov/health/yoga-effectiveness-and-safety`

improvements differed for each group when compared to baseline assessments. The astanga yoga group had decreased diastolic blood pressure and perceived stress, and increased upper body and trunk dynamic muscular strength and endurance, flexibility, and health perception. Improvements for the hatha yoga group were significant only for trunk dynamic muscular strength and endurance, and flexibility. The findings suggest that the fitness benefits of yoga practice differ by style [**?**]. The next section hence details the benefits specifically evidenced in Hatha yoga, which is the style practised in our intervention.

5.2.2 The Effectiveness of Hatha Yoga

For specifically Hatha Yoga, there are two meta analysis studies.

Hofmann [358] carried out a meta-analysis that identified 17 studies (11 waitlist controlled trials) totalling 501 participants who received Hatha yoga and who reported their levels of anxiety before and after the practice and found them reduced [358]. Furthermore, Huang et al. [359] implemented a quasi-experimental design with 63 female community residents in New Taipei City aged 40–60 years, where the Perceived Stress Scale revealed significantly lower scores after practice [359]. Again, due to the often high levels of stress experienced by software engineers, these studies promise Hatha yoga as a beneficial intervention.

Luu et al. [355] searched MEDLINE, Scopus, and PsycINFO databases for experimental studies testing the effects of Hatha yoga (acute bouts, short-term interventions, longer-term interventions) on executive function (EF). A total of 11 published studies revealed that Hatha yoga shows promise of benefit for the EF in healthy adults, children, adolescents, healthy older adults, impulsive prisoners, and medical populations (with the exception of multiple sclerosis) [355]. Given the complexity of cognitive tasks that software engineers carry out, the benefits of the executive function strongly support the choice of Hatha yoga as a well-being intervention.

5.2.3 Consequences for job performance and outcomes proposed

To assess the evidence regarding the effectiveness of yoga programmes at work, Puerto et al. identified 1343 papers, of which 13 studies met the inclusion criteria. Nine out of 13 trials were classified as having an unclear risk of bias. The overall effects of yoga on mental health outcomes were beneficial, mainly on stress. The findings of this study suggest that yoga has a positive effect on health in the workplace, particularly in reducing stress, and no negative effects were reported in any of the randomised controlled trials [360].

The dissertation by Daane [361] investigated yoga as a means of increasing job satisfaction in the workplace. Her sample of 32 yoga students was surveyed on yoga practice, exercise habits, past yoga experience, and levels of job satisfaction. It was predicted that students who had practised yoga would have increased levels of job satisfaction. Results of an independent samples t-test did not support the proposed hypothesis. Similar to this study, our results did not reveal an improvement in the participants' personal well-being.

5.2.4 Moderating Factors and Conditions for Mindfulness

Mindfulness meditation can be an on-the-spot intervention in workplace situations [362]. Hafenbrack identifies three necessary conditions for an on-the-spot mindfulness intervention to be effectively used: Employees must be aware that they are in a problem situation, they must be aware of on-the-spot mindfulness intervention as an available tool, and they must actually engage in the meditation. Hafenbrack also describes the limitations of such engagement: It is possible that some people gain less benefit from meditation than others, e.g., defensive pessimists disproportionately harness anxiety to motivate themselves to prepare for future challenges. On-the-spot mindfulness meditation may thus have more detrimental effects on their performance than for individuals who do not employ that strategy. There are also differences across national cultures in how people conceptualise time and the ways in which they are judgmental towards others. These factors may moderate the relationships between different forms of mindfulness and various outcomes [362].

5.2.5 Well-being Interventions in SE

Among the few well-being interventions in software engineering (SE) based on mindfulness practices, findings have shown positive outcomes. Penzenstadler et al. [19] ran a series of breathwork interventions with computer workers and found their well-being increasing over the course of the intervention, both qualitatively and quantitatively. Similar to that intervention, we used pre- and post-surveys and complemented them with qualitative data.

Heijer et al. [90] studied the impact of mindfulness on agile software teams in a two-month intervention, where mindfulness was practised for three minutes during stand-up meetings with 60+ participants from eight companies. The findings showed improved perceived effectiveness, decision-making, and listening. However, a limitation was the use of non-standard questionnaires. Additionally, Bernardez et al. [91, 92] conducted a series of studies on mindfulness for software engineers, showing that these interventions have positive effects on their mental well-being and self-perception.

In the article at hand, we present the **first study on using the modality of physical yoga poses, called yoga asana, with a software engineering population**.

5.3 Methodology

This section explains the design, data collection and data analysis of our study. Additionally, it also elaborates on more methodological details.

5.3.1 Research Design

This study followed the quasi-experiment mixed-method design since our main goal was to explore whether the yoga intervention positively impacted software engineers' general well-being (measured by psychometric instruments). Based on Maciejewski [119], quasi-experiments are observational studies where participants self-select to be included in an intervention (lack of randomisation), and there is a lack of a control

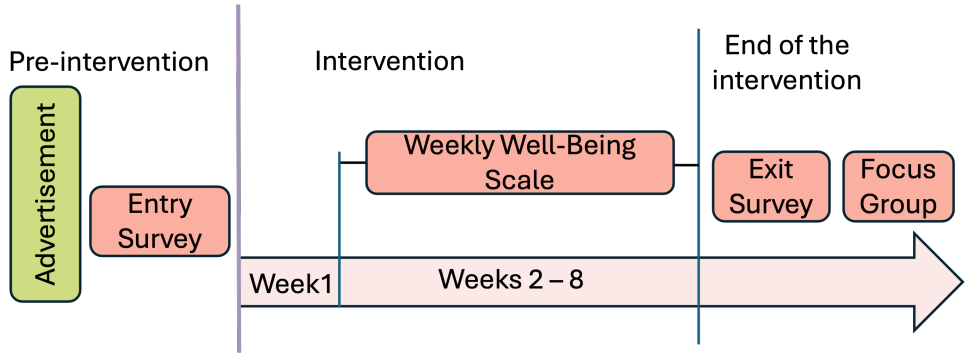


Figure 5.1: General Methodology of the Study

group. For this research, our participants were recruited by invitation from one company; however, they decided if they wanted to participate. Additionally, our (initial) control group was formed through participant self-selection. Further, our study employed a mixed-method approach, gathering qualitative and quantitative data.

5.3.2 Intervention

The intervention started with the invitation to participate in the programme. Later, participants received a link to the entry survey (including the informed consent). The programme lasted eight weeks. Participants had a 45-minute Hatha yoga session every Wednesday from 8:00 AM to 8:45 AM imparted by an experienced Yoga instructor. These sessions focused on the principles of Hatha yoga (5 min), incorporating physical postures (30 min), breathing exercises (5 min), and relaxation techniques (5 min).

Additionally, participants receive a weekly reminder and a link to fill in a weekly well-being scale to measure their well-being. After eight weeks, participants completed an exit survey. When the intervention concluded, we invited participants to give an interview. However, due to the lack of positive answers, we decided to have a focus group with the organisers who were also participants in the intervention. Figure 5.1 is the visual representation of the intervention.

5.3.3 Company, Population and Inclusion Criteria

The intervention occurred in an AI and software company dedicated to developing a comprehensive software stack for autonomous driving and advanced driver-assistance systems. The company has 501–1,000 employees, and its culture appears to be people-centred, with a strong emphasis on values-driven behaviour. Our target was software developers in general, so the advertisement was sent out to everyone in the company. The call was shared in the company's Slack space, and there were posters with the invitation in the elevators and shared coffee kitchen spaces.

5.3.4 Data Collection

This section describes the three data collection strategies we used to gather data. See Figure 5.1 to visualise the flow and organisation of our data collection process.

5.3.4.1 Entry and Exit Survey

We tailored an entry and exit survey composed of six psychometric instruments.

The first part of the survey asked participants to choose their alias and if they already have a personal well-being practice.

Later came the psychometric instruments that integrate the survey. We considered several areas that compound individual well-being to get a complete picture. Those areas were emotional well-being, which refers to understanding and managing feelings (SSEIT). Resilience (RS-14) since it is a significant psychological predictor of well-being [363]. Coping strategies (BRCS) strongly relate to positive physical and psychological health outcomes in stressful circumstances [364], leading to better long-term well-being. Self-perceived Success (The Flourishing Scale) measures an individual's self-perceived success and optimal functioning across all areas of life, reflecting the core elements of overall well-being [365]. Self-regulation (SSRQ) since higher self-regulation is linked to greater psychological well-being, including growth, purpose, relationships, and self-acceptance [366]. Finally, Self Transcendence (Self-Transcendence Scale (STS)) to measure the ability to derive a sense of well-being through cognitive, creative, social, spiritual, and introspective avenues [367, p. 1].

The Schutte Self Report Emotional Intelligence Test (SSEIT) [146] is an instrument to measure emotional intelligence developed by Dr. Nicola Schutte and her colleagues in 1998. The authors used the model of emotional intelligence of Salovey and Mayer as the conceptual foundation for the items used in the scale. It contains 33 items and uses a five-point Likert scale going from "strongly disagree" to "strongly agree".

The **14-Item Resilience Scale (RS-14)** was developed by Wagnild [147] as a shorter version of the original 25-item RS. This instrument measures five characteristics of resilience, namely: meaning and purposeful life, perseverance, equanimity, self-reliance, and existential aloneness [368]. It uses a 7-point Likert-type response format and is widely used in different fields.

Short Form Self-Regulation Questionnaire (SSRQ) [148] contains 31 items. It is the short version derived from the Self-Regulation Questionnaire (SRQ) [369] that was designed to measure self-regulation capacity across seven processes. Responses are rated on a 1–5 scale (strongly disagree to strongly agree) and can be summed up to generate a total score.

Self-Transcendence Scale (STS), "Self-transcendence" (ST) refers to the ability to broaden personal boundaries and focus on perspectives, activities, and goals beyond oneself, while still recognising the value of the self and the present context [149]. ST can result in personal transformation, enhancing well-being and improving quality of life [370]. The STS was developed by Reed in 1986 and contains 15 items that address specific behaviours or perspectives associated with expanding self-boundaries in various ways. It includes inward expansion through introspective activities, outward expansion through interactions with others, and temporal expansion by living in the present or adopting perspectives on the past and future that enrich the present [149].

The **Flourishing Scale (FS)** [141] is a concise 8-item measure that assesses the respondent's self-perceived success in key areas like relationships, self-esteem, purpose, and optimism. It yields a single score representing psychological well-being.

Brief Resilient Coping Scale (BRCS) [150] is a 4-item measure specifically designed to assess an individual's tendency to cope with stress in highly adaptive ways. Each item in this brief questionnaire targets a different aspect of adaptive coping strategies, encouraging respondents to reflect on how they manage stress in various situations

5.3.4.2 Weekly Well-being Scale

Every week, participants answered the weekly mini-survey, including the World Health Organisation-Five Well-Being Index (WHO-5) and an open question at the end for participants to elaborate on their week if they wanted to. The WHO-5 is a brief self-reported assessment of current mental well-being using five questions; these questions are answered with a six-point scale from "All of the time" to "At no time". Table 5.1 shows the questions of the weekly scale.

Table 5.1: WHO-5 Well-being Index

No.	Questions
WHO-1	I have felt cheerful in good spirits.
WHO-2	I have felt calm and relaxed.
WHO-3	I have felt active and vigorous.
WHO-4	I woke up feeling fresh and rested.
WHO-5	My daily life has been filled with things that interest me.
Open q.	Is there anything else you'd like me to now?

5.3.4.3 Focus Group

We conducted a focus group with the company's intervention coordinators to better understand the internal experts' individual experiences and evaluate the intervention. The three participants were in managerial positions and were in charge of logistics within the company. We asked them to answer the questions from two perspectives, as participants and organisers of the company's intervention. Table 5.2 shows the questions we used as an interview guide.

5.3.5 Data Analysis

This section explains how the qualitative and quantitative data were analysed.

Table 5.2: Focus group questions

No.	Questions
1	What was your personal experience of the course?
2	What is your impression of the overall group experience?
3	How does your experience in this intervention compare to other well-being practices that you do?
4	Within your company, what other well-being practices have you offered in the past, and how do you think they compare to this intervention?
5	What would you personally wish the next well-being intervention to look like?
6	What do you think the potential pool of participants will wish for?

5.3.5.1 Statistical Analysis of Instruments

Data analysis of the psychometric scales was conducted using RStudio. After cleaning the database, we obtained descriptive statistics (mean and standard deviation) and analytic statistics (normality tests and independent samples t-test). We considered the significance level of 0.05 (P = 0.05) for all statistical tests. To compare the entry and exit surveys, we initially chose the independent samples t-test since our groups had different numbers of participants due to dropouts. Additionally, we performed a paired t-test using data from participants who completed both the entry and exit surveys. This allowed us to account for within-subject differences and maximise the statistical power for this subset of participants despite the smaller sample size. We included this analysis to better understand changes among those who fully participated in the intervention. Further, since our control group was very small and became even smaller by the end of the intervention, we decided not to include it in any statistical tests, as the statistical power was already compromised. See our repository [371] for the database and code.

For the weekly scale, we only report the means per week. We calculated the scores by averaging the responses of all participants who completed the survey each week.

5.3.5.2 Qualitative Analysis of Focus Group

To analyse the data gathered from the focus group, we followed the guidelines of thematic analysis by Braun and Clarke [193]. We decided to perform it inductively, that is, codes and themes were derived directly from the data. The first and second authors went through the transcripts to become familiar with the data, as stated in the first step. Then the initial codes were generated, compared and discussed to reach agreement on their interpretation and to ensure consistency in the coding process. Later, the themes were identified, reviewed and defined to finally write up the results.

5.3.5.3 Reflexivity

The first author has a bachelor's degree in psychology and a master's degree in social work, and brings a deep understanding of human behaviour and social dynamics to the study. Her background and expertise in psychometrics equip her with the skills to explore the psychological aspects of well-being, such as stress management, coping strategies, and interpersonal relationships, which are crucial in the context of software engineering work environments.

Conversely, the second author, who holds a PhD in software engineering, offers expertise in the technical aspects of software development and extensive education as a yoga teacher. Their knowledge can shed light on the work-related factors that impact well-being, such as workload, project deadlines, and the use of technology in the workplace.

The mix of backgrounds and approaches allows for critical evaluation of interventions that address stress and well-being in the software engineering field

5.3.6 Ethical Considerations

This study adhered to the ethical research guidelines recommended by our university. Additionally, the study received approval from the country's ethics agency. All participants gave their informed consent.

Participants were comprehensively briefed on the study's objectives, methods, and potential risks. They were also informed of their right to withdraw from the study at any time without any consequences.

To ensure participants' privacy, all personal identifying information was kept strictly confidential. All collected data, including transcripts and audio recordings, was anonymised and securely stored.

5.4 Results

In this section, we report the results of the quantitative and qualitative data.

The intervention started with twenty-nine participants filling in the entry survey and finished with fourteen exit survey responses for the intervention group. For the control group, seven participants filled in the entry survey and five the exit survey.

The intervention group and the control group had similar demographics: We had a balance in terms of gender 50/50 men/women (no one identified as non-binary). All participants were at advanced stages of their careers with 10-15+ years of experience. They all held a university education (either MSc or PhD), and we had about 33% in leadership roles (program manager, project manager, product manager, engineering manager) and roughly 66% engineers. These percentages are quoted as "roughly" since some participants have overlapping functions and do not qualify as strictly one or the other. About 90% of participants were in technical roles in engineering and about 10% in human resources, communication and business management. Of the control group, 80% were in technical roles and about 20% in human resources, communication or business management.

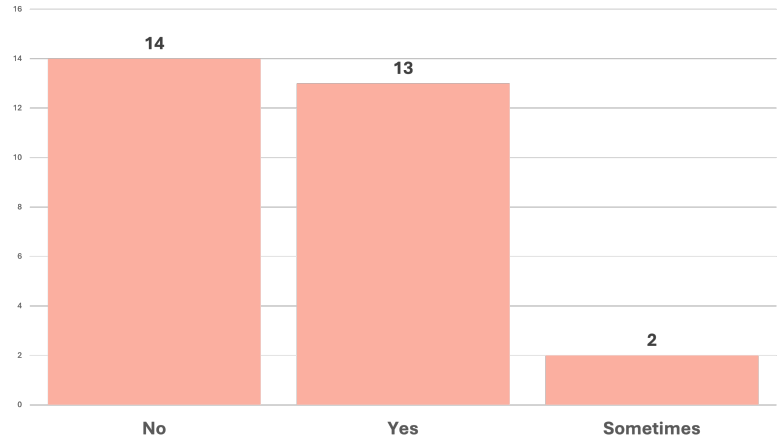


Figure 5.2: Participants' Having Well-being Practices Before the Intervention

5.4.1 Quantitative Analysis

Answers to the question about participants currently having a well-being practice are shown in Figure 5.2. The majority (14) answered with a "No", meanwhile 13 participants said they have a practice and 2 participants that only sometimes.

The results of the weekly scale are shown in Figure 5.3. It is visible that overall, participants had a higher level of general well-being at the end of the intervention compared to the initial one in week 1.

Regarding the psychometric instruments, we first performed the Shapiro test to assess the normality. Table 5.3 shows the results of the normality test for each psychometric instrument (W). The p-values are shown in brackets, all of which are greater than 0.05. Therefore, we do not reject the null hypothesis, indicating that the data can be assumed to follow a normal distribution.

Table 5.3: Shapiro-Wilk Normality Test Results

Test	Group	SSEIT	Resilience	SelfRegulation	SelfTransformation	SelfSuccess	Coping
Pre-test	**Int En**	0.982 (0.886)	0.966 (0.459)	0.969 (0.540)	0.951 (0.192)	0.972 (0.623)	0.942 (0.115)
	Cont En	0.780 (0.026)	0.937 (0.610)	0.926 (0.517)	0.967 (0.876)	0.900 (0.332)	0.915 (0.432)
Post-test	**Int Ex**	0.971 (0.895)	0.908 (0.147)	0.970 (0.875)	0.975 (0.933)	0.935 (0.356)	0.963 (0.764)
	Cont Ex	0.964 (0.838)	0.908 (0.453)	0.764 (0.040)	0.804 (0.087)	0.813 (0.104)	0.964 (0.833)

We then calculated the descriptive statistics for each scale. Figure 5.4 shows the mean scores for each psychometric instrument per group and the visual comparison of all the means. The differences between the entry and exit surveys and the control group are minimal. Based only on the means, the control group showed a better improvement in all scales in comparison to the intervention group. The difference in the means of the intervention group was slightly higher, and even one scale (STS) had a decrease after the intervention.

To explore the differences between pre- and post-intervention, we initially per-

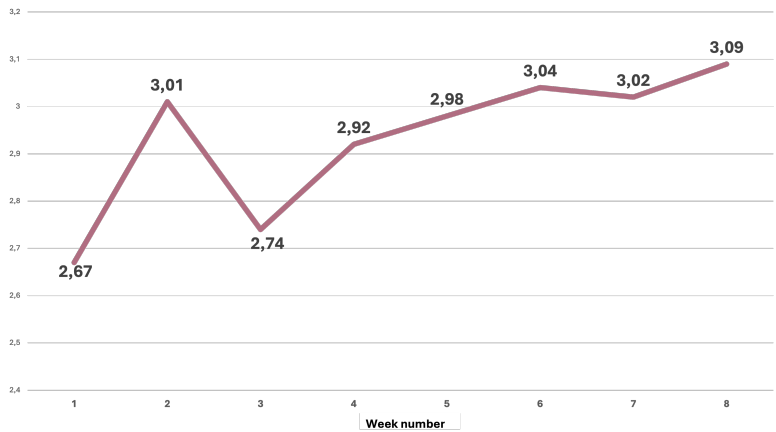


Figure 5.3: Participants' Weekly Well-being Score. We considered all participants means who answered each week.

formed an independent t-test; the results are presented in Table 5.4. There were no significant differences in any scale after the intervention finished. Then, we also performed a paired t-test with only the 14 participants who completed pre- and post-intervention surveys to gain additional insight into the data. Table 5.5 presents the results. Although the overall findings remain quantitatively non-significant, the paired t-test provided a clearer view of the data for participants who fully engaged in the intervention.

Table 5.4: Results of the Independent T-Tests for Psychometric Scales

Scale	**t value**	**df**	**p value**
Emotional Intelligence	-1.123	21.389	0.2739
Resilience	-1.2905	23.711	0.2093
Self-Regulation	-0.6949	21.601	0.4945
Self-Transcendence	-0.4783	22.945	0.6370
Self-Perceived Success	-0.9278	28.728	0.3612
Coping	0.7535	20.693	0.4597

5.4.2 Thematic Analysis

Three themes were generated from the focus group data analysis and are described below. Figures 5.5 and 5.6 are representations of the focus group's participant experience during the yoga intervention.

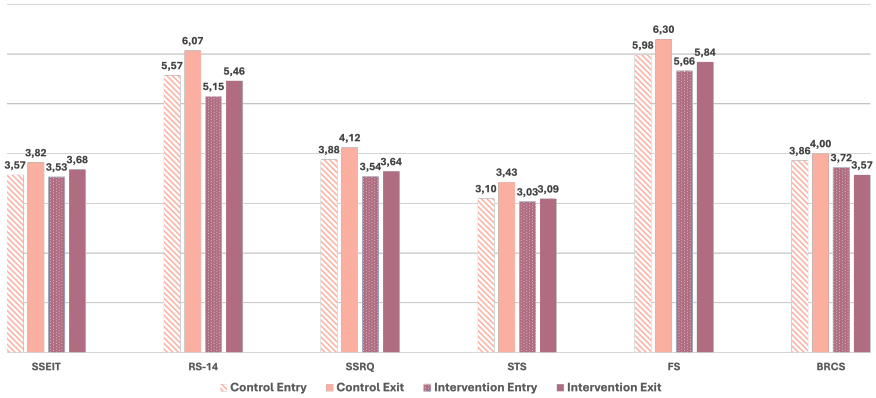


Figure 5.4: Means Comparison Between the Intervention (pre and post-programme) and Control Group

Table 5.5: Paired T-test Results for Each Scale

Scale	t value	df	p value
Emotional Intelligence	-0.75378	13	0.4644
Resilience	-0.18751	13	0.8542
Self-Regulation	-0.65387	13	0.5246
Self-Transcendence	-0.08407	13	0.9343
Self-Perceived Success	-0.30439	13	0.7657
Coping	0.25320	13	0.8041

5.4.2.1 Theme 1: Individual Benefits and Shared Reflections in Practice

This theme describes the impact of the intervention on personal and group levels. We identified three sub-themes that show how yoga influenced participants' well-being, fostered group dynamics, and evoked symbolic representations of the practice.

Sub-theme 1: Personal Benefits. This sub-theme focuses on the individual gains participants experienced from the yoga sessions, spanning physical, mental, and emotional well-being. Participants commented on how the intervention helped them manage stress and enhance emotional balance, emphasising how breathing techniques contributed to relaxation and focus.

This participant explained how yoga offered them more than physical or mental benefits. The quote below shows that yoga helped them relieve stress, promote overall well-being, and contribute to their professional life by enhancing their cognitive abilities. Specifically, yoga improved mental clarity, focus, and knowledge acquisition, which helped them perform better in their work.

> *"So I want to say that it's not only yoga and well-being. It's stress relief, but it's also cognitive input to my professional work life that helps me... it adds value to things*

other than only to my body and mind, but also to my cognition, my knowledge." —

Participants shared stories of overcoming initial hesitation towards yoga, with some noting their previous negative experiences with fast-paced classes. In contrast, this intervention's structured and mindful pace was described as relaxing and immediately impactful, encouraging participants to remain open to future yoga sessions. One participant, initially sceptical of yoga, reflected on how they overcame the barrier of waking up early to attend the sessions and found the practice deeply relaxing. These narratives explain how yoga fostered a sense of mindfulness beyond physical benefits, enabling participants to separate rational thought from emotional stress.

Sub-theme 2: Group Experience. Yoga also had a significant effect on the collective experience of participants. The organisers commented that a core group of about seven participants consistently attended the sessions and provided highly positive evaluations of the practice. While there was a drop in participation after the first two sessions, attendance stabilised, and those who continued to attend reported looking forward to the classes and appreciating their effects.

> *"I can remember a few times where someone actually either wrote on Slack or came to me saying something like: "I felt really bad in the morning and after yoga, I felt so much more ready for the day in a positive mindset"."* —

Participants gave the organisers generally good feedback, and this was shown in practice when they returned to class after missing a week and even joined online due to difficulties in commuting. The yoga sessions were beneficial on an individual level and created a shared space for relaxation within the company. One example is the "words of wisdom" (as commented by one participant) shared during the classes, which were described as having a lasting impact, with participants feeling empowered to pass on these lessons to others outside of the sessions.

Sub-theme 3: Visual and Symbolic Representations. A unique aspect of the participants' experience was how they described yoga through visual and symbolic representations. Participants used imagery to capture the mental and emotional states fostered by the sessions. For example, the colour blue (in an image done in the focus group) was repeatedly mentioned, symbolising peace and harmony, with one participant visualising blue bubbles during breathing exercises to represent a sense of calm.

> *"Then also my peace during the sessions became better. So that represents the blue dots, all the sessions we've had and that they were really like harmonised and peaceful."* —

Conversely, darker colours were used to depict confusion or unclear mental states early in practice, which gradually transitioned to lighter colours, symbolising clarity and calm as the sessions progressed. Other visual metaphors included two brains, one representing a wandering, distracted mind and the other symbolising the focused state achieved through yoga. Participants also highlighted the symbolism of yoga mats, which sparked discussions around them. The candles used in the sessions were described as a symbol of tranquillity, contrasting with chaotic external conditions, such as the inconvenience of practising near smelly shoes (week 1, due to the small room capacity). These visual and symbolic representations reflect participants' deep mental and emotional engagement with the practice.

5.4.2.2 Theme 2: Organisational Support and Logistical Challenges in Implementing the Programme

This theme explains the relationship between organisational support and the logistical challenges of implementing the yoga programme. Organisers acknowledged the company's commitment to promoting well-being, recognising its role in encouraging employee engagement. However, they also highlighted limitations within the organisation that could hinder participation.

Logistical factors, including room characteristics, scheduling preferences, and resource availability, significantly influenced participants' experiences. Additionally, the complexities of securing approval and coordinating sessions illustrated the challenges organisers faced, particularly when balancing their dual roles as both organisers and participants. Overall, this theme emphasises the need for ongoing organisational support and effective logistical planning to create an inclusive environment that encourages participation in well-being initiatives.

Sub-theme 1: Company's Role in Supporting Well-Being Initiatives. Participants commented on the company's role in supporting the yoga sessions as part of its broader well-being initiatives. The intervention was seen as an opportunity for the company to demonstrate its commitment to employee health, and several participants expressed high appreciation for the company's involvement in promoting well-being practices. Providing such interventions within the workplace was viewed favourably, with many recognising that workplace-based yoga sessions offered logistical advantages compared to external options like gym memberships or yoga studio subscriptions.

> *"Once a year we have wellness day, where we get presentations by different companies for like advertisements on well-being and what to do."* —

Despite this, there was also discussion about the limitations of the company's well-being initiatives. While the yoga sessions were well-received, participants noted that other employees' priorities or workloads might interfere with fully engaging in such programs.

> *"Some have meetings at 8:30 And then some have meetings at 9. There's always someone who has the next important meeting. It's so hard to fit everyone's discussion."* —

This points to the need for well-being interventions that not only exist but are integrated into a broader culture of health within the organisation, encouraging participation from a wider range of employees.

Sub-theme 2: Logistical and Environmental Factors. Logistics and the physical environment played a significant role in shaping participants' experiences with the yoga sessions. Participants described how environmental features like using (fake) candles helped create a calming atmosphere conducive to yoga. However, challenges related to space and resources were also mentioned, such as the availability of yoga mats and differing preferences for class locations.

> *"And then there was a huge discussion where it should be. Should it be in [room's name]? which is a big room that we have upstairs. Or should be downstairs in a smaller room?"* —

Furthermore, the hybrid option—allowing employees to participate either in person or online—was seen as a valuable addition, especially for those working from home, and it contributed to the overall accessibility of the program.

Sub-theme 3: Challenges and Efforts in Organising Well-Being Programs. Participants reflected on the challenges involved in the general planning, such as getting approval and securing funding for the sessions, with one describing the process as a "chaotic journey" that required considerable effort to bring everything into place.

> *"This was chaos for organising. And until everything fell into place, which is along the journey, it took quite a lot of effort."* —

There was initial resistance from the company, and it took time to convince key decision-makers of the value of the intervention. The level of commitment required from the organisers was also emphasised, with organisers playing dual roles, coordinating and participating in the sessions. This dual role added a layer of complexity, as the line between organiser and participant blurred. Participants noted that a significant amount of work went into ensuring the program ran smoothly, from logistical planning to recruitment and retention efforts. The organisers' high dedication was necessary to overcome these obstacles and implement the program effectively.

5.4.2.3 Theme 3: Perception and General Feedback

This theme focuses on employees' perceptions of, feedback on, and responses to the yoga programme in the company. It shows a landscape of interest, engagement, and organisational context. Participants shared varying perceptions of the company's efforts to promote yoga and well-being initiatives. They wanted more information about future courses and a greater understanding of yoga practices. There was a noticeable awareness of stress levels within the company, motivating employees to seek additional well-being strategies.

> *"Not everyone but a lot needed it. Because we are stressed and consciously, we don't admit this."* —

However, while some employees were enthusiastic and inquired about upcoming sessions, there was a recognition that participant commitment might not always match the organisers' dedication to the program. To enhance participation, several suggestions for strategies to involve more employees emerged. Additionally, participants highlighted the significance of research in evaluating these programs, noting that understanding the impact and outcomes of yoga sessions could reinforce their value within the organisation.

5.5 Discussion

In this section, we interpret the results in a wider context, argue for the relevance of the study despite non-significant statistical results, and the threats to validity.

Figure 5.5: Participant's Symbolic Journey During the Yoga Programme

Figure 5.6: Two Brains ["One representing a wandering mind, and the other, a focused mind. For me, this captures the essence of the journey"]. Image From Focus Group

5.5.1 Results in Context

The interest in studying the effects of a yoga intervention at the workplace was based on the positive benefits in different contexts [353–355]. We also considered Hafenbrack's [362] factors for on-the-spot intervention in the workplace. In the research at hand, the yoga intervention to reduce stress among software developers showed no statistically significant effect across the six psychometric scales between pre- and post-test assessments. There was a **slight increase in mean scores** for each scale; however, the changes were not large enough to reach **statistical significance**. Further, the control group was not big enough to perform statistical tests with enough power and reliability. Our quantitative findings suggest that the intervention may not have produced measurable improvements in participants' emotional intelligence, resilience, self-regulation, stress transformation, perceived success, or coping abilities, at least within the time frame and structure of the study.

We list several potential explanations for these results. First, the frequency of the intervention (one session per week) might not have been sufficient to create significant shifts in the psychometric outcomes. Other interventions with similar populations using mindfulness practices, such as meditation [91] and yoga [27], had a higher frequency (four times per week and daily practices, respectively). This suggests that our **intervention's "dosage" may not have been enough** to induce substantial changes. We did not track participants' engagement outside the weekly sessions; if participants only stuck to the weekly sessions, they may not have experienced the full benefits of yoga. In addition, participants frequently mentioned the "end of the

year" stress of having to finish a large number of tasks before the holidays. Hence, the **timing of the intervention** in the last 8 weeks before the winter break may not have been ideal (from a data collection point of view) since participants were likely to experience an increase of stress.

Another possible reason for the lack of significant findings could be the stress of the software development tasks. Software engineering is a high-strain job [87] with a combination of high demands [200], constant change [372] and technology-reliant [373]. Furthermore, it is plausible that external stressors continued to affect participants, potentially overshadowing the benefits of the yoga intervention.

Interestingly, despite the absence of significant changes in the psychometric scales, the focus group analysis revealed positive feedback from participants. The analysis showed they felt more relaxed, better able to manage stress, and more mindful after attending the sessions, aligning with existing literature that suggests yoga can improve subjective well-being [**?**, 60, 356, 357]. Looking at Figure 5.3, we notice a spike from week one to week two, which we attribute to the newness factor. Subsequently, we see a significant drop in rating in the third week, which is most likely due to a **major company milestone** where a product was going public for the first time. The instructor recalled several participants commenting on this event and how stressed they were about it. There is no way to control for such external stressors or confounding factors.

Overall, at the end, the well-being of the group finished higher than at the beginning of the intervention. This discrepancy between the quantitative and qualitative data may suggest that the intervention had subjective benefits that were not fully captured by the psychometric tools. Participants may have experienced shifts in their stress perception or management that were more subtle, context-dependent, and not easily measurable by standardised scales. Similar to the Daane [361] study and their quantitative results. This also stresses the importance of considering quantitative and qualitative outcomes when evaluating intervention programmes.

5.5.2 Importance of the Study

Despite these results, why is this study important? On a larger scientific scale, negative or null results become part of the bigger story about the intervention and what it targeted. By publishing negative results, we strengthen transparency and accountability in research. They help to interpret positive results that may have been obtained in related studies. They may adjust research designs and thereby increase the chances of success. Finally, the publication of null results will result in less bias in future meta-analysis studies, which could have incorrect conclusions if negative results are not included because they were never published. A less biased range of outcomes will ensure such meta-analyses are much more valuable.

In the particular case of our study, there are a number of confounding variables that were not possible to filter out and control in the sample size. We need these null results to redesign our experiment. We need access to negative and null results to guide us on the path to positive results.

The benefits of yoga may not be universally applicable or may require longer-term interventions, different formats, or complementary approaches to yield noticeable improvements in high-stress, cognitively demanding professions. Hence, it is important

to carefully tailor wellness interventions to the unique needs of their workforce. Rather than relying on one-size-fits-all approaches, organisations may need to explore other strategies or enhance yoga programs with additional resources like mental health support, ergonomic adjustments, or stress management training.

5.5.3 Lessons Learned

Despite the lack of measurable changes in psychometric scales, we identified several important lessons:

[L1] Stress-Management Interventions Must Address Software Engineering Workflows. The intervention was during a period of high pressure for the company, characterised by year-end deadlines and critical project milestones. However, due to the software engineering dynamics, for example, product release cycles, Agile sprints, and incident response demands, stress is always a challenge when finding the right time to start an intervention. This context might increase the difficulty of engaging participants when work-related stress peaks. Future interventions should account for these patterns and align better with project timelines. These include integrating short, stress-relief activities during sprint breaks or conducting longer sessions in less intense project phases.

[L2] Engagement Is Not Synonymous with Measurable Outcomes. Participants reported enjoying the weekly yoga sessions, but this positive reception did not translate into measurable improvements in any psychometric scale. It might suggest the need for future programmes to incorporate elements beyond enjoyment, such as tracking individual goals, providing reminders for daily practice, or connecting the intervention to broader organisational well-being strategies.

[L3] Weekly Interventions Alone Are Insufficient in High-Stress Contexts. A single weekly session, while appreciated by participants, was insufficient to counteract the acute and ongoing stressors in the software engineering workplace. This limitation stresses the importance of integrating more frequent or accessible stress-management practices into daily routines. For example, teams might benefit from micro-interventions, such as five-minute breathing exercises or mindfulness breaks incorporated into stand-ups or coding sessions.

[L4] Psychometric Scales Alone May Not Capture Software Engineering-Specific Stressors. The validated psychometric tools used in this study may not fully reflect the unique stress dynamics in software development, such as cognitive overload from debugging, context-switching, or tool-related frustrations. Although these scales measured general concepts related to well-being and resilience, their lack of sensitivity to domain-specific stressors and acute stress may have contributed to the lack of significant findings. In future interventions, we suggest additional metrics and qualitative methods tailored to the software engineering context.

[L5] Participant Dropout: a Need for Flexibility and Individualisation. We had a notable dropout rate, suggesting that the one-size-fits-all approach may not meet the diverse needs of software engineers. Participants likely struggled to balance attendance with their demanding schedules, especially during a high-pressure work period. To minimise dropouts in future interventions, offering more flexible options such as recorded sessions for asynchronous participation or shorter, on-demand activities could better accommodate varying workloads and time constraints.

In general, we learned that it is essential to tailor intervention programmes to the unique demands and context of software engineering. By aligning interventions with team dynamics, cognitive workloads, and the cyclical structure of the work, organisations can create more effective and sustainable approaches to supporting employee well-being.

5.5.4 Validity Threats

5.5.4.1 Internal Validity

Several factors were considered to address internal validity. First, the intervention was voluntary, meaning random assignment to groups was impossible. As a result, self-selection bias likely occurred, as participants had a pre-existing interest in or experience with yoga. Another challenge was controlling and measuring confounding variables, making it difficult to determine whether other factors influenced the intervention outcomes. To mitigate this, we attempted to use a control group to establish a baseline for comparison and conduct pre- and post-intervention assessments. However, the control group was ineffective, limiting its utility in the analysis.

5.5.4.2 External Validity

Our intervention was conducted in a realistic setting, making it applicable to similar work environments, aiming for generalisability. The study's conditions were designed to be replicated across different companies and everyday situations, mimicking real-world scenarios. To facilitate this, we provided a detailed methodology and a replication package [371] to allow for the reproduction of the study in diverse settings. While we acknowledge that the cultural context of the company may limit the generalizability to similar environments, the participants came from diverse backgrounds. This diversity within the participant pool may mitigate cultural constraints, suggesting that the findings could be relevant across various organisational settings, provided similar working conditions and organisational cultures exist.

5.5.4.3 Construct Validity

To ensure construct validity, we considered several actions. For example, we used psychometric standardised tools to ensure the measurement of our variables was accurate. Data collection was triangulated with qualitative data from the focus groups to complement the scales. We also implemented a longitudinal follow-up during the intervention using a weekly tune-in to monitor changes over time. Further, we had experts in psychometrics and yoga interventions to review the instruments and methodology. These combined efforts strengthened the credibility of our findings and helped ensure that the constructs were accurately captured throughout the study.

5.5.4.4 Conclusion Validity

Several challenges compromised the conclusion validity of the study. The small and ineffective control group, which was further reduced by participant dropouts, limited statistical power hindered the ability to draw reliable conclusions about the

intervention's true effects. Additionally, external stressors, such as the end-of-year workload and critical company milestones, may have confounded the results, as these factors could have overshadowed any potential benefits of the yoga intervention. Finally, while the intervention elicited subjective improvements reported in the focus group, participants willing to participate were only the organisers, adding an extra layer of bias as their vested interest in the programme's success may have influenced their feedback. This potential bias in reporting could influence the validity of the perceived benefits of the intervention.

5.6 Conclusion

In this study, we designed and implemented a mindfulness-based course, specifically yoga, to explore the benefits of workplace well-being interventions in software engineer participants. Results from the quantitative analysis showed that the impact of yoga practice in this study was not statistically significant. It is essential to clarify that a lack of statistical significance does not imply that the intervention had no positive effects. Instead, it indicates that the observed changes could not be confidently attributed to the intervention based on the quantitative data. This may be due to a small sample size, participant response variability, or other uncontrolled variables. While statistical significance is a crucial marker for determining reliable effects, it is possible that the yoga practice had subtle or individual-level benefits that were not detected in the quantitative analysis. Furthermore, the qualitative data from the focus group and the employees' feedback reported to the organisers were mainly positive. The yoga course is now an option for employees offered by the company, and they are still attending it. Employees might find other benefits that were not captured by the scales. Hence, they are still attending the course.

Future work will include making sure to control for confounding variables and to have longitudinal follow-up after intervention data collection.

5.7 Acknowledgement

We thank the company and the managers who allowed us to run the intervention with them.

Chapter 6

Paper E:

A Multimodal Approach Combining Biometrics and Self-Report Instruments for Monitoring Stress in Programming: Methodological Insights

C. Martinez, D. Grassi, N. Novielli, B. Penzenstadler

Under submission to special issue on Empirical Software Engineering Journal (EMSE), 2025

Abstract

The study of well-being, stress and other human factors has traditionally relied on self-report instruments to assess key variables. However, concerns about potential biases in these instruments, even when thoroughly validated and standardised, have driven growing interest in alternatives in combining these measures with more objective methods such as physiological measures.

We aimed to (i) compare psychometric stress measures and biometric indicators and (ii) identify stress-related patterns in biometric data during software engineering tasks.

We conducted an experiment where participants completed a pre-survey, then programmed two tasks wearing biometric sensors, answered brief post-surveys for each, and finally went through a short exit interview.

Our results showed diverse outcomes; we found no stress in the psychometric instruments. Participants in the interviews reported a mix of feeling no stress and experiencing time pressure. Finally, the biometrics showed a significant difference only in EDA phasic peaks.

We conclude that our chosen way of inducing stress by imposing a stricter time limit was insufficient. We offer methodological insights for future studies working with stress, biometrics, and psychometric instruments.

6.1 Introduction

Software engineering (SE) is a cognitively demanding profession that requires intense focus, problem-solving, and creativity. However, these tasks often come with high-stress levels due to the work characteristics, which often involve long working hours, high cognitive load, frequent interruptions, task interdependence and tight deadlines [88]. Prolonged exposure to such stressors can lead to burnout, a state of emotional, mental, and physical exhaustion that negatively impacts individual well-being and organisational productivity [374].

Understanding the stressors specific to SE tasks and accurately measuring their impact is essential for developing effective interventions to mitigate these risks. Research on emotions, affect, and stress in software engineering has mainly used self-reported instruments, such as surveys, interviews and psychometric instruments. Graziotin et al., 2014 [5], were among the first researchers in the area proposing to study human factors using psychological measurements. Studies on happiness [375], attention awareness [91], positive and negative experience, psychological well-being [27, 107], positive thinking, and self-efficacy [19] have been conducted using psychometric instruments to assess these constructs.

While these methods offer insights into subjective experiences, they are prone to biases, including recall bias, social desirability bias (SDR), and acquiescent responding (ACQ) [376]. SDR refers to the tendency to respond in a way consistent with what is perceived as desirable by salient others [377]. Meanwhile, ACQ relates to the tendency to favour the positive end of the rating scale, irrespective of the item's content [378].

Additionally, self-reported measures may not fully capture stress's physiological and cognitive aspects, essential for understanding its impact on performance and well-being. To address these limitations, recent studies have investigated the use of biometrics to recognise developers' emotions during programming tasks [109, 110, 235, 379]. What these studies have in common is the operationalisation of emotions along the dimension of valence, i.e. the (un)pleasantness of the emotional stimulus, and arousal, i.e. the level of emotional activation [380], showing promising results in their recognition through machine-learning supervised classifiers.

Despite advances in this domain, the literature reveals a significant gap, as, to the best of our knowledge, no research has specifically addressed stress. As a result, there is a growing need for more objective and reliable methods to assess stress in software engineering contexts. At the same time, recent work by Westerink and colleagues [381] provided empirical evidence that biometrics collected with non-invasive sensors can be used as a stress indicator. Inspired by these findings, we decided to perform an empirical study to fill this gap, towards enhancing the accuracy and reliability of stress measurement in software engineering. This decision was in line with our long-term goal to support early detection of stress, thus enabling interventions to prevent its long-term negative effects on well-being and productivity.

We designed and implemented an empirical study with the primary goal of investigating to what extent we can use biometrics as a proxy for stress experienced by software developers during programming tasks, to reduce the reliance on self-reported data and obtain a more comprehensive understanding of stress related to SE tasks. To this aim, we compare biometric measurements with traditional psychometric instruments as collected during programming tasks performed by ten developers in a

controlled lab environment. Although we invested considerable time and resources in the design of the empirical protocol, we obtained disappointing outcomes due to the inability to induce stress in the participants of our empirical study. This prompted us to redirect our efforts toward a comprehensive assessment of the robustness of the protocol we adopted, thus deriving methodological guidelines to inform future studies on this topic.

A key finding was that the intended stress manipulation through time pressure failed to produce measurable stress responses at the group level. This led us to conclude that time pressure alone may be insufficient to induce stress in experienced programmers. Future studies should consider multi-stressor approaches or tasks with higher personal stakes for participants. Furthermore, our findings reveal that the individual-level triangulation of data sources provided more nuanced insights than the group level. This can be observed by the combined analysis of self-reported stress measures, electrodermal activity (EDA) peaks, and qualitative interview data on a participant-by-participant basis. Finally, we discuss methodological challenges associated with distinguishing between acute and chronic stress, which might be a confounder in a lab setting focusing on stress detection during coding tasks. Specifically, we noted that while our multi-modal measurement approach showed sensitivity to stress variations, the ethical constraints of inducing stress in research settings may fundamentally limit the ability to create strong enough stressors that eventually yield actionable data without crossing ethical boundaries.

The remainder of the paper is organised as follows. In Section 6.2 , we present the background and discuss the related work on stress and biometrics in software engineering. Then, in Section 7.3 we describe the methodology, including the experimental protocol for data collection and the method of analysis of psychometrics, biometrics, and interviews with participants. Results are presented in Section 7.4 and discussed in Section 6.5, where we also present the threats to validity and the strategies adopted to mitigate them. Finally, we conclude the paper and discuss future work directions in Section 6.6.

6.2 Background and Related Work

Physiological measures, such as electroencephalography (EEG), electrodermal activity (EDA), and heart-related metrics, have become valuable tools for studying cognitive load and stress across various domains. These measures offer objective insights into mental states, offering advantages over traditional self-reported methods. In fact, biometrics hold the potential to address the limitations of self-report methods by providing objective, continuous measurement of the biometric changes that are induced by mental states [382]. Among other affective states, in this study we specifically focus on the study of stress, that is, the physiological or psychological response to internal or external triggers, involving people's bodily reactions, feelings and behaviour (see Table 6.1). In the bi-dimensional categorisation of emotions along the concepts of valence and arousal, stress is positioned in the scope of negative emotions [383] and associated with high arousal [384]. This positioning reflects the nature of stress as an unpleasant emotional state that involves high physiological and psychological activation. Stress appears near other similar emotional states such as anxiety, tension,

distress, and nervousness in this model. In the following, we report foundational related work on the use of biometrics for the study of cognitive and emotional states (Section 6.2.1). We complement this background knowledge with an overview of recent related studies in the field of software engineering (Section 6.2.2).

Table 6.1 presents definitions for the most important concepts in this study.

Table 6.1: Operationalisation of main concepts based on the American Psychological Association definitions [177]

Concept	**Definition**
Stress	"The physiological or psychological response to internal or external stressors. Stress involves changes affecting nearly every system of the body, influencing how people feel and behave."
Mental Workload	"The relative demand imposed by a particular task, in terms of mental resources required."

6.2.1 Physiological Measures of Stress and Mental Load

Stress. The link between affective states and physiological feedback, collected with biometric sensors, has been investigated for a long time by researchers in the affective computing community [385–388]. In recent years, the study of emotions and their recognition has gained attentions also in software engineering research, due to their influence on developers' wellbeing, stress levels, and cognitive performance [5, 389].

Various biometric signals have been employed to detect affective states. In particular, EEG has been widely used to analyse changes in brain activity correlated with emotional valence (pleasant vs. unpleasant emotional stimulus) and arousal (i.e., high vs. low level of emotional activation) [390]. For instance, high-frequency bands such as gamma have shown strong correlations with valence, particularly in the frontal and parietal lobes [387]. EEG also enables computation of Frontal Alpha Asymmetry, a known biomarker linked to emotional valence and stress [391] Moreover, EDA is widely adopted due to its association with the arousal dimension [392]. EDA has thus been effectively used to identify emotions [111, 235]. Its sensitivity to emotional intensity makes it a valuable, non-invasive proxy for monitoring real-time emotional fluctuations during cognitive tasks. Furthermore, HR and HRV metrics also provide insights into emotional arousal and cognitive load. Specifically, HRV indicators such as RMSSD and LF/HF ratio have been shown to reflect sympathetic and parasympathetic nervous system activity, which are modulated during emotional and stress responses [393, 394]

Similarly to what was done for the recognition of emotions, the study of biometrics has been applied to the recognition of stress episodes. In particular, EEG has been used to identify specific brainwave patterns, such as alpha and beta frequencies, which are closely linked to stress. In their study, Saeed et al. [395] found that alpha asymmetry could be a potential reliable biomarker for stress classification.

They complemented the EEG data with the Perceived Stress Scale (PSS-10) and an interview to obtain a thorough understanding of stress. A similar setup was used in our study to get a more complete view of how stress manifests physiologically and emotionally. A similar study by Chae et al. [396] looked at the relationship between stress levels and rework using EEG, EDA, and a survey, finding that all three measures consistently indicated that rework caused stress in workers. They emphasised that excessive occupational stress can negatively affect employee work performance and work-life balance. Additionally, they stressed the cognitive and emotional toll of repetitive tasks. Our study builds on this by comparing stress measurements from EEG, EDA, and psychometric instruments to enhance the understanding of workplace stress, particularly in high-pressure environments such as software engineering.

As for EDA, its link with stress episodes was demonstrated by Westerink et al. [381]. They explored the use of physiological sensors for detecting stress episodes. Their findings revealed a significant relationship between cortisol fluctuations (the primary stress hormone) and electrodermal activity (EDA) measurements. Notably, peaks in skin conductance preceded cortisol elevations, which suggests that EDA monitoring could serve as an early warning system for stress onset. In related work, Kocielnik et al. [382] developed an approach that integrated EDA measurements with calendar data to examine potential connections between daily activities and stress responses. The paper presents a framework for long-term, unobtrusive stress monitoring in workplace settings using a wearable sensor wristband (DTI-2) that measures skin conductivity. The authors proposed an approach to process EDA raw signals to identify stress levels and visualise this data in relation to users' calendar activities. Through field studies with university staff, they demonstrated that this approach helps users discover meaningful stress patterns they weren't previously aware of.

Mental Workload. Linked to stress and based on the premise that workload affects performance, Mohanavelu et al.'s [397] study focused on measuring and understanding the cognitive workload and attention during different levels of task difficulty: normal, moderate, high, and very high workloads. They used EEG to track how the brain responds under varying workloads and a NASA-Task Load Index (NASA-TLX) questionnaire to validate their findings. Results from EEG showed that the prefrontal, frontotemporal, and parietal brain regions were highly engaged under high and very high workloads; NASA-TLX results aligned well with EEG data. Considering the previous results and since software developers' work also demands a high mental workload, we used NASA TLX in this study to capture subjective workload data.

Similar to the previous study and considering sleep deprivation, which is also quite present in the software engineering field, Martínez Vásquez et al. [398] collected data from ten participants performing cognitive tasks every two hours for 24 hours to explore the relationship between brain activity (EEG) and autonomic sympathetic activity (EDA) under sleep deprivation, aiming to assess their role in determining readiness for cognitive tasks. Based on their findings, the authors proposed that the mutual information between EDA and EEG signals reported in their study indicates that examining EDA could offer a compelling alternative for studying brain activity. In our study, we collected both data to explore their relationship with cognitive and emotional processes.

6.2.2 Using Biometrics for Studying Cognitive and Affective States in Software Development

Researchers in software engineering have explored connections between developers' cognitive states—as measured through physiological indicators—and various software development dimensions, including comprehension of code [399, 400], developer productivity and interruptibility [401, 402], and the emotions experienced by developers during programming tasks [109–111, 235].

EEG. Several studies have been done to research brain activity during programming tasks, focusing on the cognitive load and mental effort involved in software development. For instance, Calcagno et al. [403] investigated brain activity during programming tasks using EEG with ten experienced software developers. Their results showed significant changes in brain activity when transitioning from a baseline condition (typing with eyes closed) to a programming task. Specifically, they observed a decrease in Alpha power and an increase in Delta, Theta, and Beta power, particularly in the frontal and parieto-occipital regions. The increase in Beta activity was most prominent at the beginning of the task, likely reflecting the heightened alertness and attention required for understanding instructions and planning code implementation. In contrast, Theta and Delta power increased during later phases, suggesting greater mental workload and working memory engagement. Their results suggest that EEG measures can provide insights into cognitive load and attentional dynamics during software development tasks.

Medeiros et al. [404] performed a controlled experiment on task comprehension with 26 programmers using three code snippets in Java with different complexity levels. The study found that features related to Theta, Alpha, and Beta brain waves were the most effective at identifying levels of mental effort required by different code lines. The EEG data indicated signs of mental effort saturation as code complexity increased. In contrast, traditional software complexity metrics did not accurately reflect the cognitive effort required for code comprehension.

Radevski et al. [401] introduced a framework that continuously monitors developers' productivity by tracking electrical activity in the brain, to assess and improve their productivity. Their proposed approach relies on off-the-shelf EEG devices to support their long-term goal of detecting negative cognitive and emotional states such as stress, fatigue, and frustration, which might emerge during programming tasks. While not being assessed for the specific task of emotion detection, the framework's usability was evaluated through a pilot user study with six participants who wore the device for an entire workday, finding it was feasible but had some comfort issues. Their study also addresses ethical considerations and user acceptance challenges that must be considered when conducting empirical studies involving the use of biometric devices.

Combining data from multiple sensors. Beyond EEG, various approaches have been proposed based on different sensor combinations for the recognition of emotional and cognitive states. Müller and Fritz [109] employed a combination of biometric indicators to assess both progress and interruptibility during small development tasks. They demonstrated that emotional states of developers during programming tasks could be classified with 71% accuracy by analysing a rich set of physiological signals, including brainwave frequencies, pupil dimensions, and heart rate. They also report achieving a comparable accuracy when predicting developers'

self-perceived progress during development tasks, though this required a distinct set of biometric indicators encompassing EDA signals, skin temperature, brainwave patterns, and pupil size variations.

In a partial replication of the original study by Müller and Fritz [109], Girardi et al. conducted an empirical investigation to identify the minimal configuration of non-invasive biometric sensors for recognizing emotions during programming tasks [235]. They developed two supervised classification models for valence and arousal dimensions using emotions self-reported by 23 participants during a Java programming assignment as a ground truth. Through experimentation with various biometric combinations, they found that developers' emotional valence and arousal could be reliably detected using a combination of electrodermal activity (EDA) and heart-related measurements, collected via the Empatica E4 wristband, suitable for emotion detection during software development activities. Using only the wristband, they achieved accuracy levels for valence (.71) and arousal (.65) comparable to those obtained with the complete sensor array (wristband + EEG helmet). Consequently, in their subsequent study, they utilized only the Empatica wristband for measuring both electrodermal activity and heart-related biometric signals [111]. Their study not only confirmed previous findings by M"uller and Fritz [109] regarding non-invasive sensors' reliability for valence classification, but also extended this work by developing an arousal dimension classifier.

Vrzakova et al. [110] combined eye tracking measurements and electrodermal activity to classify emotional valence and arousal of software developers during code review activities. They conducted an *in-situ* study with 37 professional developers engaged in code review tasks. They used features extracted from individual signal types as well as combined feature sets incorporating all available signals to train supervised machine learning models. For evaluation, they established a ground truth using binarised self-reported emotional scores for valence (positive vs.) negative and arousal (low vs. high). Their findings revealed that eye gaze measurements provided the strongest predictive capability for both emotional dimensions, achieving accuracy rates of 85.8% for valence and 76.6% for arousal. However, when incorporating features from all physiological signals, including EDA, in their supervised models, they observed a boost of classification performance for both valence and arousal dimensions, with accuracy rates reaching 90.7% and 83.9%, respectively.

6.3 Methodology

We designed an experiment with two main objectives: (1) to identify the stress levels induced by programming tasks and (2) to evaluate the accuracy of self-reported instruments in measuring stress in comparison with biometric measurements. To achieve this, we investigated the following questions:

[a] How reliable are psychometric stress measures compared to real-time biometric indicators (EEG and EDA) during software engineering tasks?

[b] What stress-related patterns can be identified in real-time biometric data (EEG and EDA) during software engineering tasks?

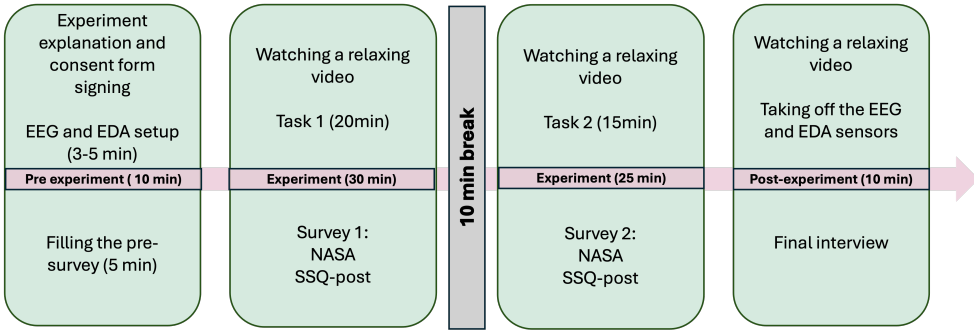


Figure 6.1: Experiment's Timeline

This section outlines the experimental protocol, the instruments used for data collection, and the approach to data analysis.

6.3.1 Participants and Recruitment Strategy

The study included ten participants: nine PhD students in computer science, artificial intelligence, and bioinformatics and one master's student in data science. They were in various academic stages, from the first to third year, and reported confidence levels in programming ranging from "somewhat confident" to "very confident." Python was the preferred programming language for most, except one who preferred Java.

6.3.2 Experiment Setup

To conduct the experiment, we recruited subjects from the Ph.D. and master's students in Computer Science who could code in Python or Java. We collected the preliminary availability of volunteers and scheduled the experimental sessions based on their agendas over a time span of two weeks.

Pre-experimental Briefing. Participants began by listening to the explanation of the experiment, reading the informed consent form, and having the opportunity to ask questions. After signing the informed consent form, participants wore biometric sensors, and the researchers made sure the signals were being captured correctly and started the recording. Subsequently, participants completed the first survey (PSS-10, and SSSQ-pre) and then watched a two-minute relaxation video to induce relaxation and establish a neutral emotional state [405]. The signal collected in this neutral emotional state is used as a baseline for each participant, which is required to preprocess the raw biometric signal, as explained in Section 6.3.4.2.

Programming Tasks and Data Collection. Participants started carrying out the first task, having 20 minutes to complete it. The tasks were a grid-based path optimisation problem requiring dynamic programming to compute the best possible resource accumulation under movement constraints (right/down or bidirectional) while handling cell-specific penalties/rewards (see tasks in the replication package [406]). Upon completing the task or reaching the time limit, they completed a second survey

(NASA TLX and SSSQ-post) reflecting on Task 1. The participants had a 10-minute break before moving on to the next task. After the break, participants watched a two-minute relaxation video, and right after, they started with Task 2, having 15 minutes to complete it. Both tasks were similar in complexity; however, Task 2 featured a shorter time limit to induce time pressure and increase stress levels. As with the first task, participants completed a survey evaluating Task 2. The experiment concluded with a final two-minute relaxation video to help the participants decompress.

Exit Interview. After participants took off the sensors, we ran a short interview to elicit their overall experience during the experiment (see interview in data collection methods). Finally, participants received a voucher for a restaurant to thank them for participating.

6.3.3 Data Collection Methods

In our study, we use a combination of biometric sensors and surveys to measure stress levels and mental workload.

6.3.3.1 Biometric Sensors

To collect data, we utilised two biometric devices: a wearable wristband for EDA and HRV acquisition and an EEG helmet. The Empatica EmbracePlus[1], as shown in Fig 6.2 (a), is a medical-grade wearable wristband, which we used for continuous, unobtrusive measurement of physiological signals. It includes a ventral EDA sensor that samples at 4 Hz and a PPG (photoplethysmography) sensor sampling at 64 Hz, from which we derived HRV metrics. The EEG data were recorded using a Neurosity Crown device[2], which measures electrical brain activity through its embedded sensors, as illustrated in Fig. 6.2 (b). The device consists of eight channels (CP3, C3, F5, PO3, PO4, F6, C4, CP4, which acquired the brain signals at a sampling rate of 256 Hz.

Figure 6.2: Wearable devices used in the study. (a) Embrace Plus by Empatica, (b) Neurosity Crown

[1]https://www.empatica.com/en-eu/embraceplus/
[2]https://neurosity.co/

6.3.3.2 Self-report Instruments

We used a combination of consolidated self-reported instruments, which are explained in detail below and are widely adopted in the literature.

Perceived Stress Scale (PSS-10) [152] is a widely used instrument to assess the degree to which individuals perceive situations in their lives as stressful. It evaluates feelings and thoughts over the past month, providing an understanding of how circumstances influence perceived stress levels.

Short Stress State Questionnaire (SSSQ) [153] is a validated instrument to assess stress states, measuring the engagement of tasks, stress, and worry. It has demonstrated sensitivity to task stressors, with different task conditions producing distinct stress patterns consistent with prior predictions. The tool includes pre- and post-task versions, making it valuable for researchers studying conscious appraisals of task-related stress.

NASA Task Load Index (NASA TLX) [154] was developed by the Human Performance Group at NASA's Ames Research Center, is a widely used tool for assessing subjective mental workload (MWL) during task performance. It evaluates MWL across six dimensions to produce an overall workload score: mental demand (cognitive effort for thinking, decision-making, or calculations), physical demand (intensity of physical activity required), temporal demand (time pressure involved), effort (exertion needed to maintain performance), performance (effectiveness in task completion), and frustration level (feelings of insecurity, discouragement, or contentment).

6.3.3.3 Post-Task Interview

The interview questions aimed to explore participants' subjective experiences during the study, focusing on their stress levels and task-related perceptions. Participants were asked to describe their overall experience, including any factors contributing to their stress, and to reflect on specific moments of increased or decreased stress during the tasks. The questions also addressed the impact of wearing EEG and EDA devices on their concentration and performance. In addition, participants were encouraged to share any strategies they used to manage stress or maintain focus and were invited to provide further comments about their experience. See the questions in Table 6.2

6.3.4 Data Analysis

The analysis of each dataset is explained in the following subsections. The goal of our analysis is twofold. First, comparing the psychometric, i.e. the self-reported stress, between Task 1 and Task 2 enables us to verify that we successfully induced stress in the participants during the second task by giving them less time for performing the coding task. Second, by comparing the biometrics collected during Task 1 and Task 2 we aim at verifying if there is any significant pattern in the physiological responses that can be used as a proxy for the self-reported level of stress at the end of each coding task. We further complement this analysis by comparing the biometrics during the pre-task and Task 1.

Table 6.2: Interview Questions

No.	Question
1	Can you describe your overall experience during the study? Did anything about the task or process contribute to your stress levels?
2	How did you feel during the tasks? Were there specific moments when you noticed increased or decreased stress levels?
3	How did you find the experience of wearing the EEG and EDA devices while completing the task? Did they interfere with your ability to concentrate or perform?
4	Did you use any strategies to manage your stress or stay focused during the task? If yes, what were they?
5	Do you want to add anything else about your experience in the study?

6.3.4.1 Psychometrics

The psychometric data were analysed using RStudio. First, we cleaned the data, inspected for missing values, standardised variable names, and reversed scale items when needed (based on the psychometric instruments guidelines). Then, we calculated descriptive statistics for each psychometric scale, including means and standard deviations. We used the unweighted average of all the scales to report descriptive statistics and to compare results across time points (see table 6.5). Later, we tested the normality of the distribution and then applied a paired t-test to assess significant differences.

6.3.4.2 Biometrics

To align the physiological signals with the different phases of the experiment (e.g., baseline, pre-task, first task, and second task), participants manually marked the start and end of each phase using the EmbracePlus wristband's event-tagging feature. These timestamps were then used to segment and synchronise the recorded physiological data (such as heart rate, skin temperature, and movement) with the corresponding experimental phases for the analysis.

To compare stress levels between the two tasks, we extracted the raw data from the two physiological sensors for a 1-minute window, starting 30 seconds before the end of each task. We chose 30 seconds for the data extraction to account for the possibility that there could be a time gap between the switching of the tasks. This approach is in line with validated practices [396] The raw data extracted from the two sensors were processed differently. The approaches to processing raw data from two different physiological sensors are presented below.

EDA. To account for individual differences in EDA signals, we standardised

the signals using z-score normalisation relative to the baseline signal collected while watching the relaxing video, following established methods adopted in related work [235]. The data was then preprocessed using the NeuroKit2 package [3]. We applied a 1Hz low-pass Butterworth filter to remove high-frequency noise, as done by Taylor et al. [407]. Next, we decomposed the filtered EDA signals into tonic (Skin Conductance Level, SCL) and phasic (Skin Conductance Response, SCR) components using the cvxEDA algorithm [408]. This preprocessing step is required to separate the slow-varying tonic components from rapid phasic responses, both of which are relevant for detecting stress-related responses [409]. From these two components, we extracted statistical features such as minimum, maximum, mean, and standard deviation (see Table 6.4), in line with previous work [410]

Since EDA peaks can be interpreted as a response to stress episodes [381], in our analysis we include consideration of such peaks. In particular, we identified EDA peaks using NeuroKit2's, following the peak detection approach proposed by Kim et al. [386]. To account for variations in task duration, we computed the number of peaks per minute by dividing the total peak count by the duration (in minutes) of each experimental phase. Similarly, we consider the duration of the pre-task step for which we also extracted the biometrics. This normalisation allows for more reliable comparisons across conditions of different durations, as in our case.

HRV. Heart rate variability was calculated using the hrvanalysis library[4], based on interbeat intervals, which represent the time intervals between successive heartbeats. The signal was preprocessed by removing outliers (interbeat outside the 300–2000 ms range), as recommended by [411]. Missing values were linearly interpolated, and ectopic beats were corrected using the Malik method [412].

For each task, we computed the RMSSD (Root Mean Square of Successive Differences), which reflects short-term heart rate variability and typically decreases under stress [413]. We also calculated the SDNN (Standard Deviation of Normal-to-Normal intervals), which has been shown to increase during stress episodes [394]. Finally, we computed the LF/HF ratio—the ratio of low-frequency (0.04–0.15 Hz) to high-frequency (0.15–0.4 Hz) components of the HRV power spectrum—which significantly increases during stress [394].

EEG. Initial inspection of the raw EEG signals revealed several instances of missing data. Specifically, in two cases (P5 and P10), the first task was missing, and in three cases (P8, P9, and P10), the second task was missing due to technical issues during data acquisition. Additionally, data for the pre-task phase were missing for two participants (P6 and P10). As a result, we had 6 data points available for analysing the comparison between the first and second tasks, and 7 data points for analysing the comparison between the pre-task and first task. Therefore, we decided not to perform any statistical analysis.

Analysis. We performed statistical analyses on EDA-derived and HRV-derived features to compare performance between the first and second tasks. Each feature was first tested for normality using the Shapiro–Wilk test. When the normality assumption held, we used paired t-tests; when it was violated, we substituted the non-parametric Wilcoxon signed-rank test.

[3] https://neuropsychology.github.io/NeuroKit/functions/signal.html
[4] https://github.com/Aura-healthcare/hrv-analysis

Task	EDA + HRV	EEG
Filling Presurvey	10	8
First Task	10	8
Second Task	10	7

Table 6.3: Number of datapoint per task (EDA, HRV and EEG)

Modality	Type	Feature	Stress	Mental Workload
EDA	Tonic	Mean, std, min, max	↑ [414]	-
	Phasic	Mean, std, min, max	↑ [413]	-
	EDA Phasic Peaks	Count	↑ [381]	-
	EDA Raw Peaks	Count	↑ [381]	-
	EDA Raw Phasic per Minute	Count	↑ [381]	-
	EDA Raw Peaks per Minute	Count	↑ [381]	-
HRV	Time domain	RMSSD	↓ [413]	↓ [415]
		SDNN	↑ [394]	-
	Frequency domain	LF/HF Ratio	↑ [394]	↑ [415]

Table 6.4: Physiological Features: EDA and HRV with Expected Behavior under Stress and Mental Workload

6.3.4.3 Interviews

The interviews were analysed following the six steps of reflexive thematic analysis by Braun and Clarke [75]. The interviews were first transcribed verbatim, the transcripts were read multiple times for familiarisation, and semantic, inductive codes were generated across the dataset. Codes were then grouped to identify potential themes, which were reviewed and refined to ensure they accurately captured patterns in the data. Themes were clearly defined and named to reflect their core meaning, and selected quotes were used to illustrate each theme in the final write-up. The process was conducted manually, with careful attention to researcher reflexivity. Since we strictly followed Braun and Clarke guidelines and aligned with Big Q qualitative values, we did not assess for inter-coder reliability [75, p. 240] as coding was treated as a flexible, interpretative process.

6.4 Results

This section presents the biometrics, psychometrics and interviews (thematic analysis) data results. In particular, we report empirical evidence from the analysis of the

psychometric and biometric indicators.

6.4.1 Psychometrics

Table 6.5 summarises the descriptive statistics (mean, standard deviation, minimum, and maximum) of the psychometric instruments (PSS-10, SSSQ and NASA-TLX) for each measurement: Pre-task, Task 1, and Task 2.

The average baseline perceived stress level (PSS-10) was in the moderate range (M=1.87, SD=0.56). Aligned with this, the pre-task SSSQ score (M = 2.35, SD = 0.38) also indicated a moderate subjective stress state before tasks began. Regarding their emotions, participants had a mean of 5.50 (SD = 1.10), reflecting relatively positive affect and moderate arousal levels before the tasks.

After Task 1, SSSQ dropped slightly to 2.25 (SD = 0.50); this change was not significant enough to impact the group stress levels. Finally, the NASA-TLX results were also moderated (M = 9.83, SD = 3.60) with considerable variability across participants. Regarding Task 2, SSSQ-post's score remained relatively stable (M = 2.24, SD = 0.33). NASA-TLX increased slightly (M = 10.82, SD = 2.81); however, there was no significant change.

Table 6.5: Descriptive statistics per group and task

Group	Instrument	Mean	SD	Min	Max
Pre-tasks	PSS-10	1.87	0.56	1.10	3.00
	SSSQ-pre	2.35	0.38	1.88	3.25
Task 1	SSSQ-post	2.25	0.50	1.71	3.13
	NASA	9.83	3.60	3.17	15.00
Task 2	SSSQ-post	2.24	0.33	1.79	2.71
	NASA	10.82	2.81	7.17	15.67

We compared participants' stress levels before and after the tasks using a paired t-test. Table 6.6 presents the results of the two comparisons, showing that none of the comparisons were significant. For Pre-task vs Task 1, the t-value was 0.592, with a p-value of 0.569, indicating no significant change in stress from the pre-task phase to Task 1. Concerning Task 1 vs Task 2, results showed a t-value of 0.137 and a p-value of 0.893, which means no statistically significant difference in stress scores between Task 1 and Task 2. Participants reported comparable levels of stress during both tasks, which suggests that the reduced time for Task 2 was not enough to induce a stress condition during the second coding task.

Furthermore, we also tested for significant differences in NASA-TLX results, presented in Table 6.7. Results showed a p-value of 0.426, implying there are no significant differences in mental workload in Task 1 and Task 2.

Table 6.6: Comparison of stress scores between tasks

Comparison	t-value	p-value	Mean Difference	Interpretation
Stress: Task 1 vs Task 2	0.137	0.893	0.017	No significant difference
Stress: Pre-task vs Task 1	0.592	0.569	0.092	No significant difference

Table 6.7: Comparison of NASA-TLX scores between tasks

Comparison	t-value	p-value	Mean Difference	Interpretation
Task_1 vs Task_2	-0.833	0.4265	-0.983	No significant difference

6.4.1.1 Closer Look per Participant

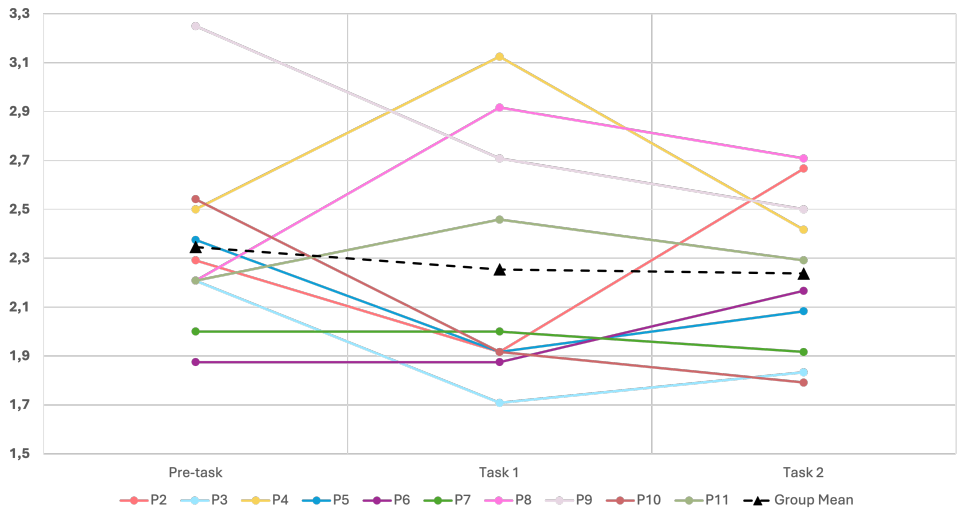


Figure 6.3: SSSQ Results per Participant. The Black line represents the group's mean.

We looked into each participant's stress levels before and after each task to better understand the low scores. Individual SSSQ results are shown in Figure 6.3.

Participants generally did not show a pattern in their stress results; they had varied trajectories from their baseline to both tasks. For example, P3, P5, P9, and P10 decreased stress with tasks, possibly due to familiarisation or engagement. On the contrary, P4, P8, and P11 increased their stress, and P2 had an interesting trajectory

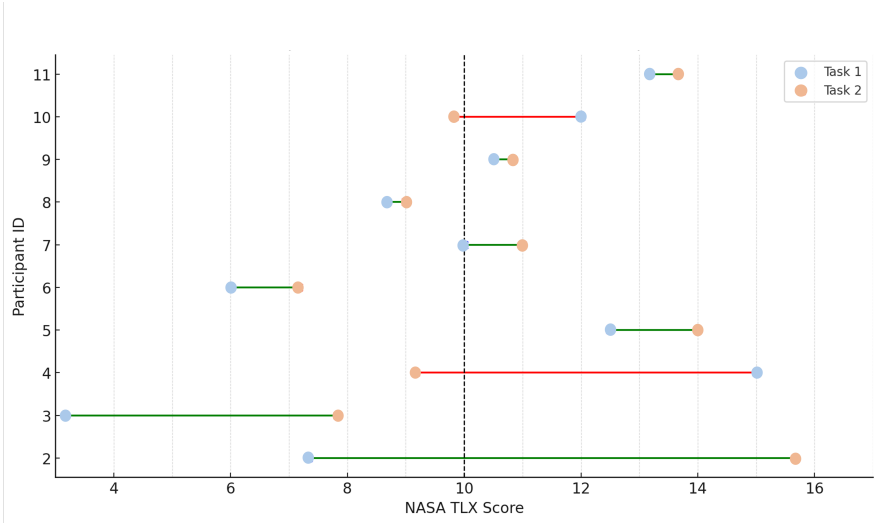


Figure 6.4: NASA-TLX results per participant. The image compares Task 1 and Task 2 scores. Green lines indicate an increase in score from Task 1 to Task 2, while red lines indicate a decrease. The vertical dashed line at score 10 represents the midpoint between low and high perceived workload.

with an initial decrease and finalising with a higher score than the baseline. Finally, P6 and P7 remained generally flat, showing stable stress during the experiment.

The variation in the responses reflects more the influence of individual differences and the task-specific experiences of participants than the tasks and the stressors we tried to add to the whole experiment.

Figure 6.4 shows NASA-TLX scores per participant. The score ranges from 0 to 20, with 10 as the midpoint between low and high workload; higher scores indicate greater mental workload. Most participants reported moderate to high workload in both tasks, with many scores clustering around or above the midpoint (10). Participants 2, 3, 5, 6 and 7 increased from Task 1 to Task 2 (green lines), which suggests Task 2 was more demanding for them. On the contrary, participants 4, 10 and 11 reported a decrease (red lines), indicating that they found Task 2 less demanding than Task 1. Two participants, P8 and P9, presented a minimal change in their scores from Task 1 to Task 2. Overall, the scores imply that Task 2 was perceived as more demanding by most participants, but responses varied considerably.

6.4.2 Biometrics

Our analyses of biometrics aimed at verifying if there are any statistically significant differences in the 15 metrics features we extracted for Task 1 vs. Task 2, and Pre-task condition vs. Task 1. The results of our paired statistical tests revealed that the participants substantially exhibit the same behaviour between the conditions, with only one variable showing significant changes across different experimental conditions (see Table 6.8).

In particular, when comparing Task 1 with Task 2, we observed a significant increase in the EDA phasic peaks per minute between the first and second tasks (Wilcoxon signed-rank test: $W = 4$, $p = 0.01$, $n = 10$), indicating a higher increase in stress during the second task performance [381], as shown in Fig. 6.5. Furthermore, we observed a statistically significant difference in EDA peaks also between the pre-task baseline and the first task. This empirical evidence is in contrast with the self-reported stress, for which we did not observe statistically significant differences across the various experimental conditions.

Comparison	**Metric**	**Statistic**	**N**	**p-value**
First vs Second Task	eda phasic peaks per minute	W = 4	10	0.01
Pre-task vs First Task	eda phasic peaks per minute	T=-4.5	10	0.00

Table 6.8: Statistical results for comparisons between tasks.

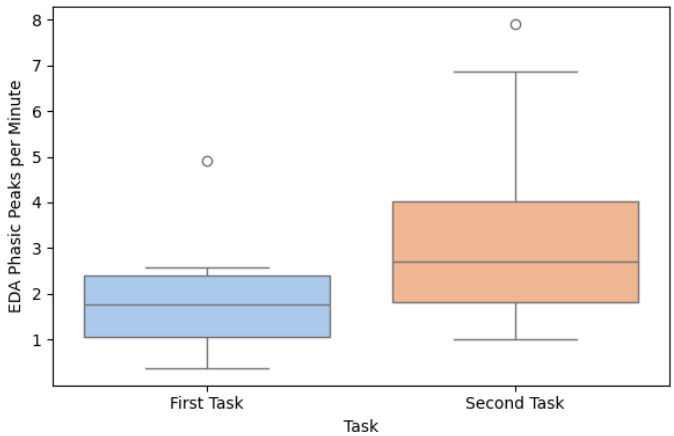


Figure 6.5: Differences between the EDA phasic peaks in the First Task vs. Second Task

We interpreted this contrasting finding as an indication of two possible problems: (i) the misalignment of self-reported and actual stress of participants, with SSSQ and biometrics indicating opposite findings; (ii) a high degree of diversity in the stress experienced by participants, as also suggested by Figure 6.3. To obtain deeper insights and in search of an explanation for this mixed evidence, we conducted a follow-up analysis to verify the alignment between the self-reported stress and the EDA peaks, which we describe in Section 6.5.1.1.

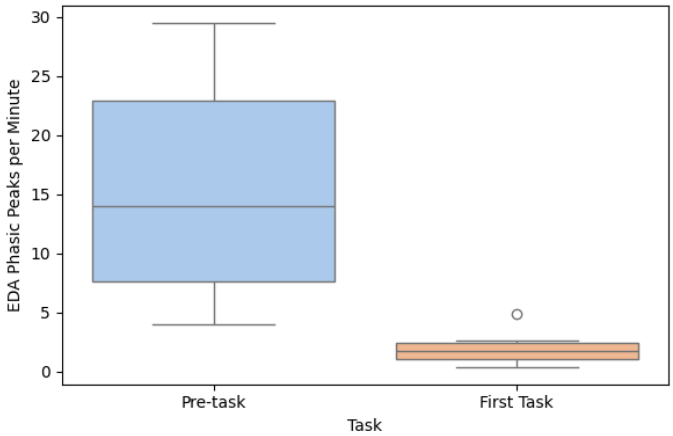


Figure 6.6: Differences between the EDA phasic peaks in the Pre-task vs. First Task

6.4.3 Thematic Analysis

Four themes were generated from the interviews with the ten participants. Below we elaborate on each of them and their corresponding sub-themes.

6.4.3.1 Theme 1: Task Impressions: Engagement and Learning

This theme describes several aspects of the experience, including engagement and skill development. Despite the challenges, many participants found the tasks engaging and appreciated the opportunity to learn and apply their skills. The theme also elaborates on the participants' perception and experience of programming while wearing devices to collect their biometrics.

Sub-theme 1: Perceived Task Structure and Difficulty

This sub-theme captures how participants perceived the tasks, including their clarity, complexity, and how their impressions evolved. In general, perception varied mainly regarding time constraints and clarity. Some found the second task straightforward, with no significant obstacles. As one participant stated:

"Everything was clear, so I didn't have any problems during the second task. Maybe I was feeling that the time was less, but just a little feeling, but nothing else". P6

The quote suggests that the task was well-structured and manageable for some despite the tighter time frame. However, other participants shared their difficulties with time pressure, which influenced their confidence and overall experience. One participant described feeling confident that they would not be able to complete the second task within the given time, leading to a particular feeling:

"In the second time in the second task, I was quite sure that I couldn't complete that in time, so I felt a bit unhappy about that". P2

The quote is a good example of how time constraints, rather than the complexity of the task itself, shaped the perceptions of difficulty. Additionally, some participants noticed structural similarities between the two tasks, which helped them refine their approach to the second one. However, this familiarity did not always mitigate concerns

about time limitations, as some questioned why a more difficult task was allotted less time. These differing perspectives suggest that task difficulty was not purely objective but influenced by individual expectations, time-related pressure, and the ability to adapt strategies based on prior experience.

Sub-theme 2: Engagement and Learning Through Task Progression

This sub-theme focuses on how participants engaged with the tasks, their sense of accomplishment, and the learning process they experienced. Participants' engagement was often shaped by their initial impressions and their ability to adapt. Some participants initially underestimated the complexity of the first task but later recognised deeper patterns that made it more challenging.

For example, Participant 10 reflected on their shift in perception, stating:

"When I read the first one at the beginning, I thought I had all figured it out, but then I read it again, and I saw some patterns that suggested that it was more difficult than I thought, and so maybe in that moment, I felt that it could be difficult and that I already knew that I could not be able to solve it in 20 minutes". P10

For Participant 10, engagement was not static but developed as they reassessed their approach and deepened their understanding of the task. The experience of completing the first task also influenced how participants approached the second. Some found familiarity with the structure helped them engage and work through the task more efficiently.

One participant described, *"Maybe because after completing the first, I reasoned better about the task because they were somehow similar. And the second, it was easier for me to solve it, and it was fine".* P5

The quote reflects how task progression supported learning, enabling participants to refine their problem-solving strategies and improve efficiency. Overall, engagement was influenced by the task challenge and the opportunity to apply and adapt knowledge. While some participants encountered unexpected difficulties that affected their confidence, others found that progressing through the tasks enhanced their ability to approach problems more effectively.

6.4.3.2 Theme 2: Emotional Responses to Challenge and Uncertainty

This theme elaborates on the participants' emotional journey as they navigated the tasks. Among the various emotions participants experienced were frustration, stress, and irritation, particularly when tasks were challenging or when they were uncertain about how to proceed. Time constraint was mentioned as a significant stressor for participants; for example, Participant 11 expressed:

"The only thing that contributed to the stress level was the time. I feel like with the task, 20 minutes was a little like the first task and the second task. 20 and 15 minutes maybe were a little bit too little". P11

Participant 11's quote is interesting because it dismisses other potential stressors like task complexity, unfamiliarity, or difficulty. It also shows how strongly time pressure could affect performance and well-being.

Other participants found the time limit particularly frustrating when encountering difficulties, such as recalling specific programming libraries or debugging errors.

Similarly, participants noted that uncertainty about their solutions contributed to their stress, mainly when they could not test or verify their code.

Despite the challenges, some participants viewed stress as a natural part of problem-solving, accepting moments of frustration as inherent to the coding process.

6.4.3.3 Theme 3: Sources of Distraction and Discomfort

Some factors influenced participants' attention, perception, and, in some cases, emotions. This theme explores how the experimental setup and environment impacted their experience. Elements within the environment disrupted their focus or contributed to discomfort, affecting their ability to fully engage with the tasks. The context of the experiment also played a role. For example, Participant 10 mentioned that they might have performed better if they had been alone. The presence of others seemed to increase their distraction, as they became more aware of how they were performing in comparison

"Maybe I would have performed better if I was alone in the room, because maybe I would have started talking by myself and so on". P10

Internal distractions, such as self-conscious thoughts about performance and concerns about how others were doing, were also noted. For example, Participant 4 mentioned:

"I have some thoughts about others or my or my results, I try to stay focused and I like, like, try to push away the thoughts and concentrate (like how I'm performing, if others are performing, well,)". P4

This participant tried to push these thoughts aside to stay focused, but they remained a persistent challenge.

These distractions and discomforts added to participants' challenges in maintaining attention and emotional balance during the study.

6.4.3.4 Theme 4: Coping Strategies and Adaptation

Participants employed various strategies to manage stress and maintain focus. This theme explains how participants adapted to the challenges of the experiment by using these strategies. They used task decomposition, deliberate attentional control, and actively ignoring feelings and physical actions.

Some participants distanced themselves from the emotional weight of the task by reminding themselves that the experiment was not an exam, thereby reducing performance pressure. Others reported ignoring negative feelings entirely and concentrating on solving the problem instead. Using a structured approach to problem-solving was also mentioned; a specific example is Participant 3, as expressed in the following quote:

"I was thinking about the best way to approach it, like if I should start by defining functions, because that's what I usually do or not, go straight forward to the code without anything in any method at all". P3

P3 decision-making process to approach the task seems to rely on prior experience and habitual strategies, notable in the phrase "because that's what I usually do". This participant showed flexibility in adapting their approach based on the task's demands. Task planning and adaptation add additional mental workload to the tasks, which could also impact our quantitative results. Overall, these adaptive behaviours allowed

participants to mitigate stress and maintain productivity within the experimental setting.

6.5 Discussion

In this section, we discuss how we answered our RQs, the key contributions of this study and the threats to validity.

While we aimed to assess the alignment of psychometric instruments and biometric data and find stress patterns in these data sources, our results did not offer consistent evidence to support clear conclusions. Nonetheless, our results offer indicative insights that may inform future research directions.

To answer our RQs:

RQ1: How reliable are psychometric stress measures compared to real-time biometric indicators (EEG and EDA) during software engineering tasks?

Psychometric results showed, at a general level, moderate stress. Furthermore, there was no increase in stress from pre-tas to Task 1 nor from Task 1 to Task 2. Mental workload results were around the moderate level, too, and showed no significant differences from Task 1 to Task 2. Aligning these results with biometrics, for EDA, only one metric (phasic peaks per minute) showed a statistically significant difference across tasks. However, this single biometric indicator did not consistently align with the psychometric instruments. For EEG, we lost several data points. Hence, we could not analyse it, and we lost that comparison.

Consequently, we cannot draw firm conclusions about the reliability of psychometric measures relative to biometric data. Hence, our findings are only indicative.

RQ2: What stress-related patterns can be identified in real-time biometric data (EEG and EDA) during software engineering tasks?

Since we could not analyse the EEG data, our observations were only on EDA metrics. Phasic peaks per minute were the only metric showing significant differences across tasks, which could suggest a stress-related pattern. However, our results are limited since we did not find any other trends in the rest of EDA features. We do not have robust enough evidence to establish precise or generalisable patterns in this context.

6.5.1 Main Contributions

This study offers the following insights.

6.5.1.1 Evidence Supporting the Alignment of Biometrics and Psychometric Instruments

One of our goals in this study was to find to what extent biometrics align with self-report. Furthermore, the mixed findings observed for the analysis of the biometrics and psychometric instruments call for further analysis of the alignment of the self-reported stress and the biometrics for each participant. In fact, as we report in the previous section, although our quantitative results of SSSQ showed no significant differences in stress levels, we observed variations of EDA peaks across the experimental condition

and, in particular, between pre-task and Task 1 and between Task 1 and Task 2 that suggest the participants might have actually experienced some stress episodes.

In search of an explanation, we performed a follow-up analysis by triangulating self-reported stress, EDA peaks and results of the qualitative coding of interviews. By applying a data triangulation across our data collection methods, we observed that our multi-modal measurement approach was sensitive to the same variations in stress responses. At the group level, there were no significant differences in stress, emotions, and mental workload levels between Tasks 1 and 2 in psychometric and biometric results. Specifically, we decided to look at each participant's behaviour with a focus on stress using self-reported stress, interviews, and EDA peaks.

In Table 6.9, we indicate the self-reported level of stress based on the answers participants provided for the PSS-10 (second column) and SSSQ (third, fourth, and fifth columns). We used the stress levels in the table, mapping the numerical answers to the corresponding levels of each psychometric instrument, for example, for PSS-10 "Low, Moderate and High" stress [416]; and for SSSQ "Not at all, Somewhat, Very much and Extremely" [153]. We remind the reader that the PSS-10 reflects perceived stress over the past month, while the SSSQ captures stress levels pre- and post-tasks. Colour coding indicates stress intensity: green = low/no stress, orange = moderate stress, red = high stress. Finally, in the last column, we report excerpts from the participants' answers during the interviews that pertain to the stress experience and the stress triggers they reported, if any. Interview quotes give contextual insight into individual experiences.

To complement this multi-modal analysis, we triangulate the data in the table with plots of the EDA peaks (see Figures 6.8 and 6.7). The green vertical lines correspond to tags created by participants during the task to contextualize events (e.g., beginning of pre-survey, beginning of baseline, start and end of first task, start and end of second task). The red dots highlight the peaks obtained following the same approach described in Section 6.3.4.2.

Table 6.9: Alignment of psychometric stress measures (PSS-10 and SSSQ) with qualitative interview excerpts across participants.

ID	**PSS-10**	**SSSQ**			**Interviews**
		Pre task	Task 1	Task 2	
P2	Moderate	Moderate	No stress	Moderate	"I was okay during the first task, even though I had the sensation of being unable to solve it, it was more or less okay. The second task was more stressful, I think because it was time-pressured."
P3	Moderate	Moderate	No stress	No stress	"It was kind of fine. I didn't really get stressed. The only part where I was a little stressed was during the code, but not significantly."

ID	PSS-10	SSSQ			Interviews
		Pre task	Task 1	Task 2	
P4	Moderate	Moderate	Moderate	Moderate	"I felt frustrated, stressed because I couldn't solve the problem, I didn't understand what was wrong and that caused me some stress. I wanted to solve it but couldn't."
P5	Moderate	Moderate	No stress	Moderate	"Overall, it was funny because I was curious about my stress level. During the second task I was more focused and that helped. The stress was moderate, nothing overwhelming."
P6	Low	No stress	No stress	Moderate	"I was pretty okay during everything about the task. I didn't feel particularly stressed, but I did notice a slight increase when the second task started, probably due to the pressure to complete it quickly."
P7	Moderate	Moderate	Moderate	No stress	"When I understand what I do, my stress goes down. In the second task, once I figured out the logic, I felt relaxed and enjoyed finishing it."
P8	Moderate	Moderate	Moderate	Moderate	"At the beginning, I didn't understand anything, which stressed me out. But after a while, I got into the flow and things became easier. It was challenging but not too difficult."
P9	High	Moderate	Moderate	Moderate	"I was not stressed at all in the first task because I found it quite simple. But in the second task, I got a bit stuck and it was a bit stressful. Still, I managed to finish."
P10	Moderate	Moderate	No stress	No stress	"I didn't feel stressed because I didn't feel like I was being evaluated. It felt like an exercise more than a test, so I remained calm throughout."

ID	PSS-10	SSSQ			Interviews
		Pre task	Task 1	Task 2	
P11	Moderate	Moderate	Moderate	Moderate	"The only thing that contributed to the stress level was the time. I felt a bit pressured to complete it fast. That made me more focused but also raised the stress a bit."

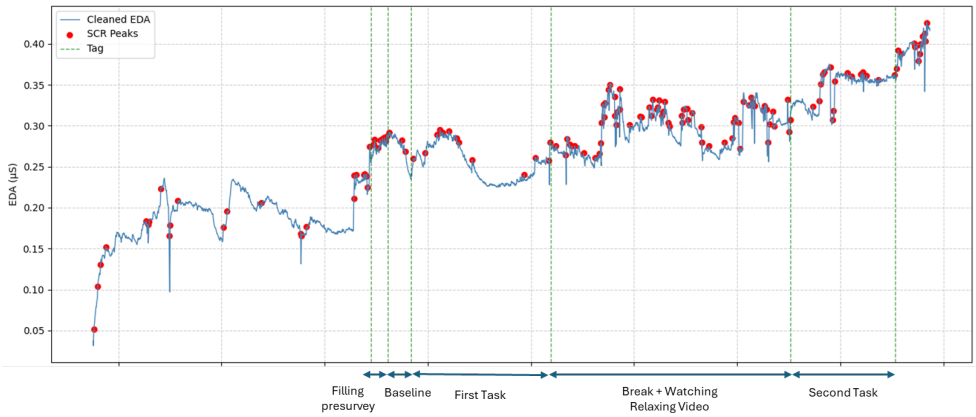


Figure 6.7: EDA signal and its peaks (red dots) across the experimental phases for P9

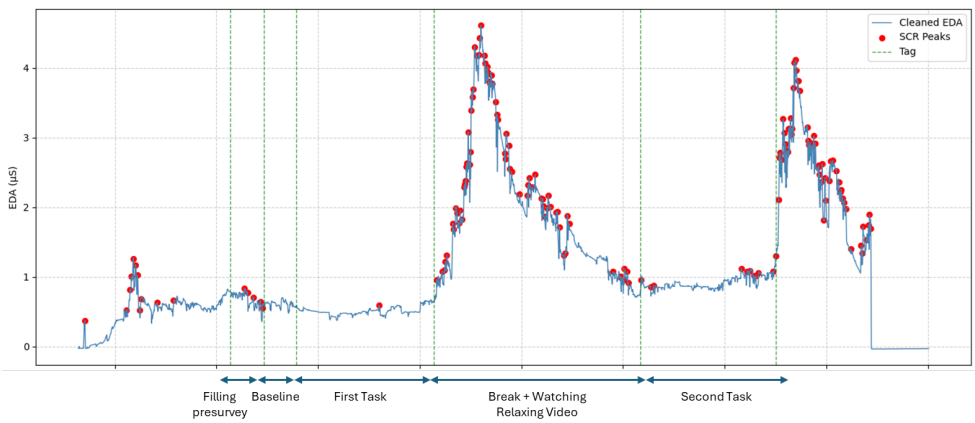


Figure 6.8: EDA signal and its peaks (red dots) across the experimental phases for P3

Looking at Table 6.9, we observe that 4 out of 10 participants (P4, P8, P9, P11) have the same stress level (SSSQ) for the entire study, including the pre-task. This

suggests that the experimental setting did not induce any changes in the self-reported stress compared to the pre-task condition. In fact, when we plot the EDA signal and its peaks for one of these participants (P9), we do not observe significant variations in the signal behaviour, with a slight increase towards the end of Task 2. This aligns with the self-report by P9, saying that '*I was not stressed at all in the first task because I found it quite simple. But in the second task, I got a bit stuck, and it was a bit stressful. Still, I managed to finish*'. Moreover, for P9, Figure 6.7 shows a gradual increase in tonic levels throughout the session, with frequent SCR peaks (marked in red). Regarding the pre-task, EDA is relatively low and stable, with a few SCRs, possibly due to the setting or anticipation. Later, in Task 1, EDA rises slightly but stabilises mid-task. Additionally, several SCR peaks are present but not densely clustered. Physiological arousal is moderate, consistent with SSSQ results. Finally, in Task 2, there is an increase in tonic EDA and dense clustering of SCR peaks, indicating possible heightened stress/arousal during this task. Psychometric results report Task 2 as "moderated" stress, with no significant changes from Task 1, as in the EDA results, and consistent with NASA-TLX (very close to the midpoint) results as well. Considering the interview quote, P9 reflects frustration and helplessness rather than classic stress, explaining why there were no changes in the SSSQ, but it is strongly visible in the EDA.

Similarly, P4, P8, and P11 report a mild experience of frustration or stress during the tasks but do not mention that they experienced different levels of stress across the conditions, which aligns with the consistent 'moderate' stress scores reported throughout the experiment. These alignments help contextualise psychometric scores. For example, participants P4 and P8 consistently reported moderate stress across the SSSQ, and their interview excerpts corroborate the presence of frustration and task-related cognitive effort.

For participants P3 and P10, we observed that they were in a pre-task stress condition, while they were not stressed while coding. This could indicate that, for some participants, stress might be induced by the idea of participating in the study, with stress subsequently dropping down during the actual coding tasks. In fact, P3 reports that '*It was kind of fine. I didn't really get stressed. The only part where I was a little stressed was during the code, but not significantly.*' When looking at P3's EDA plot, we can observe behaviour that aligns with the self-report, with peaks completely absent during both Task 1 and Task 2 (see Figure 6.8). Adding the mental workload results to the analysis (see Picture 6.4), for P3, none of the tasks was challenging; they stayed on the low-demand side for both tasks. Similarly, for P10, even though they were above the median (moderate level), they did not reach a considerably high level of mental demand.

Importantly, this alignment does not suggest perfect correlations but rather lends support to the credibility of our measurements. Research has shown that psychophysiological responses to stress are highly individual and context-dependent [417, 418], and subjective stress awareness may not always be linearly related to biometric signals.

The observed consistency across data sources reinforces their general alignment and the value of this study of using a multi-method approach for stress detection in complex, cognitively demanding tasks like programming.

Difference of Acute and Long Term Stress There are several considerations to consider when measuring a specific type of stress in experiments. Our target was to measure acute stress, which at an emotional level, refers to the appraisal resulting from situations evaluated as threatening and overwhelming based on the individual's available coping resources [419]. Reactions to this type of stress are complex and multidimensional; therefore, assessing its impact requires an equally nuanced approach [419]. We used a measurement approach to capture multiple aspects of acute stress. This included a baseline recording of EEG and EDA before the task, the PSS-10 to assess participants' general stress perception over the past month and the SSSQ pre-task.

Despite these measures, the distinction between acute, subjective and long-term (chronic) stress remains a significant challenge. Biometric tools such as EEG and EDA are well-established for detecting acute stress markers. For instance, EDA reliably reflects sympathetic arousal [167], while EEG patterns can identify emotions, vigilance, mental workload and stress levels [420].

However, longer-term stress exposure can influence biometric signals. Seo et al. [421] found that long and repetitive exposure to stress affects the ability to regulate cortisol levels. Further, since there are relationships between salivary cortisol levels and physiological variables (e.g. heart rate and galvanic skin response), chronic stress may alter baseline autonomic activity, mitigating or distorting the physiological changes typically associated with acute stress responses.

Therefore, even with a pre-task baseline, participants with high long-term stress may have exhibited attenuated or irregular acute stress responses during the task. For instance, a chronically stressed individual may be less responsive to our experimental stressor, leading to smaller physiological and psychometric shifts and potentially contributing to null findings. Moreover, our population in this study, PhD students, have a higher vulnerability to mental health difficulties compared to the general population, with multiple studies indicating elevated levels of anxiety, depression, and overall psychological distress in this group [422], which makes it even more possible the existence of long-term stress. These ongoing stress conditions could have added noise to the biometric and psychometric data. As a result, the biases introduced by chronic stress exposure may have masked clearer patterns of acute stress, limiting the sensitivity of our measures to detect short-term changes.

6.5.1.2 Methodological Insights on Recruitment and Experimental Design

We discussed several insights on the experimental design and implementation of the study and offer suggestions for improvement that we would apply in a future study.

Recruitment The participants in this study were mainly PhD students. As laid out in Sec. 6.5.1.1, PhD students tend to experience a high level of long-term stress and are more prone to stress-related mental health challenges, which may have influenced the results. We are hoping to replicate this study in an industrial practitioner context for further insights. *Suggestion:* Recruit software developers in industry instead of PhD students.

Mental Workload and Stress Both the biometric and the psychometric results show a lack of induced stress in the participants. However, while the participants did not show stress on biometric or psychometric scales, they did report stress in interviews. Hence, there is a psychological component of stress that does show up to some extent in the qualitative data, but less so in the quantitative data (only partially in the NASA-TLX results, see Fig. 6.4, but not to a statistically significant extent), see Tab. 6.7. One possible explanation is that extra effort was exerted to meet the high demands of the task by mobilising extra energy through mental effort [423]. Since our EEG data measurements were insufficient, we cannot compare directly to the results of Mohanavelu et al. [397] or Martínez Vásquez et al. [398]. *Suggestion:* Include additional instruments for differentiating mental workload from stress.

Distractions One participant mentioned that the printer starting a job was distracting. Furthermore, that sometimes led other people to enter the room. While such distracting factors, e.g., other people in the room, take away from the setting of a controlled experiment, they can be linked to a more realistic setting than participants being in a room by themselves. Hence, it leads to a better representation of a real-world scenario. *Suggestion:* Control for distractions in replication.

Participant Motivation The participants' motivation was probably not strong enough. Since participants did not have pressure to do this task well, as it had no consequences for them, this could be an indicator of why they did not get stressed as expected by the shortened time available to them for Task 2. *Suggestion:* Pick a task that the participants care about and want to see through.

Increased Stimuli to Induce Stress Other ways of inducing stress, e.g., a simulated power outage, switching off the light, loud noise, or pretending this is their exam, would be unethical. However, if the reason the time pressure failed to induce stress is really due to a lack of motivation of the participants to succeed, then none of these are likely to make a difference. *Suggestion:* Introduce a stronger incentive through a higher remuneration if a task is completed.

Technical Challenges of Sensors The practical and technical feasibility of using biometric sensors in real-world or semi-controlled software engineering tasks is limited due to the fact that sensors are not the most robust or reliable. In combination with a limited time slot that the participants were booked for, this did not allow much room for error. If it was not detected immediately when a sensor was not working, we lost data on that participant. *Suggestion:* Plan more buffer time.

6.5.1.3 General Lessons for Conducting Stress Studies

Our study, despite yielding negative results, shed light on methodological challenges in eliciting measurable stress responses within the context of software engineering experiments.

The reduction of task completion time from 20 to 15 minutes failed to elicit measurable stress responses, suggesting that time pressure alone may not serve as an adequate stressor for experienced programmers. As one participant remarked, "The

only thing that contributed to the stress level was the time," indicating that while time constraints were perceived, their impact was minimal. This aligns with prior research showing that time pressure alone often falls short of inducing significant stress responses, particularly among experienced individuals, unless combined with high-stakes outcomes or contextual disruptions [424].

Future studies could explore a multi-stressor approach, integrating time constraints with performance evaluations or unpredictable interruptions [425], to better replicate real-world stress conditions.

> Key takeaway 1: Time pressure alone may be insufficient; combine multiple stressors for more reliable stress induction.

The variability in individual self-reported stress responses (illustrated in Figure 6.3) highlights the importance of calibrating stress induction protocols through pilot testing. Participants who quickly understood tasks and found them engaging reported different experiences than those who struggled. This phenomenon mirrors what Csikszentmihalyi described as the "flow state," [426] where optimal engagement occurs when challenge levels match individual skills. Peifer et al. [427] further established connections between flow experiences and moderate stress levels, suggesting an inverted U-shaped relationship between stress and performance. Our findings suggest that matching the task complexity to participants' skill levels might help ensure enough challenge levels, thus inducing stress.

> Key takeaway 2: Adapt task difficulty to the skill of the participants.

6.5.1.4 Ethical Reflections on Stress Induction in Research

Inducing stress in controlled experiments presents a fundamental ethical tension: balancing methodological rigour with participant well-being. We employed a protocol to induce stress in participants, enabling its manipulation as an independent variable. Combining biometric (EEG, EDA) and psychometric measurements, we aimed to establish causal relationships between stress and software engineering task performance while reducing reliance on self-reported data.

Following ethical guidelines to avoid harm and long-term negative effects to our participants [428], we induced moderate, short-term stress (enough to observe effects without harming participants) by limiting the time for the second task. However, there was a risk that this stressor might be insufficient to produce measurable results. This was exactly what happened; our results did not show any stress in our data, meaning that our protocol may have fallen below the threshold needed for observable impact.

Adding to the previous, another challenge is that individuals perceive and respond to the same stressors differently. This is evident in Figure 6.3, for example, there is no clear pattern in our participants' responses to the stressor. We took this variability to reinforce the importance of post-experiment care and implemented the relaxation video at the end of the experiment to mitigate short-term discomfort [429].

Furthermore, ethical guidelines also demand careful cost-benefit analysis [429]. In this case, we might need to create more or harder stressors to get different results; however, imposing discomfort on our participants and offering them little or no immediate benefit will violate the ethical guidelines [430]. Our results denote that ethically constrained stress induction in programming experiments may be fundamentally untenable, either too weak to yield actionable data (as here) or so intense it crosses ethical boundaries.

Finally, our null results invite reflection on the trade-offs between ethical boundaries and experimental validity. Future work could explore alternative stressors (e.g., time pressure in real work environments).

6.5.2 Threats to Validity

6.5.2.1 Internal Validity

We employed a within-subject design without a control group, which limits the ability to draw causal conclusions. Additionally, there are confounding variables that we could not control and might have affected the overall results. For example, participants' prior experience with programming tasks, fatigue, long-term stress or stress unrelated to the task, could have influenced subjective and biometric responses. We tried to mitigate this using data triangulation, specifically adding an interview at the end to get participants' experiences and impressions.

Furthermore, the participants' motivation to finish or develop the task successfully could have been influenced by the lack of consequences for failure. Motivation is a highly complex variable to control without crossing ethical boundaries. We tried to mitigate this threat by explaining to the participants the importance of completing the tasks in time.

6.5.2.2 External Validity

This study was thought to be a pilot for future interventions in companies. Hence, we are aware that it is challenging to generalise findings. The number of participants is limited, and they all come from a specific population (PhD students from one university). Future studies need to recruit a more varied sample, including different levels of academic experience, diverse genders and backgrounds, and professionals from the industry to account for a more representative population. Additionally, the artificial nature of lab-like task environments may not fully replicate real-world programming stressors (further discussed in the Ecological validity section).

6.5.2.3 Construct Validity

Our small sample size limits statistical power, especially when interpreting correlations or changes across time points (for example, when comparing the change from Task 1 to Task 2). This issue was compounded by the loss of EEG data points, which reduced the sample size. While the data triangulation adds credibility, conclusions about the efficacy of biometric stress measures or the interpretability of psychometric data must be made cautiously. There is also a risk of confirmation bias in interpreting the

alignment between methods, especially when expected outcomes may unconsciously influence how data is coded or analysed.

Moreover, the type of sensors we used in this study impacted the amount and quality of collected data, specifically the EEG data. The sensor was not entirely reliable, and we lost several data points. Additionally, biometric data is influenced by environmental noise, physical movement, or individual physiology [431].

6.5.2.4 Ecological Validity

Although the study attempted to mimic realistic programming tasks in a daily work scenario, the experimental setting may have induced behaviour not reflective of natural work environments (e.g., being observed or monitored may have influenced stress levels, as one participant commented in the interview). Participants may also have responded differently, knowing there were no consequences if they did not complete the assignment. This lack of real-world accountability may have reduced the urgency or perceived importance of the task, potentially leading to lower stress levels than would be experienced in high-stakes professional contexts. Consequently, the emotional and cognitive responses observed in the study may not fully represent the stress experienced during typical workday demands, deadlines, or performance pressures.

6.6 Conclusion

6.6.1 Summary

In this article, we presented an experimental study to compare psychometric stress measures and biometric indicators and identify stress-related patterns in biometric data during software engineering tasks.

Ten participants wearing biometric sensors performed two tasks, whereby the second task had a stricter time limit.

This limitation did not stress the participants significantly, so our results remain only indicative in terms of confirming or refuting the validity of a comparison of psychometric, biometric, and qualitative data.

6.6.2 Future Work

We are considering three lines of future work:

[a] **Replication:** We are planning to replicate this study with a larger group of software developer participants in industry.

[b] **Personality and Experiences:** We are curious to explore how individual differences (e.g., personality traits, prior experiences) influence stress response and coping mechanisms.

[c] **Stress, Motivation & Performance:** We are designing a study to investigate the relationship between stress, motivation, and performance.

6.7 Declarations

6.7.1 Funding:

The research of Daniela Grassi is partially funded by D.M. 352/2022, Next Generation EU - PNRR, in the scope of the project "Recognition of emotions of cognitive workers using non-invasive biometric sensors", co-supported by Exprivia, CUP H91I22000410007. This research was co-funded by the NRRP Initiative, Mission 4, Component 2, Investment 1.3 - Partnerships extended to universities, research centres, companies and research D.D. MUR n. 341, 15.03.2022 – Next Generation EU ("FAIR - Future Artificial Intelligence Research", code PE00000013, CUP H97G22000210007), and the Complementary National Plan PNC-I.1 - Research initiatives for innovative technologies and pathways in the health and welfare sector - D.D. 931 of 06/06/2022 ("DARE - DigitAl lifelong pRevEntion initiative", code PNC0000002, CUP B53C22006420001).

6.7.2 Ethical approval:

At the time of planning the experiment, formal ethical approval was not required for this type of study. Since then, new procedures have been introduced, and we have submitted our application accordingly. Following a positive preliminary review, we are currently awaiting final approval. Protocol ID: CER_19720E5F292

6.7.3 Informed consent:

We obtained signed informed consent for all participants in the study.

6.7.4 Author Contributions [all authors should be mentioned]

Cristina Martinez Montes: Conceptualisation, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - Original Draft, Writing - Review and Editing, Visualization, Project administration.

Daniela Grassi: Conceptualisation, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - Original Draft, Writing - Review and Editing, Visualization.

Nicole Novielli: Methodology, Validation, Writing - Original Draft, Writing - Review and Editing, Supervision, Project administration, Funding acquisition

Birgit Penzenstadler: Methodology, Validation, Writing - Original Draft, Writing - Review and Editing, Supervision, Project administration, Funding acquisition

6.7.5 Data Availability Statement

The anonymised quantitative data collected are available in our repository: https://doi.org/10.5281/zenodo.15497559. Qualitative data is not openly available due to reasons of sensitivity.

6.7.6 Conflict of Interest

The authors declare that they have no conflict of interest.

6.7.7 Clinical Trial Number in the manuscript.

Clinical trial number: not applicable.

Acknowledgements

We thank Robert Feldt for contributions in designing this study and for supportive discussions during the analysis procedure. We thank our study participants for their time and effort.

Chapter 7

Paper F:

Large Language Models in Thematic Analysis: Prompt Engineering, Evaluation, and Guidelines for Qualitative Software Engineering Research

C. Martinez, R. Feldt, C. Miguel, S. Ouhbi, S. Premanandan, D. Graziotin

Under submission to Transactions on Software Engineering (TSE), 2025

Abstract

As artificial intelligence advances, large language models (LLMs) are entering qualitative research workflows–yet no reproducible methods exist for integrating them into established approaches like thematic analysis (TA), one of the most common qualitative methods in software engineering research. Moreover, existing studies lack systematic evaluation of LLM-generated qualitative outputs against established quality criteria.

We designed and iteratively refined prompts for Phases 2—5 of Braun and Clarke's reflexive TA, then tested outputs from multiple LLMs against codes and themes produced by experienced researchers. Using 15 interviews on software engineers' well-being, we conducted blind evaluations with four expert evaluators who applied rubrics derived directly from Braun and Clarke's quality criteria. Evaluators preferred LLM-generated codes 61% of the time, finding them analytically useful for answering the research question. However, evaluators also identified limitations: LLMs fragmented data unnecessarily, missed latent interpretations, and sometimes produced themes with unclear boundaries.

Our contributions are threefold. First, A reproducible approach integrating refined, documented prompts with an evaluation framework to operationalise Braun and Clarke's reflexive TA. Second, an empirical comparison of LLM- and human-generated codes and themes in software engineering data. Third, guidelines for integrating LLMs into qualitative analysis while preserving methodological rigour —clarifying when and how LLMs can assist effectively and when human interpretation remains essential.

7.1 Introduction

Recent advances in artificial intelligence (AI) have enabled Large Language Models (LLMs) to process vast amounts of text and uncover complex patterns with remarkable speed [432–434]. These capabilities make LLMs especially promising for qualitative data analysis (QDA).

While software engineering (SE) research has traditionally emphasised quantitative and experimental methods [435–437], the discipline is increasingly recognised as social, multidisciplinary, and deeply human [311, 438, 439]. Understanding software development demands attention to real-world contexts [440] and to the interplay of technical, managerial, and organisational factors [441]. As a result, qualitative methods—such as grounded theory, thematic analysis (TA), and content analysis—have gained growing traction in SE research [73, 439].

As qualitative methods gain prominence and LLMs emerge as tools for qualitative data analysis (QDA), research in this area has expanded rapidly [29, 112, 442]. Yet, several limitations and concerns persist.

Limited transparency and explainability. Many SE studies fail to report key details such as prompts, model configurations, or evaluation procedures, hindering transparency, reproducibility, and interpretability [443].

Lack of systematic evaluations on SE data. Most existing work focuses on comparing LLM outputs with human-coded results [434], overlooking other vital quality dimensions—such as the coherence and depth of themes, the transparency of coding decisions, and the usefulness of the resulting insights for addressing research questions.

Insufficient methodological rigour. Many studies adopt ad hoc or poorly justified procedures, applying LLMs to QDA without grounding their approach in established methodological frameworks. Few validate their processes against accepted qualitative standards or employ systematic checks for reliability and interpretive depth [444, 445]. In contrast, our study followed Braun and Clarke's reflexive thematic analysis (TA) framework [75] and provided detailed documentation of each step.

Narrow model scope. Prior studies often rely on a single LLM, which limits the generalisability of their findings. This also leaves open questions about how model choice influences coding quality and interpretive outcomes. Our study addressed this by evaluating several leading models. We tested different LLMs and evaluated their performance supporting more robust conclusions about LLM performance.

Beyond methodological shortcomings, researchers have also highlighted ethical and privacy risks, particularly when working with sensitive data [442].

This study addresses these gaps by empirically evaluating the use of LLMs (ChatGPT 03 mini, GPT-4o, Gemini 2.5 Pro, and Claude 4 Sonnet) in TA. TA is one of the most widely applied methods for qualitative research in SE [73]. We systematically compared human and LLM-generated codes and themes between March and July 2025. We evaluated [1] them using rubrics derived from Braun and Clarke's reflexive thematic analysis framework [75]. To enhance reliability, we iteratively refined and fully disclosed prompts to ensure transparency and reproducibility.

[1]The evaluators' role was limited to only reviewing and rating the codes and themes. All prompting, model runs, and refinement of outputs were carried out by the first and second authors.

Our study also advances methodological practice by operationalising Braun and Clarke's reflexive TA for use with LLMs. We offer practical guidelines for integrating AI in ways that support rather than replace researcher reflexivity.

This study answered the following research question:

RQ: To what extent can LLMs perform phases of reflexive TA in a way that aligns with established qualitative research standards?

Our main contributions are:

- A reproducible framework combining prompt design and rubric-based evaluation for applying reflexive thematic analysis with LLMs.
- An empirical comparison of LLM- and human-generated codes and themes in software engineering data.
- Guidelines for integrating LLMs into qualitative analysis workflows to enhance efficiency while preserving reflexivity and methodological rigour.

This study adds to the methodological discussions in empirical SE by clarifying the possibilities and the limits of AI-assisted qualitative research.

7.2 Background and Related Work

This section provides the background of the study on Thematic Analysis and the related work on using AI and LLMs in qualitative data analysis, particularly in TA.

7.2.1 Qualitative Data Analysis in Software Engineering

QDA in SE allows a deep exploration of non-technical aspects of software development [71]. Researchers find patterns, meanings, and insights by systematically interpreting rich, non-numerical data [71, 72]. QDA is particularly useful for gaining a deep understanding of software processes, tool use, and organisational or technical settings. In these cases, interpretation and contextual insights are crucial for advancing theory and practice in SE. Additionally, SE qualitative datasets often blend technical artefacts. For example, code review comments, architecture decision records, incident chats with human-centred sources (interviews, field notes). This emphasises the importance of scale management, cross-analyst consistency, and traceable decision trails for credibility.

7.2.1.1 Thematic analysis

We chose TA for this study because it is one of the most popular data analysis methods within SE research [73, 74]. We adapted the version by Braun and Clarke, Reflexive Thematic Analysis [75], for collaboration with LLMs. The six phases are detailed next:

- Phase 1. Familiarisation with the Data: Reading and re-reading the data is required to understand the content fully.

- Phase 2. Generating Initial Codes: The goal is to systematically identify and label data segments that present ideas or concepts that could help answer the research question.
- Phase 3. Generating Initial Themes: The aim is to cluster codes with similar meaning into candidate themes with patterns and broad ideas.
- Phase 4. Developing and Reviewing Themes: It extends phase 3 and does a vital check to review and explore the initial clusterings to find a better pattern development based on the research question.
- Phase 5. Refining, Defining and Naming Themes: The final themes are refined by determining the structure and flow of the analysis. It also requires writing a definition and naming them in a way that represents their content and central idea.
- Phase 6. Writing up the analysis: Involves explaining the findings in themes to answer the research questions coherently and effectively. It also includes selecting and using extracts from the data to illustrate the core parts of the themes.

7.2.2 Early AI and ML in Qualitative Data Analysis

Recent advancements in AI have sparked a growing interest in leveraging its application to qualitative data analysis. For example, Liew et al. [446] proposed a method that involves natural language processing (NLP) and machine learning (ML) to generate initial codes, which are subsequently refined by human input. Similarly, various other studies have used NLP to derive potential codes [447–450]. Other studies have instead outlined challenges of implementing ML for qualitative coding [451].

In a similar line, Towler et al. [452] proposed Machine-Assisted Topic Analysis (MATA). This is an NLP approach that combines human input with automated analysis to summarise text patterns more efficiently. MATA's features make it valuable for qualitative researchers handling large datasets. Compared to traditional TA, MATA is less time- and resource-intensive, aiding in early familiarisation and coding. A similar tool is LaMa [453], short for machine labelling. It is a web application that facilitates the handling and tracking of labels and changes. It makes it easy for researchers to group labels, create themes, and collaborate. However, unlike our study, these tools do not generate codes or themes. Our approach allows the LLM to propose codes and themes. Having initial codes provides an initial analytical layer for researchers and supports a more comprehensive initial capture of the data.

While traditional supervised and unsupervised ML techniques have been widely employed in qualitative analysis [454–460], significant gaps remain in addressing challenges such as privacy, model bias, quality control, and reproducibility [442]. In this study, we addressed these concerns by implementing stepwise human oversight, transparent prompt design, and systematic evaluation procedures.

7.2.3 LLMs in Qualitative Analysis

Several studies have investigated LLMs' role in supporting QDA. A first set of works has explored the potential of LLMs for inductive or deductive coding when doing TA. De Paoli [29] explored the application of ChatGPT 3.5-Turbo to conduct inductive TA in semi-structured interviews. Using open-access interviews previously analysed by human researchers, De Paoli showed the capacity of LLMs to infer main themes from prior research contexts. Moreover, the study emphasises the LLMs' capability to identify relevant themes that might have eluded human analysts. However, methodological rigour across TA phases was not assessed. Our study evaluated analytic quality and coherence across Phases 2–5 using a rubric-based assessment and disclosed prompts. In the same way, Xiao et al. [28] focused on evaluating the deductive coding agreement between LLMs and human analysts. They also investigated how the design of prompts influences analysis outcomes. Their focus was on agreement and prompt effects; we additionally test interpretive depth, theme coherence, and reflexive alignment with the RQs.

Wen et al. [112] extended this line of inquiry, testing LLMs to perform inductive and deductive coding with a large-scale case study in the charity sector. They achieved strong semantic alignment with human coding and sentiment analysis, yet also found inconsistencies in excerpt extraction and the heavy need for human validation. In our approach, We mitigated these issues by coding interviews segment by segment and integrating continuous human feedback rather than relying on post-hoc validation.

Beyond coding, researchers have also explored how LLMs might contribute to higher-level analytic tasks. Tabone and de Winter [461] showed that GPT-generated sentiment ratings and summaries were often consistent with human outputs. However, results varied depending on the prompt, and GPT sometimes produced themes absent in human analyses. Their outputs may be useful, but they risk inconsistency or distortion without systematic evaluation. Our work focuses on the analytical core of TA. We incorporated human feedback to ensure depth and trustworthiness while maintaining the final interpretative step as fully human as possible.

These previous studies have proven that LLMs can assist with coding, summarisation, and collaborative workflows. Still, most studies focused on isolated tasks, expressed prompt sensitivity, or stopped short of assessing analytic rigour. Our study contributes to the existing body of literature and addresses these limitations. We applied human and rubric-based evaluations between steps and embedded transparency and reflexivity in the prompts.

7.2.4 Generative AI Tools and Frameworks in QDA

Researchers have started to propose tools and frameworks to integrate LLMs as collaborators in QDA. For example, CollabCoder [113] is a one-stop, end-to-end workflow used for inductive coding. It offers AI-generated code suggestions, facilitates iterative discussions using quantitative metrics, and provides primary code suggestions for creating codebooks. However, this tool focuses only on one specific type of coding (inductive) and does not consider the RQs, which are the drivers in qualitative analysis. In our study, our prompt included the RQ. Moreover, the tool only addresses three analysis steps; in contrast, our study also covers the themes' creation step. In addition, CollabCoder run under certain assumptions that are not always true when conducting

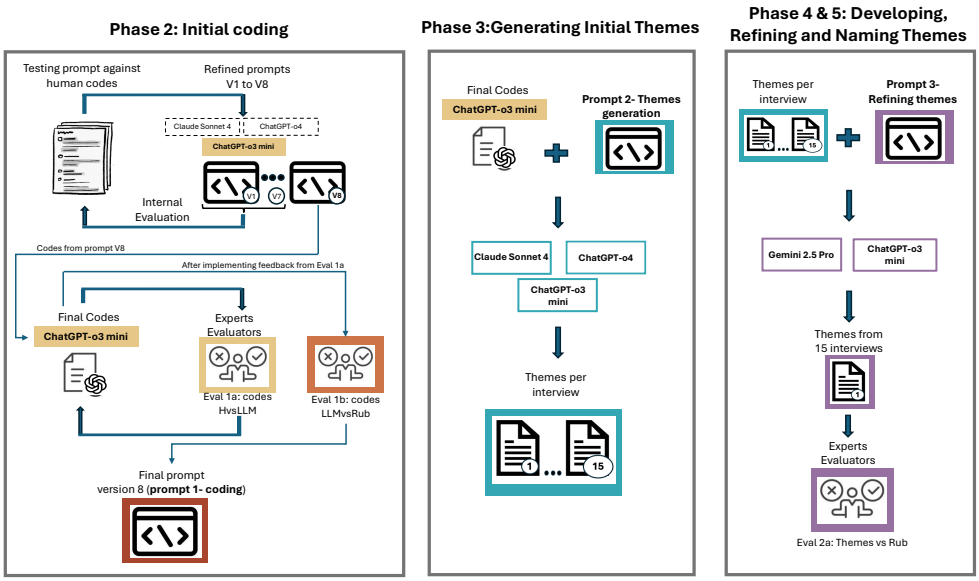


Figure 7.1: Study's methodology overview. The dataset was coded in Phase 2 with Prompt 1, evaluated by experts, and refined iteratively to obtain the final version (v8). In Phase 3, Prompt 2 generated themes per interview, and in Phases 4–5, Prompt 3 created themes across 15 interviews. Final outputs from different model pipelines were again reviewed by experts using Braun and Clarke's quality criteria.

TA (the presence of two coders, raw data consisting of semantically distinct units, and data segments each conveying a single meaning). Our tool is not tied to a certain number of researchers and can handle any type of raw data. For example, interviews, where a single segment can have more than one meaning.

In a similar line, Gebreegziabher et al. proposed Patat [114]. This tool learns patterns from user-annotated codes and recommends new ones. It also creates codebooks and helps the user learn data characteristics. Patat improves explainability compared to CollabCoder by showing users what the model has learned. In line with this, our prompt also explains codes and themes. A limitation of Patat is that it supports only one user at a time, which is problematic since qualitative analysis is typically conducted collaboratively by multiple researchers. Although it offers several features, its creators acknowledge that this makes the tool complex and challenging to learn. Unlike Patat, our tool supports collaborative analysis, extends across multiple TA phases, and evaluates analytic quality using reflexive rubrics.

Drápal and Savelka's [462] framework, designed for legal experts to collaborate with OpenAI's GPT-4 model, is the closest to our study, although in a different field. They covered phases 2 (generating initial codes), integrating users' feedback, 3 (searching for themes), and 4 (generating initial themes). However, their work was restricted to a legal domain and stopped short of evaluating methodological. By contrast, our study not only covered the same steps but also added explanations of

the model decisions to improve trust. We also tested several models in parallel for comparison, and incorporated extra tasks such as defining themes, mapping codes to themes, and flagging sensitive segments. Furthermore, unlike Drápal and Savelka's framework, we embedded step-by-step evaluations with human experts based on Braun and Clarke's guidelines. Additionally, we refined prompts iteratively throughout the process to ensure transparency and reproducibility.

Taken together in our study, we integrate systemic human evaluation using tailored rubrics based on Braun and Clarke's TA framework, making the assessment rigorous. Furthermore, we engineered prompts for steps 2 to 5 in reflexive TA and fully disclosed them, thereby advancing transparency and methodological clarity beyond prior work.

7.3 Methodology

This section elaborates on the steps we followed in the study, the design, implementation, and evaluation. We adopted a mixed-methods approach, combining qualitative and quantitative data collection. We aimed to assess the capabilities of LLMs in conducting TA following Braun and Clarke's framework.

7.3.1 Dataset

Our dataset consisted of 15 semi-structured interviews that explored factors that influence the well-being of software engineers, collected from a previous study [**?**]. Each interview lasted between 40 and 75 minutes and was audio-recorded, transcribed, and anonymised. The transcription resulted in 177 pages in Word using a font size of 12. Two researchers previously inductively coded the interviews following Braun and Clarke's guidelines [75]. We used this set of codes (human-generated) to compare with the LLM codes at multiple stages.

7.3.2 Study Design: Mapping TA Phases to Human vs LLM Roles

The first two authors (referred to as "we" from now on) structured the study to match Phases 2 to 5 of Braun and Clarke's Reflexive TA framework (see Figure 7.1 for the overview of the experiment):

- Phase 2: Generating Initial Codes
- Phase 3: Generating Initial Themes (per interview)
- Phase 4: Developing and Reviewing Themes
- Phase 5: Refining, Defining, and Naming Themes

We left phases 1 (Familiarisation) and 6 (Producing the Report) entirely human-led. However, we implicitly embedded familiarisation within the initial prompting of Phase 2 by providing the LLM interview and problem context. Phase 6 requires human interpretation, as it involves contextualising themes and ensuring alignment with the RQs and qualitative standards. The final write-up must be grounded in

Prompt 1- coding	
Core element	**Instruction**
Role & expertise	"You are a world-class SE researcher ..."
Study's goal	"identify the factors that influence the well-being"
RQ	"What factors influence the well-being..."
Methodological orientation	Inductive · Semantic · Experiential
Coding principles	Succinct, specific, gerunds, informative
Evaluation criteria	Clarity, relevance, semantic, specificity, theme potential, alignment, labels
Context	Context, for the reply below:
Output format	JSON with fields: segment, codes, explanation, sensitive
Response to be coded	"Response, that you should code:"

Figure 7.2: Core elements and short examples of prompt 1, final version for creating codes.

Prompt 2- Themes generation	
Core element	**Instruction**
Role & expertise	"You are a world-class SE researcher ..."
Goal and scope	Generate a set of initial themes from a single...
RQ	"What factors influence the well-being..."
Methodological orientation	Inductive · Semantic · Experiential
Instructions	1.Generate initial themes 2...
Internal Quality Check	"Perform a hidden self-assessment...('Excellent' column from rubric)"
Output format	Return a single JSON object with a key "themes"

Figure 7.3: Core elements and short examples of prompt 2 for creating themes per interview.

Table 7.1: Rubric used to evaluate the quality of initial codes generated by LLMs during Phase 2 of thematic analysis. The rubric includes eight criteria adapted from Braun and Clarke's guidelines, each rated on a 4-point scale from "Poor" (1) to "Excellent" (4).

Criteria	Excellent (4)	Good (3)	Fair (2)	Poor (1)
Clarity of Meaning	Codes are exceptionally clear, specific, and unambiguous, capturing distinct and well-defined meanings within the data.	Codes are mostly clear and specific, with minor ambiguities that do not hinder understanding.	Codes are somewhat unclear, leading to some ambiguity in meaning and interpretation.	Codes are unclear, vague, or ambiguous, failing to capture distinct meanings.
Relevance to Research Question	Codes are highly relevant, directly addressing and reflecting the research question with strong alignment to the data.	Codes are mostly relevant, with a few minor deviations from the research question.	Codes are somewhat relevant, but fail to capture important aspects of the research question fully.	Codes are largely irrelevant, showing little to no connection to the research question.
Balance of Latent and Semantic Meanings	Codes effectively capture surface-level and deeper meanings, demonstrating a strong balance between the two.	Codes capture either surface-level or deeper meanings effectively, but not both equally.	Codes focus primarily on surface-level meanings, neglecting deeper insights.	Codes fail to capture both surface-level and deeper meanings, lacking depth.
Specificity	Codes are precise, capturing narrow and distinct meanings that do not overlap with other codes.	Codes are mostly precise, with occasional overlaps that may cause some confusion.	Codes lack precision, with significant overlaps leading to unclear distinctions.	Codes are imprecise and broad, with substantial overlap, making distinct meanings unclear.
Potential for Theme Development	Codes provide a robust foundation for meaningful theme development, reflecting diverse insights.	Codes mostly support theme development, but may lack some diversity in insights.	Codes provide limited potential for theme development, lacking diversity and clarity.	Codes do not support theme development, reflecting a narrow range of insights.
Alignment with Data	Codes are closely aligned with the dataset content, accurately reflecting the meaning of the data.	Codes are mostly aligned, with minor discrepancies in reflecting the data's meaning.	Codes show some misalignment with the data, leading to inaccurate representations.	Codes are poorly aligned, failing to reflect the dataset's meaning accurately.
Good Labels	Code labels offer a concise, pithy, and insightful shorthand for broader ideas, enhancing understanding.	Code labels are mostly concise and insightful, but some could be improved for clarity.	Code labels are somewhat vague or overly broad, lacking clarity in labelling.	Code labels are unclear and lengthy, failing to provide effective shorthand for ideas.
Explanation of Interview Segment Selection	The explanation is exceptionally clear and logical. It directly and explicitly connects the coded interview segment to the research questions and convincingly demonstrates its importance to the main topic. The reasoning is thorough and well-justified, leaving no ambiguity.	The explanation is clear and mostly logical, with minor areas that could be more detailed. It connects the coded interview segment to the research questions and shows its importance to the main topic; however, the link could be more explicit or compelling.	The explanation is somewhat clear but confusing in parts. It partially connects the coded interview segment to the research questions and mentions its importance to the main topic; however, the relevance and justification are weak or underdeveloped.	The explanation is unclear, disjointed, or difficult to follow. It fails to connect the coded piece to the research questions or demonstrate its importance to the main topic. The reasoning is missing, vague, or irrelevant.

Prompt 3- Refining themes	
Core element	**Instruction**
Role & expertise	"You are a world-class SE researcher ..."
Goal and scope	Your goal is to merge and then rank themes...
RQ	"What factors influence the well-being..."
Methodological orientation	Inductive · Semantic · Experiential
Task specification	1. Identify, merge, and (if needed) split themes...
	2. Theme structure...
Output format	Return a single JSON object with a key "themes"

Figure 7.4: Core elements and short examples of prompt 3 for creating themes, including 15 interviews.

a solid theoretical foundation and developed through a rich, rigorous interpretative process [75]. This makes the task more challenging for LLMs. We used four LLMs across the study. Table 7.3 presents the characteristics of each model and indicates the phases in which they were implemented.

7.3.3 Prompting Strategy, Application and Evaluation

We constructed, tested, evaluated, refined and rewrote several prompts for each phase (see the prompts in the online appendix [463]). The evaluations combined quantitative tools, such as rubrics, with qualitative data from the evaluators' comments.

We created two rubrics based on Braun and Clarke's quality benchmarks for TA outputs. The rubric for evaluating the quality of the codes from Phase 2 is presented in Table 7.1. It follows Braun and Clarke's guidelines for Phase 2 of reflexive thematic analysis ('generating initial codes') [75]. It includes eight criteria: clarity of meaning, relevance to the research question, balance of latent and semantic meanings, specificity, potential for theme development, alignment with data, quality of code labels, and explanation of interview segment selection. The rating has a four-point scale from "Poor" (1) to "Excellent" (4).

Meanwhile, the rubric for evaluating the quality of themes is shown in Table 7.2. It defines eight criteria: coherence, relevance, boundary clarity, data support, definition, naming, analytical contribution, and use of subthemes. Each criterion is rated on a four-point scale from "Poor" (1) to "Excellent" (4).

The rubrics facilitated a systematic, transparent, and comparable evaluation of codes and themes against qualitative research standards. The combination of data supported that assessments were standardised. The whole process we followed is

Table 7.2: Rubric to evaluate the quality of themes. The rubric includes eight criteria adapted from Braun and Clarke's framework: coherence, relevance, boundary clarity, data support, definition, naming, analytical contribution, and use of subthemes. Each criterion is rated on a 4-point scale from "Poor" (1) to "Excellent" (4).

Criteria	Excellent (4)	Good (3)	Fair (2)	Poor (1)
Central Organising Concept and Conceptual Coherence	The theme has a coherent, clear, distinct, and well-defined central organising concept that seamlessly ties all data and codes.	The theme has a central organising concept that ties most data and codes together, with minor gaps in coherence.	The theme has a central organising concept, but is somewhat vague or inconsistently applied.	The theme lacks a coherent, clear central organising concept.
Meaningfulness and Relevance	The theme captures something highly meaningful and relevant to the research questions.	The theme captures something meaningful and relevant, but the connection to the research questions could be more explicit or detailed.	The theme captures some meaningful aspects, but its relevance to the research questions is unclear or weakly argued.	The theme does not capture anything meaningful or relevant to the research questions.
Clarity of Boundaries	The theme has clear and well-defined boundaries. It is distinct from other themes, with no overlap or confusion.	The theme has mostly clear boundaries, with minor overlaps or ambiguities that do not significantly detract from its distinctiveness.	The theme has somewhat unclear boundaries, with noticeable overlaps or ambiguities that weaken its distinctiveness.	The theme lacks clear boundaries. It overlaps significantly with other themes or is too broad/vague to be distinct.
Data Support and Evidence	Strongly supported by meaningful and sufficient data, with diverse yet coherent evidence.	Supported by sufficient data, but some data points could be more strongly aligned.	Partially supported by data, but with gaps or inconsistencies in alignment.	Lacks sufficient or meaningful data support; data are sparse, irrelevant, or misaligned
Theme Definition	The definition clearly outlines the theme's central organising concept, boundaries, and uniqueness.	The definition outlines the central concept and boundaries, but could be sharper.	The definition partially explains the central concept and boundaries but lacks depth or clarity.	The definition is missing, unclear, or fails to explain the theme's central concept, boundaries, or uniqueness.
Theme Name	The name is informative, concise, and catchy.	The theme name is clear and informative, but could be more concise or engaging.	The theme name is somewhat unclear or generic.	The theme name is vague or uninformative.
Contribution to Overall Analysis	The theme significantly and uniquely contributes to the overall analysis. It adds depth, insight, and clarity to the research questions and findings.	The theme contributes to the overall analysis, but its unique contribution could be more explicitly stated or developed.	The theme contributes partially to the overall analysis, but its role is unclear or underdeveloped.	The theme does not contribute to the overall analysis. It seems redundant, irrelevant, or disconnected from the research questions and findings.
Subthemes (if existent)	Subthemes are conceptually clear, non-overlapping, and each captures a distinct facet of the central organising concept. They enhance the narrative's meaning.	Subthemes are relevant and mostly well-aligned with the central concept. Minor overlap or lack of distinctness, but they still support theme clarity.	Subthemes are weakly connected to the central theme or to each other. They show some redundancy or confusion, weakening coherence.	Subthemes are misaligned, redundant, vague, or unnecessary. They add little value and may introduce confusion.

explained next:

7.3.3.1 Prompting Strategy

Creating and refining prompts was an iterative process that involved testing, assessing, implementing feedback, and retesting.

The final prompts were intentionally lengthy, as we included clear and consistent guidance throughout the analytic steps. However, the trade-off is that long prompts might reduce flexibility, increase computational cost, and make replication or adaptation more difficult. We included the full prompts in the online appendix [463]. In the paper, only the core parts are presented for brevity and overview. We elaborate in each prompt next:

Table 7.3: LLMs characteristics and dates of access per TA phase. The Phase is represented by "P"+ number.

Model	Settings (Temp, Max Tokens)	Phases and Dates of Access
ChatGPT o3-mini	Temp = default, Max tokens = 200k	P:2, date 21 of May, 25 P:3, date: 6 of July, 25 P:4&5, date 9 of july, 25
GPT-4o	Temp = default, Max tokens = 64,000	P:2, date 21 of March, 25 P:3, date: 6 of July, 25
Claude Sonnet 4	Temp = default, Max tokens = 64,000	P:2, date 21 March, 25 P:3, date: 23 of May, 25
Gemini 2.5 Pro	Temp = default, Max tokens = 1,048,576	P:3, date: 8 of July, 25 P:4&5, date date 9 of july, 25

Prompt 1- coding went through eight versions. We began with short instructions for the LLM on generating codes from interview responses in line with Braun Clarke's reflexive TA framework. Preliminary outputs had issues, including overly broad codes, hallucinated segments, and insufficiently descriptive labels. We addressed these by adding key elements and conducting internal dry runs with ChatGPT o3-mini, Claude Sonnet 4, and GPT-4o. The final prompt, version 8, generated the codes to be sent to the evaluators. This version requested the model to act as a qualitative researcher, identifying meaningful data segments in participants' responses and code labels. Each coded segment included the verbatim quote, a brief explanation of its relevance, and a note on any sensitive information. The prompt requested the LLM to write the specific interview number and line from which the code was taken. With this, we made sure that all codes were real and avoided hallucinations. We performed quality checks following this step by cross-verifying the excerpts with the original transcripts to confirm accuracy and consistency. The prompt also included a coding quality check using the "Excellent" column from the rubric in Table 7.5. Additionally, it specified detailed methodological guidance and structured JSON output for later

theme development. Since the prompt was long, 1485 tokens in total, we show an overview in Figure 7.2. See the online appendix [463] for the full prompt.

Prompt 2- Themes generation instructed the LLM to act as a qualitative software engineering researcher applying Braun and Clarke's TA. It requested the production of initial themes from the coded interview by Prompt 1-coding. The model received coded segments and must derive themes that directly address the study's RQ. It requested to group related codes into coherent, data-driven themes and sub-themes. It also asked to provide concise, meaningful theme names and write detailed definitions. The prompt included methodological constraints, a style example for depth and tone, and an internal self-check rubric based on the "Excellent" column from the rubric in Table 7.2. All of this presented in a JSON output structure.

Prompt 2 tested whether the LLM could generate themes across a full transcript. This step ensured that the model could handle larger amounts of qualitative data while maintaining alignment with Braun and Clarke's TA framework. We tested the prompt with ChatGPT o3-mini, GPT-4o and Claude Sonnet 4. The goal was to assess each model's ability to identify patterns across individual interviews before proceeding to the entire set. The prompt consists of 1280 tokens. The overview is shown in Figure 7.3, and the full prompt is provided in the online appendix [463].

Prompt 3- Refining themes was to generate themes for all 15 interviews, building on the sets of themes produced by Prompt 2. Whereas Prompt 2 generated themes per interview, Prompt 3 synthesised patterns across the full dataset. The aim was to produce a coherent set of candidate themes that answer the RQ.

This prompt asked the LLM to do the theme aggregation and refinement stage of TA. It used the initial themes from individual interviews and guided the model to merge themes. It also requested to produce a coherent, ranked set of overarching themes. The model assessed the significance of each theme based on its explanatory power, frequency, and diversity of supporting evidence. It also assigned ranks to high/medium/lower tiers, and records the source themes. For each theme and sub-theme, a detailed, human-quality definition was required. It described its central organising concept, boundaries, uniqueness, and contribution to answer the RQ.

We used the prompt in different LLM pipelines to produce 5 sets of candidate themes. Prompt 3 has 1319 tokens. See Figure 7.4 for an overview and the online appendix [463] for the full text).

7.3.3.2 Initial Coding (Phase 2)

We used the final version (V8) of prompt 1 to generate the codes for TA phase 2. The codes were then assessed and evaluated.

Evaluation. We evaluated this phase in two steps:

Eval 1a: codes HvsLLM: First, we compared human vs LLM. A total of 96 sets of codes were presented to evaluators randomly and blindly. Each set had between 1 and 4 codes. The sets were formed by human and LLMs codes from different interview segments. Evaluators received the interview transcript and the RQ to gain

Code 5

Interviewer:

How does your company promote diversity, equity and inclusion in the workplace?

Respondent:

They don't. If they do, it's invisible to me.

CODER A		CODER B	
Segment	**Code**	**Segment**	**Code**
They don't. if they do, it's invisible to me.	Perception on diversity, equity and inclusion	They don't. if they do, it's invisible to me.	Not perceiving DEI initiatives

Figure 7.5: Example of human- and LLM-generated codes presented in the blind evaluation. Evaluators chose between coder A and coder B and provided a brief justification for their decision. Here, the whole reply was coded, an uncommon case, shown for brevity and clarity.

a comprehensive understanding of the context. They also received an online survey, with each question addressing one set. Evaluators could see both sets (human and LLM) simultaneously. Figure 7.5 shows how evaluators saw each set and answered the survey. Then they assessed each set, selecting which codes better captured meaning and addressed the RQ. They also provided written justifications for their choices. Each set consisted of the interview question, the participant's response, and the corresponding code (human and LLM-generated).

Eval 1b: codes LLMvsRub: Second, we compared LLM vs rubric. After refining the prompt and generating additional codes, 24 sets of LLM-only codes were evaluated using a rubric (see Table 7.1). The rubric criteria reflect Braun and Clarke's qualities as essential for high-quality code generation. The four evaluators independently rated the codes and provided written feedback, yielding quantitative scores and qualitative insights.

7.3.3.3 Generating, Reviewing, Refining and Naming Themes (Phases 3-5)

For phase 3 (generation of initial themes), we used prompt 2 to generate the first themes per interview in ChatGPT o3-mini, GPT-4o, Claude 4 Sonnet models. This phase was not evaluated because it generated per-interview themes. These were intermediate analytic artefacts that required further refinement in Phases 4 and 5 before they could be meaningfully interpreted. Evaluating Phase 3 alone would not provide valid insight into analytical quality, as reflexive thematic analysis treats theme development as an iterative and cumulative process. Consequently, we focused our

Table 7.4: Pipelines tested in phases 4 and 5 to create the final themes. Pipeline 5 produced the best set of themes to answer the RQ

Pipeline	Phase 2	Phase 3	Phase 4 & 5
P1	ChatGPT o3-mini	Gemini 2.5 Pro	Gemini 2.5 Pro
P2	ChatGPT o3-mini	ChatGPT o3-mini	Gemini 2.5 Pro
P3	ChatGPT o3-mini	ChatGPT o3-mini	ChatGPT o3-mini
P4	ChatGPT o3-mini	Claude Sonnet 4	ChatGPT o3-mini
P5	ChatGPT o3-mini	Claude Sonnet 4	Gemini 2.5 Pro

evaluation efforts on the more stable outputs from Phases 2 and 4 & 5, which better reflect interpretive rigour and final analytic quality.

Then, we used prompt 3 to cover phases 4 and 5 (refined, defined and named the themes). We used five different LLM pipelines (see Methodology, Figure 7.1) across all 15 interviews. The pipelines combined ChatGPT o3-mini, GPT-4o, Claude 4 Sonnet, and Gemini 2.5 Pro.

Five sets of themes were generated and assessed for completeness, code quality, structural coherence, and the presence of subthemes. Pipelines that omitted interviews or lacked a clear structure were excluded. The best-performing pipeline presented all the assessment criteria. Table 7.4 shows all the pipelines we used to generate theme sets. Then we selected the most promising set, produced by P5, to be scored by the evaluators. Nine distinct themes formed the resulting set of themes.

Eval 2a: Themes vs Rub: The final set of themes was independently evaluated by three experts. They used the rubric in Table 7.2 and provided free-text comments for overall feedback.

7.3.4 Data analysis

For the quantitative results from the rubrics, we summed the evaluators' scores for each criterion and calculated averages. We conducted a content analysis for the qualitative feedback (free-text justifications and comments), grouping statements into the rubric in Table 7.2 criteria.

7.3.5 Evaluators Team

One evaluator is a full professor of information systems and digital technologies. They research on human cognition, behaviour, and social interactions in software engineering. This evaluator brought extensive experience in QDA and research of human aspects of digitalisation and software development.

Another evaluator is a social scientist, an associate professor with a background in media and communications and marketing. They specialise in the social impacts of digital media and emerging technologies. Their research includes intimacy, online privacy, AI, humanoid robots, influencer marketing, and digital nomadism. They have extensive experience in QDA, specifically in thematic analysis.

One more evaluator holds a PhD in information systems and has over a decade of research and teaching experience. Their work spans interdisciplinary projects in digital technologies, healthcare, and mental health. They have expertise in design methods, technology adoption across cultural contexts, and connected health. They have experience in mixed methods and have performed thematic analysis extensively.

The last evaluator is a senior lecturer in computing science with expertise in software engineering, human–machine interaction, and sustainability. Their research addresses education, health, and well-being domains, where digitalisation reshapes traditional services and user expectations.

The evaluators represented a complementary mix of expertise in SE, qualitative research methods, human–technology interaction, healthcare, and socio-cultural studies. They were chosen to provide a broad yet rigorous perspective. They ensured that evaluations of LLM outputs considered methodological quality alongside human, organisational, and societal dimensions.

7.3.6 Ethical Considerations

We obtained informed consent from all interviewees to use their data. Transcripts were anonymised before analysis. We used the LLMs following their respective terms of service. Evaluators were blinded to the source of coded outputs to minimise bias in comparative assessment.

7.4 Results

This section presents the results of our evaluations of LLMs' performance in conducting Phases 2 through 5 of TA, following Braun and Clarke's six-phase framework. The section is organised in a chronological order, hence the code evaluations are presented first and then the theme evaluations.

7.4.1 Evaluation Phase 2: Human vs LLM (1a)

As explained in Eval 17.3.3.2, evaluators had to choose blindly between human and LLM codes. They preferred LLM-generated codes more often, 58 times out of 95 (since Evaluator 1 did not rate the final set of codes). The total rate was 61%. Meanwhile, Human-generated codes were selected 37 times (39%). Figure 7.6 summarises their preferences across all segments. The choice for LLM codes was relatively consistent across evaluators. However, there was some variation; for example, Evaluator 3 selected human codes more often than others. Complete agreement among the four evaluators was shown only for LLM-generated codes five times.

The evaluation's **qualitative** part is summarised in Figure 7.7 for human codes and Figure 7.8 for LLM-generated codes. The full table with all the comments is available in the online appendix [463].

Despite the lower overall selection rate, evaluators identified several strengths in **human-generated codes**. They noted that the codes demonstrated thematic depth. The codes captured behaviours, interpersonal dynamics, motivations, and outcomes without excessive fragmentation. Hence, codes offered breadth and clearer

Set	Eval 1	Eval 2	Eval 3	Eval 4	MV
1	LLM	LLM	LLM	LLM	**LLM**
2	LLM	H	H	LLM	
3	LLM	LLM	H	H	
4	LLM	LLM	LLM	LLM	**LLM**
5	LLM	LLM	H	LLM	**LLM**
6	LLM	LLM	LLM	H	**LLM**
7	H	LLM	H	H	**H**
8	LLM	LLM	LLM	LLM	**LLM**
9	LLM	LLM	LLM	H	**LLM**
10	LLM	H	H	H	**H**
11	LLM	H	H	LLM	
12	LLM	LLM	LLM	LLM	**LLM**
13	H	H	H	LLM	**H**
14	LLM	H	H	H	**H**
15	H	LLM	LLM	H	
16	LLM	LLM	H	H	
17	LLM	LLM	LLM	H	**LLM**
18	LLM	LLM	LLM	LLM	**LLM**
19	LLM	LLM	H	LLM	**LLM**
20	LLM	H	LLM	LLM	**LLM**
21	LLM	H	H	H	**H**
22	LLM	LLM	H	LLM	**LLM**
23	LLM	H	H	LLM	
24	X	H	H	H	**H**
LLM (n)	20	15	10	13	
H (n)	3	9	14	11	

Figure 7.6: Evaluator preferences for human and LLM-generated codes across 24 interview segments. Each column corresponds to one of four evaluators (Eval 1–4), and each row represents one set of codes. Cells indicate whether the evaluator chose the human (H) or LLM (LLM) codes. Evaluator 1 declined to rate the final segment, judging that none of the codes aligned with the RQ. The majority voting (MV) column indicates majority preferences. H (n) and LLM (n) show how often each evaluator chose the H or LLM code.

connections within the data. Evaluators also noted that the codes demonstrated clarity and analytical usefulness. Labels were concise and insightful, going beyond mere description to contextualise data and extract the core meaning of interviewee responses. Experts also mentioned codes' practical application, being easy to categorise, consistent in structure, and actionable in use. Similarly, they found the codes relevant and focused, since they aligned closely with the RQ and directly identified influences on well-being. Finally, evaluators mentioned that codes showed consistency across the sets.

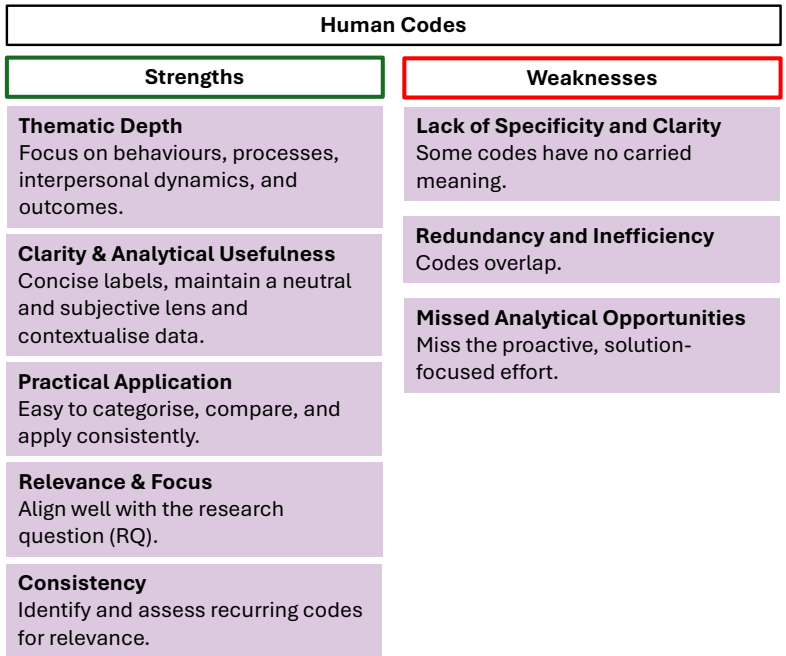


Figure 7.7: Summary of strengths and weaknesses identified by evaluators for human-generated codes, based on the Phase 2 (1a) evaluation.

At the same time, evaluators identified several areas of opportunity. A main concern was the lack of specificity and clarity. Some codes were deemed overly broad, generic, or descriptive, lacking nuance and depth, and occasionally failing to reflect participants' expressions accurately. Related to this, evaluators pointed to redundancy and inefficiency, with overlapping codes that reduced analytical sharpness. In addition, they noted that some codes missed analytical opportunities. This was particularly in capturing proactive or solution-oriented perspectives and clarifying how specific aspects of participant perceptions influenced outcomes.

Regarding **LLM-generated codes**, Figure 7.8 presents evaluators' qualitative feedback. They observed substantial analytical depth and thematic insight. The codes extended beyond surface-level descriptions to reveal underlying factors, tensions, and contradictions. LLM outputs were noted for their ability to understand metaphors, identify inferred concepts, and link physical and mental aspects of well-being. Sometimes codes situated participant experiences within broader theoretical frameworks. This interpretive characteristic gave the codes a sense of "telling a story" rather than offering only descriptive labels.

Another strength was clarity and specificity, reflected in the rubric scores for Clarity of meaning WA = 2.94 (Figure 7.9). The codes frequently provided precise descriptions of important concepts, used concise action-oriented phrasing, and aligned well with participants' intentions. The codes also displayed good structure and organisation. Codes broke experiences into meaningful parts, showing progression and

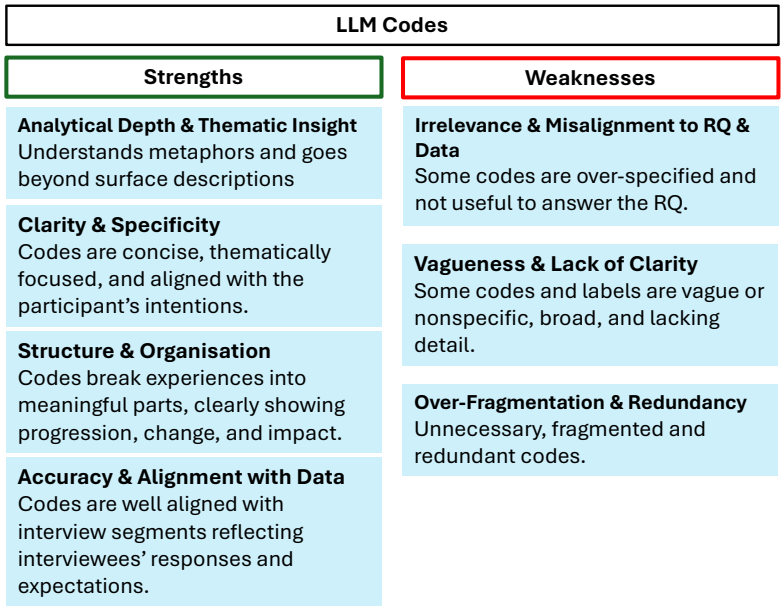


Figure 7.8: Summary of strengths and areas of opportunity identified by evaluators for LLM-generated codes, based on the Phase 2 (1a) evaluation.

change, and supporting comparison across diverse experiences. Finally, evaluators mentioned accuracy and alignment with the data. Many codes captured brief but significant excerpts and reflected participants' responses faithfully.

Evaluators also identified areas of opportunity. One recurring issue concerned relevance and alignment to the RQ and data. This improved in the second evaluation (WA = 3.04). Some codes were judged not directly helpful in answering the research question. Others were overly specific, focused too much on individual experiences rather than team or organisational perspectives, or misrepresented participant accounts. Evaluators also noted vagueness and lack of clarity. Some labels were broad, insufficiently action-oriented, or ambiguous in conveying causality and influence. Finally, problems of over-fragmentation and redundancy were reported. Some codes overlapped unnecessarily, while others fragmented experiences, reducing coherence.

Key finding: Our results showed that, for Phase 1 (generating initial codes), LLMs can produce codes that are competitive with, and sometimes preferred over, codes produced by experienced human researchers.

In this evaluation, evaluators chose LLM codes 61% of the time. The reason for this high percentage can be due to the surface-level readability and polish of LLM outputs. Evaluators consistently described LLM codes as concise, specific, and well-formulated. This made codes appear clearer and easier to apply, even when they sometimes lacked deeper alignment with participants' intent. By contrast, while strongly relevant to the RQ and contextually sensitive, human codes were often longer, less standardised, and occasionally redundant. These characteristics can make human codes appear less

precise in side-by-side comparisons, even if they carry greater reflexive or interpretive weight.

7.4.2 Evaluation Phase 2: LLM Codes Rubric-Based Evaluation (1b)

Figure 7.9 shows the ratings distribution across all evaluated code sets for the second part of the phase 2 evaluation (see section 7.3.3.3). Overall, most criteria received a "Good" rating. The better-rated dimensions were "Explanation of Interview Segment Selection" (Excellent = 42.7, WA = 3.21) and "Relevance to RQ" (Excellent = 27.1, WA = 3.04). This indicated the LLM's capabilities to justify segment selection and alignment with the RQ and the study's analytical goals. The lowest rated criterion was "Balance between Latent and Semantic Codes" (Poor = 13.5, WA = 2.59). This showed that LLMs may have difficulties interpreting segments, which is expected in human-led qualitative analysis. Similarly, "Potential for Theme Development" (WA = 2.91) was rated moderately, implying room for improvement in analytical precision and thematic extrapolation.

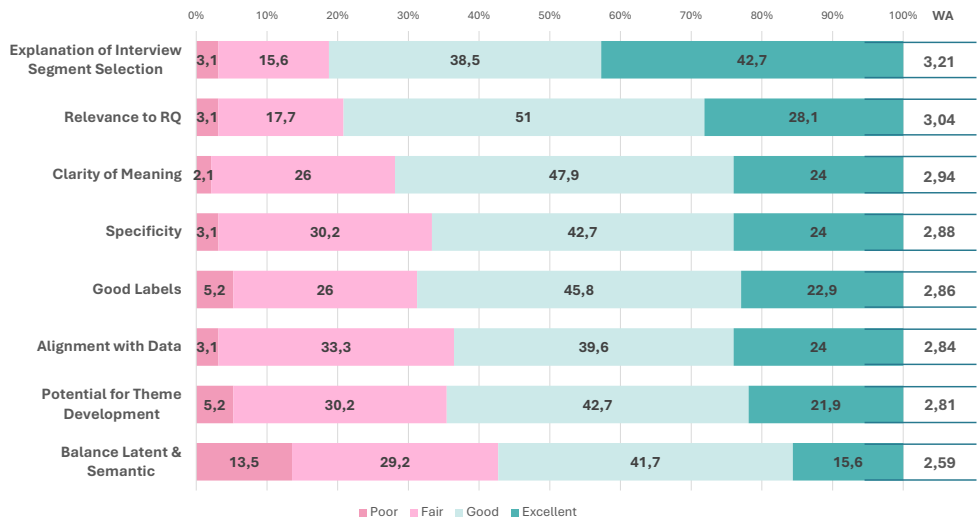


Figure 7.9: Rubric-based evaluation of LLM-generated codes across eight criteria (see Table 7.1). Based on evaluations from four raters, bars indicate the percentage of ratings for each quality level (Poor, Fair, Good, Excellent). The weighted average (WA) per criterion is shown on the right.

Similar to the previous evaluation, we asked evaluators to give feedback by explaining their assessment. Table 7.5 presents the analysis of evaluators' comments across eight criteria. The full table is in the online appendix [463]. In this evaluation, experts focused mainly on the weak aspects of the codes; hence, the table is more populated on the negative side.

The two criteria with balanced comments were clarity of meaning and explanation of the interview segment selection. This finding aligned well with the quantitative evaluation in Figure 7.9 (WA= 3.21). Regarding clarity of meaning, evaluators stressed the LLMs' ability to produce clear and coherent interpretations. LLMs performed well on this even when working with fragmented or ambiguous participant responses. However, they also commented on missed nuances, vague terminology, and occasional mismatches between the code's emphasis and the segment's actual intent.

More positively, the 'explanation of interview segment selection' was consistently commented to be clear, coherent, and persuasive framing. The positive comments aligned well with the positive results from the rubric scores. At the same time, evaluators identified that strong justifications sometimes masked weak codes or introduced assumptions not fully grounded in the data.

Key finding: Codes were good at clarity and relevance but weaker at nuance and latent interpretation.

Our results indicate that LLMs can produce codes that meet several foundational standards of TA, particularly in terms of clarity and relevance. However, more abstract or interpretive dimensions, such as latent insight to create codes and interpretive nuance, seem more challenging to achieve.

Table 7.5: Qualitative feedback from evaluators on the performance of LLM-generated codes, grouped by rubric criterion. Positive comments start with + and negatives ones with -.

Criteria	Comments
Clarity of Meaning	+ Codes align closely with the expected interpretation + LLM interpreted a fragmented and incoherent response surprisingly well - Missed some details and do not reflect suggestions - Inappropriate or unclear terminology
Relevance to Research Question	+ Some codes are relevant to RQ - Some codes are not clearly linked to the RQ - Codes miss expressive or subjective content and nuanced segments
Balance of Latent and Semantic	+ Overall performance is fair in balancing latent and semantic content + Most codes stay too close to the surface meaning of the text - There is untapped potential for deeper interpretation of participant meaning

Criteria	Comments
Specificity	+ Descriptive codes match a descriptive response - Certain codes are too broad, redundant or unfocused and miss nuances - Codes are mostly descriptive and lack depth and detail - Codes are either too granular or lack granularity; the middle point is not there
Potential for Theme Development	- Missed opportunity to explore causes or implications - Codes may overlap in meaning; merging could improve thematic clarity
Alignment with Data	- Codes miss contextual references and key details from interviewees - Code and explanation ignore the question context
Good Labels	+ Code label summarises content effectively + Avoids direct quotation from the text - Code label may narrow the meaning and reduce accuracy - Unclear if code reflects an observed event, an experience or a suggestion - Labels omit key details relevant for further analysis and are ambiguous or misleading
Explanation of Interview Segment Selection	+ Good explanation, is persuasive and compelling + Explanation of code selection was strong and well-articulated - Risk of overestimating code quality due to strong explanation - Explanation includes assumptions not grounded in data, over-interprets participants' words and does not consider the full context

7.4.3 Evaluation Phase 3-5 Generating, Refining and Naming Themes (2a)

Figure 7.10 summarises the themes' evaluation results. The figure shows a decreasing tendency in the evaluations. It starts from a well-evaluated theme, "The Team as a Protective Sanctuary", to a theme with low scores, "Supportive Organisational Infrastructure and Leadership". The scores aligned well with the ranks and tiers proposed by the LLM. The first five themes were placed in the High tier, showing consistently strong performance across dimensions.

Three themes were placed in the Medium tier. These typically showed solid organising concepts but were weak in clarity of boundaries. The lower score was in data support and evidence, since some codes that formed this theme were vague.

The last theme, "Supportive Organisational Infrastructure and Leadership", fell into the Lower tier due to limited analytical sharpness and inconsistent naming. It showed weak differentiation from other themes, unclear subtheme use, and insufficient clarity in tone and contribution to the overarching analysis.

These findings indicate that the selected LLM pipeline was capable of creating

themes that meet many of the criteria of TA. While not all themes achieved equal strength, the best-performing ones have structural clarity, interpretive coherence, and alignment with participants' meaning.

On the weak side, experts commented on the apparent over-fragmentation of the theme structure. For example, Theme 9 could be merged with Theme 3 as a subtheme. Although Theme 9's topic is important, it overlaps with the central organising concept of Theme 3. The experts proposed creating a subtheme titled "Organisational Infrastructure and Company Support" and including it under Theme 3.

Besides that suggestion of restructuring, the themes grouping seems complete in topics that help to answer the RQ.

Key finding: LLM themes were strong overall but sometimes fragmented; experts suggested merging overlaps.

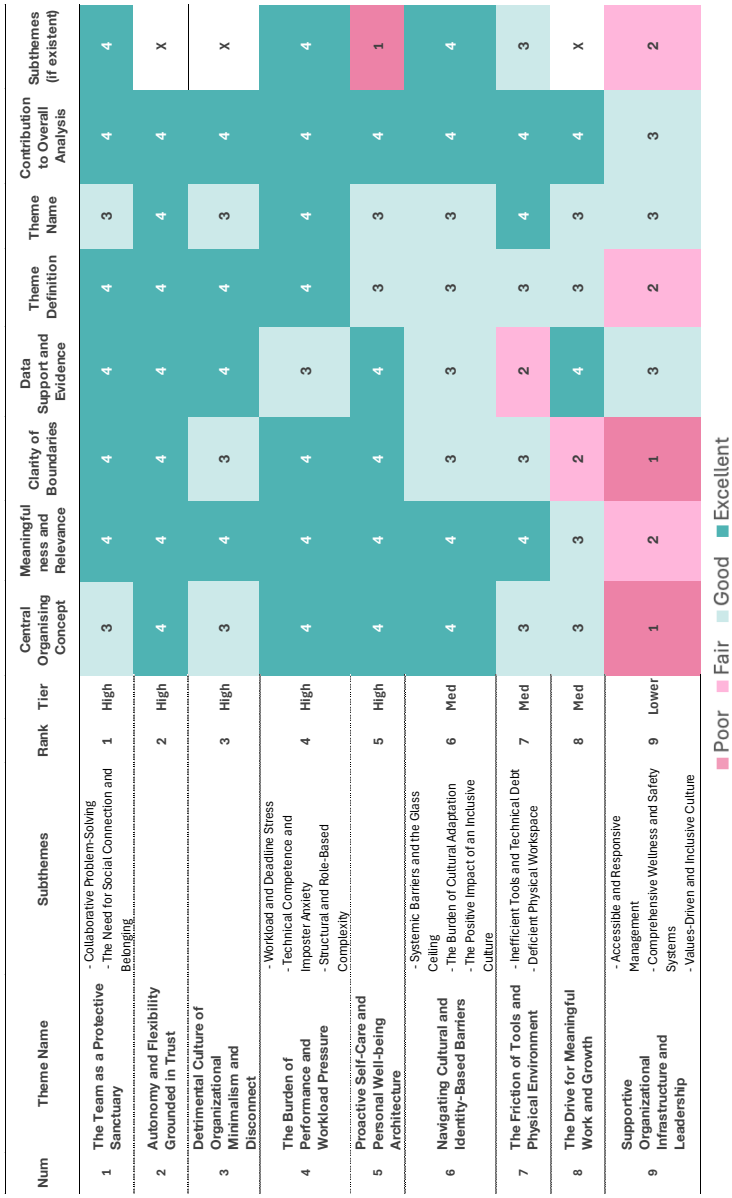
7.5 Discussion

Following the current discussions around the inclusion of AI in QDA, we adapted Braun and Clarke's TA framework to be performed partially by LLMs. Our study addresses the call to design and tailor prompts for QDA [464]. In this section, we discuss the implications, limitations, benefits, methodological aspects, and recommendations to consider when using LLMs in QDA.

7.5.1 LLMs as Analytical Assistants in SE Research

Based on our results, we conclude that LLMs can perform several phases of TA. In particular, they are effective in initial coding and theme generation. Evaluators rated the LLMs' outcome quality highly, specifically the codes, which were preferred over human-generated ones. These findings align with previous positive outcomes of conducting QDA using AI [457, 460, 462, 465]. This thereby points to LLMs as viable analytical assistants in qualitative research. In contexts with large volumes of qualitative data or with a single researcher, LLMs can help reduce time and manual workload. They assist in these scenarios by offering codes and candidate themes that researchers can refine. However, this help comes with the need for caution. LLMs lack a critical and analytical perspective, which may compromise the method's rigour. Experts' reflexive supervision is necessary during the LLM's pre-coding and clustering. Similarly, during the interpretation phases, the model can generate convincing outputs that are not always grounded in the data. Because qualitative data analysis involves reflexive, complex, and continuous meaning-making [466], automating the entire process is not recommended. LLMs can surface patterns, candidate framings, and alternative readings, but they do not make meaning in the epistemological sense; that remains the role of reflexive human interpretation. Accordingly, LLM outputs should be positioned as scaffolds for researcher sense-making (e.g., prompts to compare, contest, or refine interpretations), not as substitutes for interpretive judgement or context stewardship.

The following subsections elaborate more on the role and potential collaboration



Num	Theme Name	Subthemes	Rank	Tier	Central Organising Concept	Meaningfulness and Relevance	Clarity of Boundaries	Data Support and Evidence	Theme Definition	Theme Name	Contribution to Overall Analysis	Subthemes (if existent)
1	The Team as a Protective Sanctuary	- Collaborative Problem-Solving - The Need for Social Connection and Belonging	1	High	3	4	4	4	4	3	4	4
2	Autonomy and Flexibility Grounded in Trust		2	High	4	4	4	4	4	4	4	X
3	Detrimental Culture of Organizational Minimalism and Disconnect		3	High	3	4	3	4	4	3	4	X
4	The Burden of Performance and Workload Pressure	- Workload and Deadline Stress - Technical Competence and Imposter Anxiety - Structural and Role-Based Complexity	4	High	4	4	4	3	4	4	4	4
5	Proactive Self-Care and Personal Well-being Architecture		5	High	4	4	4	4	3	3	4	1
6	Navigating Cultural and Identity-Based Barriers	- Systemic Barriers and the Glass Ceiling - The Burden of Cultural Adaptation - The Positive Impact of an Inclusive Culture	6	Med	4	4	3	3	3	3	4	4
7	The Friction of Tools and Physical Environment	- Inefficient Tools and Technical Debt - Deficient Physical Workspace	7	Med	3	4	3	2	3	4	4	3
8	The Drive for Meaningful Work and Growth		8	Med	3	3	2	4	3	3	4	X
9	Supportive Organizational Infrastructure and Leadership	- Accessible and Responsive Management - Comprehensive Wellness and Safety Systems - Values-Driven and Inclusive Culture	9	Lower	1	2	1	3	2	3	3	2

Figure 7.10: Rubric-based evaluation of nine themes produced by the best-performing LLM pipeline. The first five columns ("Num" to "Tier") are LLM's outputs. The remaining columns show human evaluations based on the rubric in Table 7.2. Three evaluators rated each theme on a scale of 1 (poor) to 4 (excellent). The values shown are median scores across raters.

humans and LLMs can have when doing QDA together. Figure 7.11 presents LLM tasks as an analyst assistant.

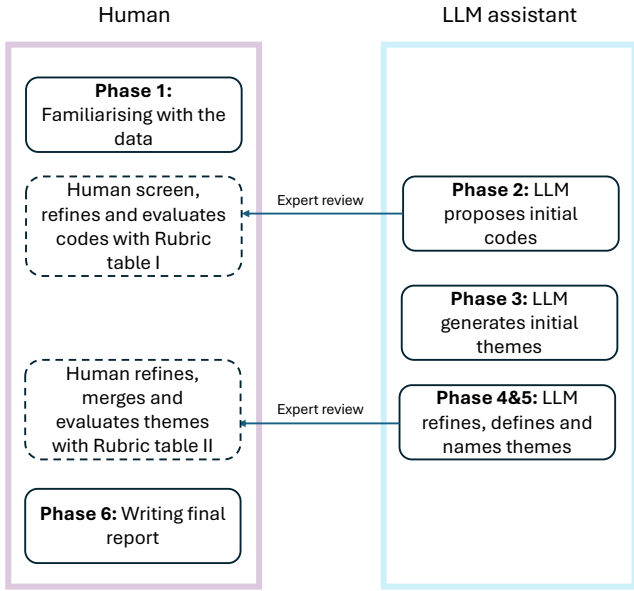


Figure 7.11: Proposal of implementation of LLM in TA. LLM contributes to TA Phases 2–5 as an assistant; the human leads Phases 1 and 6 and gates progression using rubric-based evaluations. Dashed boxes indicate areas that require human evaluation and refinement.

Based on our discussion, we present in Figure 7.12 the benefits, risks and guidelines when integrating LLMs in QDA. It aims to support researchers in designing their studies more transparently and rigorously. It can be used as a reference point for evaluating and reporting LLM involvement in future empirical SE work.

7.5.1.1 Human Oversight Remains Essential

Our findings reaffirm that while LLMs can support qualitative analysis, they cannot replace human researchers. In line with prior work [442, 460, 467], it is recommended to assign LLMs the role of assistant or collaborator instead of a substitute. One of the reasons is that we found that LLMs struggled to balance semantic and latent coding [467]. This matters because latent coding in SE research often requires connecting technical knowledge with implicit beliefs, values, or organisational dynamics. Such interpretations rely on human judgment and reflexivity [75]. Human oversight is essential for ensuring nuanced and valid interpretations [460].

Furthermore, we found that LLMs risk oversimplifying or misinterpreting without feedback and supervision, consistent with [467]. Essential aspects such as tone, intention, and power dynamics remain accessible only through human engagement in the research process. As we advocate for a collaboration with LLMs in QDA, we also emphasise that researchers must remain knowledgeable about their dataset. They have to be able to guide the LLM, and vigilant about bias and alignment with research goals [468]. Transparency about the involvement of LLMs is also a key responsibility.

Dimension	Benefits	Threats	Guidelines for Future Auto-QA in SE Research
Codes quality	Produces clear, relevant, structured and well-justified codes; often preferred over human ones; aligns with RQ when prompted.	Risk of vague, redundant, or fragmented codes; weak in latent/interpretive meaning. Inclusion of biases in codes. Hallucination of codes or segments.	Use LLMs for *initial coding*, but require human review for interpretive depth and alignment with context. Apply rubric in Table I to refine codes.
Themes' structure	Can cluster codes into meaningful candidate themes; some strong organising concepts.	Theme boundaries sometimes unclear; limited handling of subthemes; over-fragmented theme structure.	Apply rubric in Table II to check theme coherence; merge or refine with human expertise.
Analytical depth	Detects metaphors, contradictions, inferred links; adds interpretive richness. Possibility of conversation engagement to find new interpretations and angles.	Misses cultural, contextual, and theoretical nuance; outputs may seem convincing but shallow.	Researchers must add context, reflexivity and theoretical framing. Treat LLM outputs as starting points or proposals from a junior colleague.
Efficiency and scalability	Faster coding and theme generation, which reduces manual workload. Handles large datasets or single researcher studies.	Over-reliance may reduce human immersion in data, critical role, and reflexive thinking.	Use LLMs to *augment*, not replace, human immersion; balance efficiency with engagement and refine output with human supervision.
Transparency and transferability	Full disclosure of prompts. Explanations of coding decisions to enhance transparency.	Without detailed reporting, reproducibility, rigour and validity are undermined.	Disclose LLM integration, prompts, models, and evaluation procedures; disclose human intervention.
Traceability	Tailored prompt to trace every code to specific interview segments to avoid hallucinations.	Lack of traceability can miss possible hallucinations.	Audit and document code traceability. Record how many codes were verified against the original transcripts and report the percentage of confirmed links.
Ethics	Flags sensitive and distressing content; supports anonymisation and ethical safeguards	Risks if privacy, data leak, consent, or sensitive data are mishandled. Inclusion of biases in data.	Always keep humans responsible for sensitive content handling before loading the transcripts to the LLM, anonymising data, obtaining consent, and revising all flagged content.
Methodological rigour	Follows a specific framework and guidelines and considers the research question(s) to guide the analysis. It can serve as a quality checker by providing alternative readings of the data.	Without grounding in established frameworks and clear RQs, outputs risk being superficial, irreproducible, and misaligned with the research aims. Minimise researcher involvement can risk methodological rigour.	Add to the prompt the specific framework and parameters to follow and the research question(s). Ensure alignment of codes, completeness of the themes, iteration and reflexivity.

Figure 7.12: Reflection-based summary of the benefits, risks, and methodological guidelines for integrating LLMs in thematic analysis within SE research.

7.5.2 Strengths and Limitations of LLM Outputs in Engineering Contexts

The strengths of LLM-generated codes (see Figure 7.8) include Analytical Depth & Thematic Insight, and Structure & Organisation, which are important characteristics when doing QDA. Evaluators even commented on the capacity to understand metaphors and identify inferred concepts. This is notable given the topic in our dataset.

Moreover, the LLM could clearly justify codes creation in phase 1. During the evaluation, this criterion received the highest score in "Excellent" (see Figure 7.9, Explanation of Interview Segment Selection WA = 3.21). It could also explain why a particular quote or excerpt was used in a theme. This results from tailoring the prompt with a theoretical position to simulate "choosing" segments based on their research purpose. However, it is important to remember that this is only a mimic, not an alignment with epistemological expectations.

Relevance to Research Question WA = 3.04 was also among the highest scores. The LLM could, in general, align codes with the RQ and avoid tangential interpretations. We included the RQ in the prompt along with specific guidelines of what type of codes we expected. This guided the LLM in prioritising analytical decisions for the study's goal. With a non-tailored prompt, the LLM might fail to choose relevant codes and output surface-level or not well-aligned segments. It is important to note that evaluators commented that some codes did not appear to be fully aligned with the RQ. This was in phase 2, where the first coding round was done. Braun and Clarke clarify that codes and even themes are not final until the end of the analysis. Having tangential codes that could potentially be added later to the analysis does not translate into a problem. In this case, it reinforces the need for researchers to review, refine and guide the iterative stages of analysis.

However, there are limitations regarding Vagueness and lack of Clarity and Redundancy. Especially in Phases 4–5, themes occasionally lacked clear boundaries. LLMs did not create subthemes when there was room for the,m and appeared too fragmented in their theme structure. These shortcomings can reduce the coherence of the analysis and fail to capture nuances in the data. As result, this can compromise the narrative to answer the research questions.

Furthermore, the lowest rubric score was Balance of Latent and Semantic, WA = 2.59. It is essential to have both types of codes in thematic analysis to create a wholesome analysis. LLM did a good job at the semantic level. However, creating latent codes requires theoretical reading, ideological critique, and cultural and contextual sensitivity, which the LLM fails to mimic. This result shows the limitations and risks of relying only on LLMs to create codes. It once more stresses the importance of researchers in identifying deeper meanings, questioning assumptions, and situating findings within broader theoretical and socio-cultural frameworks.

One more factor to consider is the clarity and formulation quality in LLM's outputs. LLM outputs are often linguistically polished and consistent, which can influence judgments of quality even when interpretive depth is limited. Thus, evaluator preference does not necessarily equate to epistemic adequacy. This situation might be more critical when qualitative data comes from a context where meaning is expressed through colloquial, non-standard, or culturally embedded language.

7.5.3 Strategies to Ensure Methodological Rigour

We implemented the following strategies based on Christou's [468] suggestion of the need to ensure accuracy and credibility in all AI-generated content by cross-referencing it.

a) The dataset we used and the prompt had cues that helped us **trace** back the source data (each segment for each code per interview). By instrumenting and cross-referencing the prompt, we **mitigated hallucinations, lack of transparency, inconsistency and increased trust and auditability**. To strengthen this even more, we required the LLM to work segment by segment and to generate codes anchored in specific data excerpts to increase data fidelity.

b) We also prompted the LLMs to write an explanation of the selected codes and the reason each code helped answer the RQ. With this, we aimed to improve **transparency** in each step of the process. We also enable more precise alignment between the data, the generated codes, and the main research question.

c) To ensure **methodological accuracy**, we instructed the LLM specifically with the Braun and Clarke Thematic Analysis approach. We included the definition and examples of codes and specifications of the type of coding, inductive, in this case. To complement this, we tailored our prompt to have a specific research purpose by giving it a research question.

d) Additionally, we requested the LLM to flag sensitive content in the interview segments to ensure the **ethical handling** of potentially distressing material and support AI's responsible use in qualitative analysis.

Combining these strategies, we aimed to provide a structured and transparent method to use LLMs in TA. We included strategies to ensure methodological rigour and to address AI's common challenges [442]. These relate to the credibility, ethical integrity, and trustworthiness of the analysis process. Process rigour is maintained by positioning the LLM as an analytical assistant while the researcher ensures reflexivity, iterative engagement, and interpretive decision-making.

7.5.4 Implications for Empirical SE Methods

As the integration of LLMs in qualitative research, particularly in QDA, appears inevitable [465], it is essential to define and assess their role and limitations. We must also examine the methodological implications of their use in SE research and beyond. We propose to have them as analytical assistants in qualitative research. This implies rethinking current strategies for ensuring quality and trustworthiness in qualitative research. Such quality features are credibility, transferability, dependability and confirmability [468, 469].

- Credibility: Researchers need to familiarise themselves with their data to give feedback to the LLM throughout the QDA process. Regarding reflexivity,

it must expand to critically assess the AI's biases, limitations, and how its outputs influence interpretations. Researchers need to reflect not only on their subjectivity but also on how the AI shapes the analytic process and outcomes.

- Transferability: Clear reporting standards should now include details about the research context, model specification, and human involvement. They should also describe the prompt structure and how the AI output was integrated into the analysis.
- Dependability: The methodological documentation should explicitly detail the AI components, preprocessing steps, and how researchers validated or modified AI outputs. Audit trails must register AI interactions, results and decision-making by the researcher based on them. A change tracker on codes or themes changes can aid with this marker.
- Confirmability: For peer debriefing, discussions about the AI's role and input need to be included. Building on the previous point, the influence of AI in data interpretation needs to be questioned. Member-checking practices also need adaptation. If the study's participants are asked to review AI-generated summaries or themes, researchers must clearly explain the role of AI in the process. Additionally, it is necessary to consider participants' views on AI's interpretative role. Finally, researchers' reflexive journals should include insights on AI-related challenges and decisions. They should also acknowledge how AI's presence shaped the researcher's thought process and interpretations.

Integrating LLMs as analytical assistants in SE can help with specific field challenges. Such challenges are managing massive, technically rich, and inherently socio-technical data. Since SE qualitative data often combines technical artefacts with human-centric sources. LLMs can help by triangulating insights more efficiently and effectively with large amounts of data. In the requirements engineering area, for example, LLMs can easily and consistently trace data. This will allow human researchers to focus on higher-level interpretative work. Additionally, they can maintain consistency across long or multi-researcher projects, aid in reducing mental exhaustion, and assist the researcher in identifying biases. Prompts can be tailored to try different QDA, which gives the researcher a perspective on data analysis decisions.

Finally, it is equally important to address ethical and privacy aspects. In this study, we handled all data under informed consent and anonymisation protocols. LLMs processed no personal identifiers. All models were accessed under institutional terms of service, ensuring compliance with privacy standards. We, researchers, were responsible for ethical oversight and for verifying sensitive segments flagged by the models. However, more researcher is needed to tailor formal frameworks for responsible AI-assisted qualitative analysis. This is especially important in cases where qualitative data contains sensitive or personal information.

7.5.5 Threats to Validity

We took several measures to strengthen the validity of our findings across the four standard categories: internal, external, construct, and conclusion validity. We elaborate on them in the following paragraphs.

Internal validity: Our dataset included a ground truth of pre-coded interviews by two experienced researchers. Additionally, we employed blinded comparative evaluation during Phase 2. This ensured that evaluators did not know the code generators. As a result, it helped isolate the effect of code quality from potential biases related to authorship. We also iteratively refined prompts and evaluation procedures to reduce confounding variables related to prompt phrasing or interpretation. Having expert evaluators with prior qualitative research experience increased consistency in applying evaluation criteria, reducing potential noise in the assessment process.

Construct validity: Our prompts were lengthy, as we included a significant amount of information and quality checks. This might make them complex and harder to use. To mitigate this, we structured the prompts clearly, provided step-by-step instructions, and tested them iteratively to ensure usability and effectiveness.

Rubrics were grounded in Braun and Clarke's Reflexive TA framework, ensuring alignment with TA rigour standards. Additionally, we prompted the LLMs to explain their coding decisions and how they related to the RQ to ensure that outputs reflected more than superficial content features.

Conclusion validity: We triangulated evaluators' feedback by having experts from social sciences and SE. Furthermore, we collected quantitative and qualitative data from their evaluations to make more robust inferences about LLM performance. This allowed us to cross-check results across different evaluative lenses. Furthermore, we evaluated five complete LLM pipelines, rather than relying on isolated examples. We selected the best-performing one through a systematic comparison, which strengthened our conclusions.

External validity: The prompt was tailored in a way that can be used with any qualitative dataset. We acknowledge that LLM performance may vary across topics and datasets, and future work should explore broader generalisability. Additionally, we only tested code quality within a specific topic and a homogeneous population. It is necessary to account for differences in colloquial language, culturally embedded experiences, and diverse types of populations.

7.6 Conclusion

This study provided one of the first systematic, rubric-based evaluations of LLMs as analytical assistants in qualitative SE research. Our results showed that LLMs can produce analytically applicable and often well-structured codes and themes. Human evaluators preferred LLM-generated codes in most cases (61%), confirming their potential to augment human interpretation.

In this study, we offer a documented, reproducible framework including:

- Complete prompts for use in LLMs.
- Tailored rubrics based on Braun and Clarke's TA to evaluate the quality of codes and themes.
- Result-based guidelines to integrate LLMs into QDA.

We particularly stress the methodological grounding of our study to ensure rigour and trustworthiness. We conclude that LLMs can assist in QDA in SE when used

as collaborators. This is effective as long as they are embedded within well-defined methodological and ethical boundaries. To examine the process integrity of using LLMs as analytical assistants in thematic analysis, future work should conduct parallel analyses on new data, comparing a fully human analysis with an LLM-assisted one.

7.7 Authors' Contributions

C.M.M. and R.F. conceived and designed the study. R.F. implemented and ran the LLM prompts, provided feedback, and contributed to shaping the study design. C.M.M. performed the data analysis, integrated the evaluation results, and prepared the manuscript drafts. Both authors discussed the findings and contributed to the interpretation of the results.

The evaluators, C.M.M., S.P., S.O., and D.G., conducted the blind evaluations of the codes and themes and revised the final paper version, but did not participate in the study design, or prompt development.

Bibliography

[1] A. Brand, L. Allen, M. Altman, M. Hlava, and J. Scott, "Beyond authorship: Attribution, contribution, collaboration, and credit." *Learned Publishing*, vol. 28, no. 2, 2015.

[2] C. Amrit, M. Daneva, and D. Damian, "Human factors in software development: On its underlying theories and the value of learning from related disciplines. a guest editorial introduction to the special issue," pp. 1537–1542, 2014.

[3] L. M. Restrepo-Tamayo and G. P. Gasca-Hurtado, "Human aspects in software development: A systematic mapping study," in *International Conference on Collaboration Technologies and Social Computing*. Springer, 2022, pp. 1–22.

[4] O. Hazzan and I. Hadar, "Why and how can human-related measures support software development processes?" *Journal of Systems and Software*, vol. 81, no. 7, pp. 1248–1252, 2008.

[5] D. Graziotin, X. Wang, and P. Abrahamsson, "Happy software developers solve problems better: psychological measurements in empirical software engineering," *PeerJ*, vol. 2, p. e289, 2014.

[6] A. Salas-Vallina, J. Alegre, and R. F. Guerrero, "Happiness at work in knowledge-intensive contexts: Opening the research agenda," *European research on management and business economics*, vol. 24, no. 3, pp. 149–159, 2018.

[7] M. E. García-Buades, J. M. Peiró, M. I. Montañez-Juan, M. W. Kozusznik, and S. Ortiz-Bonnín, "Happy-productive teams and work units: A systematic review of the 'happy-productive worker thesis'," *International journal of environmental research and public health*, vol. 17, no. 1, p. 69, 2020.

[8] D. Graziotin, F. Fagerholm, X. Wang, and P. Abrahamsson, "What happens when software developers are (un) happy," *Journal of Systems and Software*, vol. 140, pp. 32–47, 2018.

[9] J. Michaelson, "National accounts of well-being," in *Encyclopedia of Quality of Life and Well-Being Research*. Springer, 2024, pp. 4571–4577, https://www.nationalaccountsofwellbeing.org/learn/download-report.html.

[10] E. Sagone and M. E. De Caroli, "A correlational study on dispositional resilience, psychological well-being, and coping strategies in university students," *American journal of educational research*, vol. 2, no. 7, pp. 463–471, 2014.

[11] M. Knobelsdorf and R. Romeike, "Creativity as a pathway to computer science," in *Proceedings of the 13th annual conference on Innovation and technology in computer science education*, 2008, pp. 286–290.

[12] L. Gonçales, K. Farias, B. da Silva, and J. Fessler, "Measuring the cognitive load of software developers: A systematic mapping study," in *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)*. IEEE, 2019, pp. 42–52.

[13] F. Fagerholm, M. Felderer, D. Fucci, M. Unterkalmsteiner, B. Marculescu, M. Martini, L. G. W. Tengberg, R. Feldt, B. Lehtelä, B. Nagyváradi *et al.*, "Cognition in software engineering: A taxonomy and survey of a half-century of research," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–36, 2022.

[14] S. K. G. Dommata and S. C. H. Konagala, "Impact of group dynamics on teams working in software engineering," 2014.

[15] K. Kaur, A. Mishra, and P. Chopra, "An amalgamation of cognitive aspects in software engineering: A content analysis," *Expert Systems with Applications*, vol. 246, p. 122968, 2024.

[16] P. N. Robillard and M. P. Robillard, "Types of collaborative work in software engineering," *Journal of Systems and Software*, vol. 53, no. 3, pp. 219–224, 2000.

[17] N. Kameo, "A culture of uncertainty: Interaction and organizational memory in software engineering teams under a productivity scheme," *Organization studies*, vol. 38, no. 6, pp. 733–752, 2017.

[18] N. Zaidman and H. Cohen, "Micro-dynamics of stress and coping with cultural differences in high tech global teams," *Journal of International Management*, vol. 26, no. 3, p. 100772, 2020.

[19] B. Penzenstadler, R. Torkar, and C. Martinez Montes, "Take a deep breath: Benefits of neuroplasticity practices for software developers and computer workers in a family of experiments," *Empirical Software Engineering*, vol. 27, no. 4, pp. 1–64, 2022.

[20] S. Rasnayaka, G. Wang, R. Shariffdeen, and G. N. Iyer, "An empirical study on usage and perceptions of llms in a software engineering project," in *Proceedings of the 1st International Workshop on Large Language Models for Code*, 2024, pp. 111–118.

[21] B. Jeffery, B. Weddle, J. Brassey, and S. Thaker, "Thriving workplaces: How employers can improve productivity and change lives," https://www.mckinsey.com/mhi/our-insights/thriving-workplaces-how-employers-can-improve-productivity-and-change-lives, January 16 2025, accessed: 2025-11-08.

[22] N. Cassee, "Sentiment in software engineering: detection and application," in *Proceedings of the 30th ACM Joint European Software Engineering Conference*

and Symposium on the Foundations of Software Engineering, 2022, pp. 1800–1804.

[23] A. C. C. França, T. B. Gouveia, P. C. Santos, C. A. Santana, and F. Q. da Silva, "Motivation in software engineering: A systematic review update," in *15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011)*. IET, 2011, pp. 154–163.

[24] T. R. Tulili, A. Capiluppi, and A. Rastogi, "Burnout in software engineering: A systematic mapping study," *Information and Software Technology*, vol. 155, p. 107116, 2023.

[25] M.-A. Storey, T. Zimmermann, C. Bird, J. Czerwonka, B. Murphy, and E. Kalliamvakou, "Towards a theory of software developer job satisfaction and perceived productivity," *IEEE Transactions on Software Engineering*, vol. 47, no. 10, pp. 2125–2142, 2021.

[26] M. C. Makel, M. S. Meyer, M. A. Simonsen, A. M. Roberts, and J. A. Plucker, "Replication is relevant to qualitative research," *Educational Research and Evaluation*, vol. 27, no. 1-2, pp. 215–219, 2022.

[27] C. M. Montes and B. Penzenstadler, "Piloting a well-being and resilience intervention in a course on digitalization for sustainability." in *ICT4S (Doctoral Symposium, Demos, Posters)*, 2023, pp. 105–118.

[28] Z. Xiao, X. Yuan, Q. V. Liao, R. Abdelghani, and P.-Y. Oudeyer, "Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding," in *Companion proceedings of the 28th international conference on intelligent user interfaces*, 2023, pp. 75–78.

[29] S. De Paoli, "Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach," *Social Science Computer Review*, p. 08944393231220483, 2023.

[30] A. P. Association, "Apa dictionary of psychology," https://dictionary.apa.org/, n.d.

[31] H. H. Publishing. (2024, Apr. 3) Understanding the stress response: Chronic activation of this survival mechanism impairs health. https://www.health.harvard.edu/staying-healthy/understanding-the-stress-response. Harvard Health Publishing. Reviewed by Howard E. LeWine, MD. [Online]. Available: https://www.health.harvard.edu/staying-healthy/understanding-the-stress-response

[32] J. Nestor, *Breath: The new science of a lost art*. Penguin UK, 2020.

[33] G. Mark, D. Gudith, and U. Klocke, "The cost of interrupted work: more speed and stress," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2008, pp. 107–110.

[34] J. Suárez and A. Vizcaíno, "Stress, motivation, and performance in global software engineering," *Journal of Software: Evolution and Process*, vol. 36, no. 5, p. e2600, 2024.

[35] P. Sterling, "Allostasis: a model of predictive regulation," *Physiology & behavior*, vol. 106, no. 1, pp. 5–15, 2012.

[36] B. S. McEwen, "Neurobiological and systemic effects of chronic stress," *Chronic stress*, vol. 1, p. 2470547017692328, 2017.

[37] B. S. McEwen and J. C. Wingfield, "The concept of allostasis in biology and biomedicine," *Hormones and behavior*, vol. 43, no. 1, pp. 2–15, 2003.

[38] J. A. Hinds and E. R. Sanchez, "The role of the hypothalamus–pituitary–adrenal (hpa) axis in test-induced anxiety: assessments, physiological responses, and molecular details," *Stresses*, vol. 2, no. 1, pp. 146–155, 2022.

[39] C. Ayada, Ü. Toru, and Y. Korkut, "The relationship of stress and blood pressure effectors," *Hippokratia*, vol. 19, no. 2, p. 99, 2015.

[40] B. S. McEwen, C. Nasca, and J. D. Gray, "Stress effects on neuronal structure: hippocampus, amygdala, and prefrontal cortex," *Neuropsychopharmacology*, vol. 41, no. 1, pp. 3–23, 2016.

[41] P. S. Kumar, "Technostress: A comprehensive literature review on dimensions, impacts, and management strategies," *Computers in Human Behavior Reports*, vol. 16, p. 100475, 2024.

[42] M. Tarafdar, Q. Tu, B. S. Ragu-Nathan, and T. Ragu-Nathan, "The impact of technostress on role stress and productivity," *Journal of management information systems*, vol. 24, no. 1, pp. 301–328, 2007.

[43] R. Ayyagari, V. Grover, and R. Purvis, "Technostress: Technological antecedents and implications," *MIS quarterly*, pp. 831–858, 2011.

[44] D. Fucci, G. Scanniello, S. Romano, and N. Juristo, "Need for sleep: the impact of a night of sleep deprivation on novice developers' performance," *IEEE Transactions on Software Engineering*, 2018.

[45] D. Graziotin, F. Fagerholm, X. Wang, and P. Abrahamsson, "On the unhappiness of software developers," in *Proceedings of the 21st international conference on evaluation and assessment in software engineering*, 2017, pp. 324–333.

[46] K. R. Head, "Minds in crisis: How the ai revolution is impacting mental health," *Health*, vol. 9, no. 3, pp. 34–44, 2025.

[47] K. T. Kalam, J. M. Rahman, M. R. Islam, and S. M. R. Dewan, "Chatgpt and mental health: Friends or foes?" *Health Science Reports*, vol. 7, no. 2, p. e1912, 2024.

[48] M. Mantzios and K. Giannou, "A real-world application of short mindfulness-based practices: a review and reflection of the literature and a practical proposition for an effortless mindful lifestyle," *American Journal of Lifestyle Medicine*, vol. 13, no. 6, pp. 520–525, 2019.

[49] J. Kabat-Zinn and T. N. Hanh, *Full catastrophe living: Using the wisdom of your body and mind to face stress, pain, and illness*. Delta, 2009.

[50] J. Kabat-Zinn, "Mindfulness-based interventions in context: past, present, and future," *Clinical psychology: Science and practice*, vol. 10, no. 2, pp. 144–156, 2003.

[51] P. Grossman, L. Niemann, S. Schmidt, and H. Walach, "Mindfulness-based stress reduction and health benefits: A meta-analysis," *Journal of psychosomatic research*, vol. 57, no. 1, pp. 35–43, 2004.

[52] S. Joyce, F. Shand, J. Tighe, S. J. Laurent, R. A. Bryant, and S. B. Harvey, "Road to resilience: a systematic review and meta-analysis of resilience training programmes and interventions," *BMJ open*, vol. 8, no. 6, p. e017858, 2018.

[53] G. W. Fincham, C. Strauss, J. Montero-Marin, and K. Cavanagh, "Effect of breathwork on stress and mental health: A meta-analysis of randomised-controlled trials," *Scientific Reports*, vol. 13, no. 1, p. 432, 2023.

[54] (1972–2019) Definition: Breathwork, noun. Oxford English Dictionary. [Online]. Available: https://www.oed.com/

[55] P. Philippot, G. Chapelle, and S. Blairy, "Respiratory feedback in the generation of emotion," *Cognition & Emotion*, vol. 16, no. 5, pp. 605–627, 2002.

[56] B. Banushi, M. Brendle, A. Ragnhildstveit, T. Murphy, C. Moore, J. Egberts, and R. Robison, "Breathwork interventions for adults with clinically diagnosed anxiety disorders: A scoping review," *Brain Sciences*, vol. 13, no. 2, p. 256, 2023.

[57] T. M. Leyro, M. V. Versella, M.-J. Yang, H. R. Brinkman, D. L. Hoyt, and P. Lehrer, "Respiratory therapy for the treatment of anxiety: Meta-analytic review and regression," *Clinical psychology review*, vol. 84, p. 101980, 2021.

[58] M. Garfinkel and H. R. Schumacher Jr, "Yoga," *Rheumatic Disease Clinics of North America*, vol. 26, no. 1, pp. 125–132, 2000.

[59] S. Khalsa, "Yoga as a therapeutic intervention," *Principles and practice of stress management*, vol. 3, pp. 449–462, 2007.

[60] A. Ross and S. Thomas, "The health benefits of yoga and exercise: a review of comparison studies," *The journal of alternative and complementary medicine*, vol. 16, no. 1, pp. 3–12, 2010.

[61] X. Li, Y. Zhou, C. Zhang, H. Wang, and X. Wang, "Neural correlates of breath work, mental imagery of yoga postures, and meditation in yoga practitioners: a functional near-infrared spectroscopy study," *Frontiers in neuroscience*, vol. 18, p. 1322071, 2024.

[62] L. J. Dimitroff, L. Sliwoski, S. O'Brien, and L. W. Nichols, "Change your life through journaling–the benefits of journaling for registered nurses," *Journal of Nursing Education and Practice*, vol. 7, no. 2, pp. 90–98, 2017.

[63] J. W. Pennebaker, *Opening up: The healing power of expressing emotions*. Guilford Press, 2012.

[64] T. K. Blake, "Journaling; an active learning technique." *International Journal of Nursing Education Scholarship*, vol. 2, no. 1, 2005.

[65] F. B. King and D. LaRocco, "E-journaling: A strategy to support student reflection and understanding," *Current Issues in Education*, vol. 9, 2006.

[66] A. P. Association, "Emotion," n.d., accessed: 2024-10-05. [Online]. Available: https://dictionary.apa.org/emotion

[67] B. Crawford, R. Soto, C. L. de la Barra, K. Crawford, and E. Olguín, "The influence of emotions on productivity in software engineering," in *International Conference on Human-Computer Interaction*. Springer, 2014, pp. 307–310.

[68] M. V. Kosti, R. Feldt, and L. Angelis, "Personality, emotional intelligence and work preferences in software engineering: An empirical study," *Information and Software Technology*, vol. 56, no. 8, pp. 973–990, 2014.

[69] P. Dewan, "Towards emotion-based collaborative software engineering," in *2015 IEEE/ACM 8th International Workshop on Cooperative and Human Aspects of Software Engineering*. IEEE, 2015, pp. 109–112.

[70] G. Willcox, "The feeling wheel: A tool for expanding awareness of emotions and increasing spontaneity and intimacy," *Transactional Analysis Journal*, vol. 12, no. 4, pp. 274–276, 1982.

[71] C. B. Seaman, "Qualitative methods in empirical studies of software engineering," *IEEE Transactions on software engineering*, vol. 25, no. 4, pp. 557–572, 1999.

[72] C. Treude, "Qualitative data analysis in software engineering: Techniques and teaching insights," in *Handbook on Teaching Empirical Software Engineering*. Springer, 2024, pp. 155–176.

[73] P. Lenberg, R. Feldt, L. Gren, L. G. Wallgren Tengberg, I. Tidefors, and D. Graziotin, "Qualitative software engineering research: Reflections and guidelines," *Journal of Software: Evolution and Process*, vol. 36, no. 6, p. e2607, 2024.

[74] Y. Dittrich, M. John, J. Singer, and B. Tessem, "Editorial for the special issue on qualitative software engineering research," *Information and software technology*, vol. 49, no. 6, pp. 531–539, 2007.

[75] V. Braun and V. Clarke, "Thematic analysis: A practical guide," 2021.

[76] T. Rath and J. K. Harter, *Wellbeing: The five essential elements*. Simon and Schuster, 2010.

[77] M. E. Seligman, *Flourish: A visionary new understanding of happiness and well-being*. Simon and Schuster, 2011.

[78] L. Michels, A. Petkova, M. Richter, A. Farley, D. Graziotin, and S. Wagner, "Overwhelmed software developers," *IEEE Software*, vol. 41, no. 04, pp. 51–59, jul 2024.

[79] D. Ford and C. Parnin, "Exploring causes of frustration for software developers," in *2015 IEEE/ACM 8th international workshop on cooperative and human aspects of software engineering*. IEEE, 2015, pp. 115–116.

[80] K. Madampe, R. Hoda, and J. Grundy, "Addressing bad feelings in agile software project contexts: Considering team welfare and developer mental health," *IEEE Software*, vol. 41, no. 04, pp. 44–50, jul 2024.

[81] D. Grassi, F. Lanubile, A. Motca-Schnabel, and N. Novielli, "A cluster-based approach for emotion recognition in software development," in *2025 IEEE/ACM 18th International Conference on Cooperative and Human Aspects of Software Engineering (CHASE)*. IEEE, 2025, pp. 239–247.

[82] C. M. Hicks, C. S. Lee, and M. Ramsey, "Developer thriving: Four sociocognitive factors that create resilient productivity on software teams," *IEEE Software*, vol. 41, no. 04, pp. 68–77, jul 2024.

[83] M.-A. Storey, T. Zimmermann, C. Bird, J. Czerwonka, B. Murphy, and E. Kalliamvakou, "Towards a theory of software developer job satisfaction and perceived productivity," *IEEE Transactions on Software Engineering*, vol. 47, no. 10, pp. 2125–2142, 2019.

[84] G. Dwomoh and A. Barcomb, "Advancing more inclusive tech careers: How people develop their potential and thrive," *IEEE Software*, vol. 41, no. 04, pp. 60–67, jul 2024.

[85] C. Martinez Montes, J. Johansson, and E. Dunvald, "Factors influencing gender representation in it faculty programmes: Insights with a focus on software engineering in a nordic context," in *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, 2025, pp. 772–782.

[86] B. Santana, S. Freire, J. Santos, and M. Mendonca, "Psychological safety in the software work environment," *IEEE Software*, vol. 41, no. 04, pp. 86–94, jul 2024.

[87] N. Wong, V. Jackson, A. Van Der Hoek, I. Ahmed, S. M. Schueller, and M. Reddy, "Mental wellbeing at work: Perspectives of software engineers," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–15.

[88] P. Godliauskas and D. Šmite, "The well-being of software engineers: a systematic literature review and a theory," *Empirical Software Engineering*, vol. 30, no. 1, p. 35, 2025.

[89] C. Martinez Montes and R. Khojah, "Emotional strain and frustration in llm interactions in software engineering," in *Proceedings of the 29th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 193–204. [Online]. Available: https://doi.org/10.1145/3756681.3756951

[90] P. den Heijer, W. Koole, and C. J. Stettina, "Don't forget to breathe: a controlled trial of mindfulness practices in agile project teams," in *International Conference on Agile Software Development*. Springer, 2017, pp. 103–118.

[91] B. Bernárdez, J. I. Panach, J. A. Parejo, A. Durán, N. Juristo, and A. Ruiz-Cortés, "An empirical study to evaluate the impact of mindfulness on helpdesk employees," *Science of Computer Programming*, vol. 230, p. 102977, 2023.

[92] B. Bernardez, A. Durán, J. A. Parejo, N. Juristo, and A. Ruiz-Cortés, "Effects of mindfulness on conceptual modeling performance: A series of experiments," *IEEE Transactions on Software Engineering*, vol. 48, no. 2, pp. 432–452, 2020.

[93] S. Romano, A. Conforti, G. Guidetti, S. Viotti, R. Ceschin, and G. Scanniello, "Mbsr at work: Perspectives from an instructor and software developers," *arXiv preprint arXiv:2506.11588*, 2025.

[94] S. Beecham, N. Baddoo, T. Hall, H. Robinson, and H. Sharp, "Motivation in software engineering: A systematic literature review," *Information and software technology*, vol. 50, no. 9-10, pp. 860–878, 2008.

[95] H. Sharp, N. Baddoo, S. Beecham, T. Hall, and H. Robinson, "Models of motivation in software engineering," *Information and software technology*, vol. 51, no. 1, pp. 219–233, 2009.

[96] S. S. Cruz, F. Q. da Silva, C. V. Monteiro, C. Santos, and M. Dos Santos, "Personality in software engineering: Preliminary findings from a systematic literature review," in *15th annual conference on Evaluation & assessment in software engineering (EASE 2011)*. IET, 2011, pp. 1–10.

[97] S. Cruz, F. Q. Da Silva, and L. F. Capretz, "Forty years of research on personality in software engineering: A mapping study," *Computers in Human Behavior*, vol. 46, pp. 94–113, 2015.

[98] A. Tarasov, "Assessing job satisfaction of software engineers using gqm approach," in *International Conference on Objects, Components, Models and Patterns*. Springer, 2019, pp. 121–135.

[99] T. Sorg, A. Abbad-Andaloussi, and B. Weber, "Towards a fine-grained analysis of cognitive load during program comprehension," in *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2022, pp. 748–752.

[100] M. Andrzejewska and A. Skawińska, "Examining students' intrinsic cognitive load during program comprehension–an eye tracking approach," in *International Conference on Artificial Intelligence in Education*. Springer, 2020, pp. 25–30.

[101] I. Mistrík, J. Grundy, A. Van der Hoek, and J. Whitehead, "Collaborative software engineering: challenges and prospects," *Collaborative software engineering*, pp. 389–403, 2010.

[102] J. Whitehead, "Collaboration in software engineering: A roadmap," in *Future of Software Engineering (FOSE'07)*. IEEE, 2007, pp. 214–225.

[103] G. Ruhe, "Software engineering decision support–a new paradigm for learning software organizations," in *International Workshop on Learning Software Organizations*. Springer, 2002, pp. 104–113.

[104] S. Baltes and S. Diehl, "Towards a theory of software development expertise," in *Proceedings of the 2018 26th acm joint meeting on european software engineering conference and symposium on the foundations of software engineering*, 2018, pp. 187–200.

[105] W. Groeneveld, L. Luyten, J. Vennekens, and K. Aerts, "Exploring the role of creativity in software engineering," in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, 2021, pp. 1–9.

[106] C. M. Montes, B. Penzenstadler, and R. Feldt, "The factors influencing well-being in software engineers: A mixed-method study," *ACM Trans. Softw. Eng. Methodol.*, Oct. 2025. [Online]. Available: https://doi.org/10.1145/3770074

[107] C. M. Montes, F. Sjögren, A. Klevfors, and B. Penzenstadler, "Qualifying and quantifying the benefits of mindfulness practices for it workers," in *2024 10th International Conference on ICT for Sustainability (ICT4S)*. IEEE, 2024, pp. 272–281.

[108] L. Cerqueira, S. Freire, D. Neves, J. Bastos, B. Santana, R. Spinola, M. Mendonca, and J. Santos, "Empathy and its effects on software practitioners' well-being and mental health," *IEEE Software*, vol. 41, no. 04, pp. 95–104, jul 2024.

[109] S. C. Müller and T. Fritz, "Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress," in *ICSE*, 2015, pp. 688–699.

[110] H. Vrzakova, A. Begel, L. Mehtätalo, and R. Bednarik, "Affect recognition in code review: An in-situ biometric study of reviewer's affect," *J. Syst. Softw.*, vol. 159, 2020.

[111] D. Girardi, F. Lanubile, N. Novielli, and A. Serebrenik, "Emotions and perceived productivity of software developers at the workplace," *IEEE Transactions on Software Engineering*, vol. 48, no. 9, pp. 3326–3341, 2021.

[112] C. Wen, P. Clough, R. Paton, and R. Middleton, "Leveraging large language models for thematic analysis: a case study in the charity sector," *AI & SOCIETY*, pp. 1–18, 2025.

[113] J. Gao, Y. Guo, G. Lim, T. Zhang, Z. Zhang, T. J.-J. Li, and S. T. Perrault, "Collabcoder: a lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–29.

[114] S. A. Gebreegziabher, Z. Zhang, X. Tang, Y. Meng, E. L. Glassman, and T. J.-J. Li, "Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–19.

[115] P. J. Pelto, "What is so new about mixed methods?" *Qualitative Health Research*, vol. 25, no. 6, pp. 734–745, 2015.

[116] J. Creswell and J. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, 2022. [Online]. Available: https://books.google.se/books?id=Rkh4EAAAQBAJ

[117] A. Bryman, *Social Research Methods 4th Ed.* Oxford University Press, 2012.

[118] B. Capili and J. K. Anastasi, "An introduction to types of quasi-experimental designs," *The American Journal of Nursing*, vol. 124, no. 11, pp. 50–52, 2024.

[119] M. L. Maciejewski, "Quasi-experimental design," *Biostatistics & Epidemiology*, vol. 4, no. 1, pp. 38–47, 2020.

[120] V. Braun and V. Clarke, *Thematic analysis.* American Psychological Association, 2012.

[121] A. N. Ghazi, K. Petersen, S. S. V. R. Reddy, and H. Nekkanti, "Survey research in software engineering: Problems and mitigation strategies," *IEEE Access*, vol. 7, pp. 24 703–24 718, 2018.

[122] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén *et al.*, *Experimentation in software engineering*. Springer, 2012, vol. 236.

[123] M. Brysbaert, "Designing and evaluating tasks to measure individual differences in experimental psychology: A tutorial," *Cognitive Research: Principles and Implications*, vol. 9, no. 1, p. 11, 2024.

[124] K.-J. Stol and B. Fitzgerald, "Guidelines for conducting software engineering research," in *Contemporary Empirical Methods in Software Engineering*. Springer, 2020, pp. 27–62.

[125] L. D. Wijsen, D. Borsboom, and A. Alexandrova, "Values in psychometrics," *Perspectives on Psychological Science*, vol. 17, no. 3, pp. 788–804, 2022.

[126] K. Swan, R. Speyer, M. Scharitzer, D. Farneti, T. Brown, V. Woisard, and R. Cordier, "Measuring what matters in healthcare: a practical guide to psychometric principles and instrument development," *Frontiers in Psychology*, vol. 14, p. 1225850, 2023.

[127] D. E. Gray, "Doing research in the real world," 2021.

[128] E. Piciucco, E. Di Lascio, E. Maiorana, S. Santini, and P. Campisi, "Biometric recognition using wearable devices in real-life settings," *Pattern Recognition Letters*, vol. 146, pp. 260–266, 2021.

[129] Y. Borgianni and L. Maccioni, "Review of the use of neurophysiological and biometric measures in experimental design research," *Ai Edam*, vol. 34, no. 2, pp. 248–285, 2020.

[130] U. Kuckartz and S. Radiker, "Qualitative content analysis: Methods, practice and software," 2023.

[131] Swedish Ethical Review Authority, "Ethical review authority," https://etikprovningsmyndigheten.se/en/, accessed: 2026-01-27.

[132] E. Emanuel, E. Abdoler, and L. Stunkel, "Research ethics: How to treat people who participate in research." 2016.

[133] M. Wilkinson and A. Moore, "Inducement in research," *Bioethics*, vol. 11, no. 5, pp. 373–389, 1997.

[134] R. Müller, D. Schischke, B. Graf, and C. H. Antoni, "How can we avoid information overload and techno-frustration as a virtual team? the effect of shared mental models of information and communication technology on information overload and techno-frustration," *Computers in Human Behavior*, vol. 138, p. 107438, 2023.

[135] Z. Zheng, K. Ning, Q. Zhong, J. Chen, W. Chen, L. Guo, W. Wang, and Y. Wang, "Towards an understanding of large language models in software engineering tasks," *Empirical Software Engineering*, vol. 30, no. 2, p. 50, 2025.

[136] T. Ahmed, P. Devanbu, C. Treude, and M. Pradel, "Can llms replace manual annotation of software engineering artifacts?" in *2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR)*. IEEE, 2025, pp. 526–538.

[137] A. Cambon, B. Hecht, B. Edelman, D. Ngwe, S. Jaffe, A. Heger, M. Vorvoreanu, S. Peng, J. Hofman, A. Farach *et al.*, "Early llm-based tools for enterprise information workers likely provide meaningful boosts to productivity," *Microsoft Research. MSR-TR-2023*, vol. 43, 2023.

[138] C. Kobiella, T. Mitrevska, A. Schmidt, and F. Draxler, "When efficiency meets fulfillment: Understanding long-term llm integration in knowledge work," in *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, 2025, pp. 1–15.

[139] S. S. Maitipe, "The psychological and workplace impact of large language models (llms) on it professionals: A survey-based study on professionals in the it industry," 2025.

[140] K. W. Brown and R. M. Ryan, "The benefits of being present: mindfulness and its role in psychological well-being." *Journal of personality and social psychology*, vol. 84, no. 4, p. 822, 2003.

[141] E. Diener, D. Wirtz, R. Biswas-Diener, W. Tov, C. Kim-Prieto, D.-w. Choi, and S. Oishi, "New measures of well-being," in *Assessing well-being*. Springer, 2009, pp. 247–266.

[142] R. C. Kessler, C. Barber, A. Beck, P. Berglund, P. D. Cleary, D. McKenas, N. Pronk, G. Simon, P. Stang, T. B. Ustun *et al.*, "The world health organization health and work performance questionnaire (hpq)," *Journal of Occupational and Environmental Medicine*, vol. 45, no. 2, pp. 156–174, 2003.

[143] J.-P. Ostberg, D. Graziotin, S. Wagner, and B. Derntl, "A methodology for psycho-biological assessment of stress in software engineering," *PeerJ Computer Science*, vol. 6, p. e286, 2020.

[144] B. Bernárdez, J. A. Parejo, M. Cruz, S. Muñoz, and A. Ruiz-Cortés, "On the impact and lessons learned from mindfulness practice in a real-world software company," in *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 2023, pp. 1–12.

[145] C. Martinez Montes and B. Penzenstadler, "Evaluating the impact of a yoga-based intervention on software engineers' well-being," in *Proceedings of the 29th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 227–237. [Online]. Available: https://doi.org/10.1145/3756681.3756950

[146] N. S. Schutte, J. M. Malouff, L. E. Hall, D. J. Haggerty, J. T. Cooper, C. J. Golden, and L. Dornheim, "Development and validation of a measure of emotional intelligence," *Personality and individual differences*, vol. 25, no. 2, pp. 167–177, 1998.

[147] G. M. Wagnild, *The resilience scale user's guide: For the US English version of the Resilience Scale and the 14-item Resilience Scale (RS-14)*. Resilience center, 2011.

[148] K. B. Carey, D. J. Neal, and S. E. Collins, "A psychometric analysis of the self-regulation questionnaire," *Addictive behaviors*, vol. 29, no. 2, pp. 253–260, 2004.

[149] G. Haugan, T. Rannestad, H. Garåsen, R. Hammervold, and G. A. Espnes, "The self-transcendence scale: an investigation of the factor structure among nursing home patients," *Journal of Holistic Nursing*, vol. 30, no. 3, pp. 147–159, 2012.

[150] V. G. Sinclair and K. A. Wallston, "The development and psychometric evaluation of the brief resilient coping scale," *Assessment*, vol. 11, no. 1, pp. 94–101, 2004.

[151] World Health Organization, *Well-being measures in primary health care: The DepCare project*. Copenhagen, Denmark: WHO Regional Office for Europe, 1998. [Online]. Available: https://www.corc.uk.net/en/outcome-measures-guidance/directory-of-outcome-measures/the-world-health-organisation-five-well-being-index-who-5/

[152] E.-H. Lee, "Review of the psychometric evidence of the perceived stress scale," *Asian nursing research*, vol. 6, no. 4, pp. 121–127, 2012.

[153] W. S. Helton and K. Näswall, "Short stress state questionnaire," *European Journal of Psychological Assessment*, 2015.

[154] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[155] A. B. Bakker and E. Demerouti, "Job demands–resources theory: taking stock and looking forward." *Journal of occupational health psychology*, vol. 22, no. 3, p. 273, 2017.

[156] U. Bronfenbrenner, *Ecological systems theory.* American Psychological Association, 2000.

[157] T. Keever, "The whole well-being model: A layered framework for thriving people, systems, and planet," *Global Advances in Integrative Medicine and Health*, vol. 14, p. 27536130251364869, 2025.

[158] S. L. Gilbert and E. K. Kelloway, "Leadership, recognition and well-being: A moderated mediational model," *Canadian Journal of Administrative Sciences/Revue canadienne des sciences de l'administration*, vol. 35, no. 4, pp. 523–534, 2018.

[159] V. Lewis, G. Marsh, J. Macmillan, and R. Mullins, "Does acknowledgement matter? an exploratory study of the value of a healthy workplace recognition scheme," *Health Promotion Journal of Australia*, vol. 31, no. 3, pp. 518–524, 2020.

[160] F. Klonek and S. Parker, "Does ai at work increase stress? text mining social media about human–ai team processes and ai control," *Journal of Organizational Behavior*, 2025.

[161] C. M. Montes, D. Grassi, N. Novielli, and B. Penzenstadler, "A multimodal approach combining biometrics and self-report instruments for monitoring stress in programming: Methodological insights," *arXiv preprint arXiv:2507.02118*, 2025.

[162] J. R. Kuntz, S. Malinen, and K. Näswall, "Employee resilience: Directions for resilience development." *Consulting Psychology Journal: Practice and Research*, vol. 69, no. 3, p. 223, 2017.

[163] P. Catapano, S. Cipolla, G. Sampogna, F. Perris, M. Luciano, F. Catapano, and A. Fiorillo, "Organizational and individual interventions for managing work-related stress in healthcare professionals: A systematic review," *Medicina*, vol. 59, no. 10, p. 1866, 2023.

[164] A. Segura-Camacho, J.-J. García-Orozco, and G. Topa, "Sustainable and healthy organizations promote employee well-being: The moderating role of selection,

optimization, and compensation strategies," *Sustainability*, vol. 10, no. 10, p. 3411, 2018.

[165] U. R. Hülsheger, H. J. Alberts, A. Feinholdt, and J. W. Lang, "Benefits of mindfulness at work: the role of mindfulness in emotion regulation, emotional exhaustion, and job satisfaction." *Journal of applied psychology*, vol. 98, no. 2, p. 310, 2013.

[166] M. A. Neff, "The feelings wheel," 2024, accessed: 2024-10-27. [Online]. Available: https://neurodivergentinsights.com/blog/the-feelings-wheel

[167] H. D. Critchley, "Electrodermal responses: what happens in the brain," *The Neuroscientist*, vol. 8, no. 2, pp. 132–142, 2002.

[168] A. Moravcsik, *Transparency in qualitative research*. SAGE Publications Limited London, 2020.

[169] H. Aguinis and A. M. Solarino, "Transparency and replicability in qualitative research: The case of interviews with elite informants," *Strategic management journal*, vol. 40, no. 8, pp. 1291–1315, 2019.

[170] T. Jowsey, V. Braun, V. Clarke, D. Lupton, and M. Fine, "We reject the use of generative artificial intelligence for reflexive qualitative research," *Qualitative Inquiry*, p. 10778004251401851, 2025.

[171] C. M. Montes, R. Feldt, C. M. Martos, S. Ouhbi, S. Premanandan, and D. Graziotin, "Large language models in thematic analysis: Prompt engineering, evaluation, and guidelines for qualitative software engineering research," *arXiv preprint arXiv:2510.18456*, 2025.

[172] J. D. Wallach, S. Serghiou, L. Chu, A. C. Egilman, V. Vasiliou, J. S. Ross, and J. P. Ioannidis, "Evaluation of confounding in epidemiologic studies assessing alcohol consumption on the risk of ischemic heart disease," *BMC medical research methodology*, vol. 20, no. 1, p. 64, 2020.

[173] D. Russo, P. H. Hanel, S. Altnickel, and N. van Berkel, "Predictors of well-being and productivity among software professionals during the covid-19 pandemic–a longitudinal study," *Empirical Software Engineering*, vol. 26, no. 4, p. 62, 2021.

[174] J. L. Magyar and C. L. Keyes, *Positive psychological assessment: A handbook of models and measures (2nd ed., pp. 389–415)*. American Psychological Association, 2019, ch. Defining, measuring, and applying subjective well-being.

[175] E. Diener, *The science of well-being: The collected works of Ed Diener*. Springer, 2009, vol. 37.

[176] A. McNaught, "Defining wellbeing," *Understanding wellbeing: An introduction for students and practitioners of health and social care*, pp. 7–23, 2011.

[177] American Psychological Association, *APA Dictionary of Psychology*. Washington, DC: American Psychological Association, 2018. [Online]. Available: https://dictionary.apa.org

[178] K. Anjali and D. Anand, "Intellectual stimulation and job commitment: A study of it professionals," *IUP Journal of Organizational Behavior*, vol. 14, no. 2, p. 28, 2015.

[179] A. Sokół and I. Figurska, "Creativity as one of the core competencies of studying knowledge workers," *Entrepreneurship and Sustainability Issues*, vol. 5, no. 1, pp. 23–35, 2017.

[180] K. Nielsen, "Work and well-being in teams," Ph.D. dissertation, University of Nottingham, 2003.

[181] K. Leifels and R. P. Zhang, "Cultural diversity in work teams and wellbeing impairments: A stress perspective," *International Journal of Cross Cultural Management*, vol. 23, no. 2, pp. 367–387, 2023.

[182] J. K. Harter, F. L. Schmidt, and C. L. Keyes, "Well-being in the workplace and its relationship to business outcomes: A review of the gallup studies." *Flourishing: Positive psychology and the life well-lived*, 2003.

[183] N. P. Leme, J. Arriel, A. Garcia, J. Azevedo, T. Sousa, L. Bueno, J. Godinho, and J. A. Pereira, "Mental health and productivity in software development: A study with the bravo central platform," in *Proceedings of the 20th Brazilian Symposium on Information Systems*, 2024, pp. 1–11.

[184] O. Sghaier, J. Boudrias, and H. Sahraoui, "Toward optimal psychological functioning in ai-driven software engineering tasks: The software evaluation for well-being and optimal psychological functioning in a context-aware environment assessment framework," *IEEE Software*, vol. 41, no. 04, pp. 105–114, jul 2024.

[185] P. Ralph, S. Baltes, G. Adisaputri, R. Torkar, V. Kovalenko, M. Kalinowski, N. Novielli, S. Yoo, X. Devroey, X. Tan *et al.*, "Pandemic programming: How COVID-19 affects software developers and how their organizations can help," *Empirical Software Engineering*, 2020.

[186] R. Santos, C. Magalhaes, and C. Franca, "Hybrid work well-being: Software professionals finding equilibrium," *IEEE Software*, vol. 41, no. 04, pp. 78–85, jul 2024.

[187] A. Singh, P. R. Anish, and S. Ghaisas, "Softment: Detecting mental health and wellbeing of women in the software sector," in *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2024, pp. 405–411.

[188] D. Stokols, "Social ecology and behavioral medicine: implications for training, practice, and policy," *Behavioral medicine*, vol. 26, no. 3, pp. 129–138, 2000.

[189] ——, "Establishing and maintaining healthy environments: Toward a social ecology of health promotion." *American psychologist*, vol. 47, no. 1, p. 6, 1992.

[190] M. Denscombe, *Ground rules for social research: Guidelines for good practice*. McGraw-Hill Education (UK), 2009.

[191] C. Martinez Montes, B. Penzenstadler, and R. Feldt, "Online appendix of the paper: The factors influencing well-being in software engineers: A mixed-method study [data set]," aug 2025. [Online]. Available: https://doi.org/10.5281/zenodo.17055685

[192] D. G. Oliver, J. M. Serovich, and T. L. Mason, "Constraints and opportunities with interview transcription: Towards reflection in qualitative research," *Social forces*, vol. 84, no. 2, pp. 1273–1289, 2005.

[193] V. Clarke and V. Braun, *Thematic analysis: a practical guide*. SAGE Publications Ltd, 2021.

[194] K. Ruggeri, E. Garcia-Garzon, Á. Maguire, S. Matz, and F. A. Huppert, "Well-being is more than happiness and life satisfaction: a multidimensional analysis of 21 countries," *Health and quality of life outcomes*, vol. 18, pp. 1–16, 2020.

[195] A. Tsatsoulis and S. Fountoulakis, "The protective role of exercise on stress system dysregulation and comorbidities," *Annals of the New York Academy of Sciences*, vol. 1083, no. 1, pp. 196–213, 2006.

[196] A. L. T. Hirschle and S. M. G. Gondim, "Stress and well-being at work: a literature review," *Ciência & Saúde Coletiva*, vol. 25, pp. 2721–2736, 2020.

[197] D. W. de Guerre, M. Emery, P. Aughton, and A. S. Trull, "Structure underlies other organizational determinants of mental health: recent results confirm early sociotechnical systems research," *Systemic Practice and Action Research*, vol. 21, pp. 359–379, 2008.

[198] D. S. Syahreza, P. Lumbanraja, R. F. Dalimunthe, and Y. Absah, "Compensation, employee performance, and mediating role of retention: A study of differential semantic scales," 2017.

[199] N. P. R. de Souza and K. Gama, "Diversity and inclusion: Culture and perception in information technology companies," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 15, no. 4, pp. 352–361, 2020.

[200] J. Teevan, N. Baym, J. Butler, B. Hecht, S. Jaffe, K. Nowak, A. Sellen, L. Yang, M. Ash, K. Awori *et al.*, "Microsoft new future of work report 2022," *Jenna Butler (Senior Applied Research Scientist)*, 2022.

[201] D. Scholarios and A. Marks, "Work-life balance and the software worker," *Human Resource Management Journal*, vol. 14, no. 2, pp. 54–74, 2004.

[202] T. C. D'Oliveira and L. Persico, "Workplace isolation, loneliness and wellbeing at work: The mediating role of task interdependence and supportive behaviours," *Applied Ergonomics*, vol. 106, p. 103894, 2023.

[203] A. Nawrat, "Tech overload is killing productivity," https://www.unleash.ai/wellbeing/tech-overload-is-killing-productivity/, 2023, accessed: 2023-12-11, 15:52 CEST.

[204] A. Petermans and R. Cain, *Design for Wellbeing: An Applied Approach*, 2020.

[205] M. for better mental health, "How to promote wellbeing and tackle the causes of work-related mental health problems," https://www.mind.org.uk/media-a/4662/resource3_howtopromotewellbeingfinal.pdf.

[206] B. Maxwell, M. Martin, and D. Kelly, "Translation and cultural adaptation of the survey instruments," *Third international mathematics and science study (TIMSS) technical report*, vol. 1, pp. 159–169, 1996.

[207] O. Asare, M. Nagappan, and N. Asokan, "Is github's copilot as bad as humans at introducing vulnerabilities in code?" *Empirical Software Engineering*, vol. 28, no. 6, p. 129, 2023.

[208] S. Lubos, A. Felfernig, T. N. T. Tran, D. Garber, M. El Mansi, S. P. Erdeniz, and V.-M. Le, "Leveraging llms for the quality assurance of software requirements," in *2024 IEEE 32nd International Requirements Engineering Conference (RE)*. IEEE, 2024, pp. 389–397.

[209] R. Khojah, M. Mohamad, P. Leitner, and F. G. de Oliveira Neto, "Beyond code generation: An observational study of chatgpt usage in software engineering practice," *Proceedings of the ACM on Software Engineering*, vol. 1, no. FSE, pp. 1819–1840, 2024.

[210] M.-A. Storey and A. Zagalsky, "Disrupting developer productivity one bot at a time," in *Proceedings of the 2016 24th ACM SIGSOFT international symposium on foundations of software engineering*, 2016, pp. 928–931.

[211] J. D. Weisz, M. Muller, S. I. Ross, F. Martinez, S. Houde, M. Agarwal, K. Talamadupula, and J. T. Richards, "Better together? an evaluation of ai-supported code translation," in *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 2022, pp. 369–391.

[212] A. M. Dakhel, V. Majdinasab, A. Nikanjam, F. Khomh, M. C. Desmarais, and Z. M. J. Jiang, "Github copilot ai pair programmer: Asset or liability?" *Journal of Systems and Software*, vol. 203, p. 111734, 2023.

[213] J. Wang, Y. Huang, C. Chen, Z. Liu, S. Wang, and Q. Wang, "Software testing with large language models: Survey, landscape, and vision," *IEEE Transactions on Software Engineering*, 2024.

[214] J. J. Norheim, E. Rebentisch, D. Xiao, L. Draeger, A. Kerbrat, and O. L. de Weck, "Challenges in applying large language models to requirements engineering tasks," *Design Science*, vol. 10, p. e16, 2024.

[215] J. G. Meyer, R. J. Urbanowicz, P. C. Martin, K. O'Connor, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez *et al.*, "Chatgpt and large language models in academia: opportunities and challenges," *BioData Mining*, vol. 16, no. 1, p. 20, 2023.

[216] A. P. Association, "Frustration," n.d., accessed: 2025-04-12. [Online]. Available: https://dictionary.apa.org/frustration

[217] M. A. Neff, "Emotional identification," 2023, accessed: 2024-10-07. [Online]. Available: https://neurodivergentinsights.com/blog/emotional-identification

[218] C.-H. Chang, R. E. Johnson, and L.-Q. Yang, "Emotional strain and organizational citizenship behaviours: A meta-analysis and review," *Work & Stress*, vol. 21, no. 4, pp. 312–332, 2007.

[219] M. Salanova, S. Llorens, and M. Ventura, "Technostress: The dark side of technologies," in *The impact of ICT on quality of working life*. Springer, 2014, pp. 87–103.

[220] S. C. Srivastava, S. Chandra, and A. Shirish, "Technostress creators and job outcomes: theorising the moderating influence of personality traits," *Information Systems Journal*, vol. 25, no. 4, pp. 355–401, 2015.

[221] R. A. Calvo and D. Peters, *Positive computing: technology for wellbeing and human potential*. MIT press, 2014.

[222] G. Mark, S. Iqbal, M. Czerwinski, and P. Johns, "Focused, aroused, but so distractible: Temporal perspectives on multitasking and communications," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015, pp. 903–916.

[223] J. Wester, T. Schrills, H. Pohl, and N. van Berkel, ""as an ai language model, i cannot": Investigating llm denials of user requests," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–14.

[224] N. R. Ferreri and C. B. Mayhorn, "Identifying and understanding individual differences in frustration with technology," *Theoretical Issues in Ergonomics Science*, vol. 24, no. 4, pp. 461–479, 2023.

[225] C. Foster and J. Sayers, "Exploring physiotherapists' emotion work in private practice," *New Zealand Journal of Physiotherapy*, vol. 40, no. 1, pp. 17–23, 2012.

[226] M. V. Mäntylä, K. Petersen, T. O. Lehtinen, and C. Lassenius, "Time pressure: a controlled experiment of test case development and requirements review," in *Proceedings of the 36th International Conference on Software Engineering*, 2014, pp. 83–94.

[227] D. Graziotin, X. Wang, and P. Abrahamsson, "Software developers, moods, emotions, and performance," *arXiv preprint arXiv:1405.4422*, 2014.

[228] E. Marcos, R. Hens, T. Puebla, and J. M. Vara, "Applying emotional team coaching to software development," *IEEE Software*, vol. 38, no. 4, pp. 85–93, 2020.

[229] M. Mäntylä, B. Adams, G. Destefanis, D. Graziotin, and M. Ortu, "Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity?" in *Proceedings of the 13th international conference on mining software repositories*, 2016, pp. 247–258.

[230] J. A. Russell, "Culture and the categorization of emotions." *Psychological bulletin*, vol. 110, no. 3, p. 426, 1991.

[231] D. Girardi, F. Lanubile, N. Novielli, and A. Serebrenik, "Emotions and perceived productivity of software developers at the workplace," *IEEE Transactions on Software Engineering*, vol. 48, no. 9, pp. 3326–3341, 2022.

[232] M. Sánchez-Gordón and R. Colomo-Palacios, "Taking the emotional pulse of software engineering—a systematic literature review of empirical studies," *Information and Software Technology*, vol. 115, pp. 23–43, 2019.

[233] D. Graziotin, X. Wang, and P. Abrahamsson, "Are happy developers more productive? the correlation of affective states of software developers and their self-assessed productivity," in *Product-Focused Software Process Improvement: 14th International Conference, PROFES 2013, Paphos, Cyprus, June 12-14, 2013. Proceedings 14*. Springer, 2013, pp. 50–64.

[234] S. C. Müller and T. Fritz, "Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress," in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, vol. 1. IEEE, 2015, pp. 688–699.

[235] D. Girardi, N. Novielli, D. Fucci, and F. Lanubile, "Recognizing developers' emotions while programming," in *Proceedings of the ACM/IEEE 42nd international conference on software engineering*, 2020, pp. 666–677.

[236] R. A. Stebbins, *Exploratory research in the social sciences*. Sage, 2001, vol. 48.

[237] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili *et al.*, "A survey on large language models: Applications, challenges, limitations, and practical usage," *Authorea Preprints*, vol. 3, 2023.

[238] C. Martinez Montes and R. Khojah, "Replication Package for The Study "Emotional Strain and Frustration in LLM Interactions in Software Engineering" ," Jan. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.14766606

[239] S. Baltes and P. Ralph, "Sampling in software engineering research: A critical review and guidelines," *Empirical Software Engineering*, vol. 27, no. 4, p. 94, 2022.

[240] C. Erlingsson and P. Brysiewicz, "A hands-on guide to doing content analysis," *African journal of emergency medicine*, vol. 7, no. 3, pp. 93–99, 2017.

[241] HuggingFace, "Transformers documentation: Llm tutorial," 2023, accessed: 2024-11-04.

[242] R. Khojah, F. G. de Oliveira Neto, and P. Leitner, "From human-to-human to human-to-bot conversations in software engineering," in *Proceedings of the 1st ACM International Conference on AI-Powered Software*, ser. AIware 2024. New York, NY, USA: Association for Computing Machinery, 2024, p. 38–44. [Online]. Available: https://doi.org/10.1145/3664646.3664761

[243] S. Fitzgerald, G. Lewandowski, R. McCauley, L. Murphy, B. Simon, L. Thomas, and C. Zander, "Debugging: finding, fixing and flailing, a multi-institutional study of novice debuggers," *Computer Science Education*, vol. 18, no. 2, pp. 93–116, 2008.

[244] Y. Liu, T. Le-Cong, R. Widyasari, C. Tantithamthavorn, L. Li, X.-B. D. Le, and D. Lo, "Refining chatgpt-generated code: Characterizing and mitigating code quality issues," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 5, p. 26, 2024.

[245] I. Ceaparu, J. Lazar, K. Bessiere, J. Robinson, and B. Shneiderman, "Determining causes and severity of end-user frustration," *International journal of human-computer interaction*, vol. 17, no. 3, pp. 333–356, 2004.

[246] J. Lazar, A. Jones, and B. Shneiderman, "Workplace user frustration with computers: An exploratory investigation of the causes and severity," *Behaviour & Information Technology*, vol. 25, no. 03, pp. 239–251, 2006.

[247] J. Prather, B. N. Reeves, P. Denny, B. A. Becker, J. Leinonen, A. Luxton-Reilly, G. Powell, J. Finnie-Ansley, and E. A. Santos, ""it's weird that it knows what i want": Usability and interactions with copilot for novice programmers," *ACM Trans. Comput.-Hum. Interact.*, vol. 31, no. 1, Nov. 2023. [Online]. Available: https://doi.org/10.1145/3617367

[248] F. Eshraghian, N. Hafezieh, F. Farivar, and S. de Cesare, "Ai in software programming: understanding emotional responses to github copilot," *Information Technology & People*, 2024.

[249] G. A. Opoku-Boateng, "User frustration in hit interfaces: Exploring past hci research for a better understanding of clinicians'experiences," in *AMIA Annual Symposium Proceedings*, vol. 2015, 2015, p. 1008.

[250] D. S. Tawfik, A. Sinha, M. Bayati, K. C. Adair, T. D. Shanafelt, J. B. Sexton, and J. Profit, "Frustration with technology and its relation to emotional exhaustion among health care workers: cross-sectional observational study," *Journal of medical Internet research*, vol. 23, no. 7, p. e26817, 2021.

[251] C. França, H. Sharp, and F. Q. Da Silva, "Motivated software engineers are engaged and focused, while satisfied ones are happy," p. 8, 2014.

[252] X. Wang, X. Li, Z. Yin, Y. Wu, and J. Liu, "Emotional intelligence of large language models," *Journal of Pacific Rim Psychology*, vol. 17, p. 18344909231213958, 2023. [Online]. Available: https://doi.org/10.1177/18344909231213958

[253] L. Erlenhov, F. G. de Oliveira Neto, R. Scandariato, and P. Leitner, "Current and future bots in software development," in *2019 IEEE/ACM 1st International Workshop on Bots in Software Engineering (BotSE)*. IEEE, 2019, pp. 7–11.

[254] S. L. Levine, C. J. Brabander, A. M. Moore, A. C. Holding, and R. Koestner, "Unhappy or unsatisfied: distinguishing the role of negative affect and need

frustration in depressive symptoms over the academic year and during the covid-19 pandemic," *Motivation and Emotion*, vol. 46, no. 1, pp. 126–136, 2022.

[255] K. G. Barman, N. Wood, and P. Pawlowski, "Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for llm use," *Ethics and Information Technology*, vol. 26, no. 3, p. 47, 2024.

[256] P. Ladwig, K. E. Dalrymple, D. Brossard, D. A. Scheufele, and E. A. Corley, "Perceived familiarity or factual knowledge? comparing operationalizations of scientific understanding," *Science and Public Policy*, vol. 39, no. 6, pp. 761–774, 2012.

[257] H. Herrman, D. E. Stewart, N. Diaz-Granados, E. L. Berger, B. Jackson, and T. Yuen, "What is resilience?" *The Canadian Journal of Psychiatry*, vol. 56, no. 5, pp. 258–265, 2011.

[258] J. M. Smith, "Is computational thinking critical thinking?" in *Expanding Global Horizons Through Technology Enhanced Language Learning*. Springer, 2021, pp. 191–201.

[259] S. Easterbrook, "From computational thinking to systems thinking," in *The 2nd international conference ICT for Sustainability (ICT4S), Stockholm*, 2014.

[260] A. Cockburn, *Agile software development: the cooperative game*. Pearson Education, 2006.

[261] J. D. Herbsleb and D. Moitra, "Global software development," *IEEE software*, vol. 18, no. 2, pp. 16–20, 2001.

[262] F. Derksen, J. Bensing, and A. Lagro-Janssen, "Effectiveness of empathy in general practice: a systematic review," *British Journal of General Practice*, vol. 63, no. 606, pp. e76–e84, 2013.

[263] B. J. Broome, "Building shared meaning: Implications of a relational approach to empathy for teaching intercultural communication," *Communication education*, vol. 40, no. 3, pp. 235–249, 1991.

[264] H. Riess, "The science of empathy," *Journal of patient experience*, vol. 4, no. 2, pp. 74–77, 2017.

[265] A. N. Meyer, "Fostering software developer productivity through awareness increase and goal-setting," Ph.D. dissertation, University of Zurich, 2019.

[266] S. Konrath, "The empathy paradox: Increasing disconnection in the age of increasing connection," in *Handbook of research on technoself: Identity in a technological society*. IGI Global, 2013, pp. 204–228.

[267] N. Donnelly and S. B. Proctor-Thomson, "Disrupted work: home-based teleworking (hbtw) in the aftermath of a natural disaster," *New Technology, Work and Employment*, vol. 30, no. 1, pp. 47–61, 2015.

[268] B. Penzenstadler, "What is your remedy to cognitive overload?" *IEEE Software Blog*, 2020, http://blog.ieeesoftware.org/2020/03/what-is-your-remedy-to-cognitive.html?m=1.

[269] T. Medvedyk, I. Antoniuk, and S. Lebid, "Influence of stress factors on cognitive tasks performance," in *2019 IEEE 20th International Conference on Computational Problems of Electrical Engineering (CPEE)*. IEEE, 2019, pp. 1–4.

[270] T. Maudgalya, S. Wallace, N. Daraiseh, and S. Salem, "Workplace stress factors and 'burnout' among information technology professionals: A systematic review," *Theoretical Issues in Ergonomics Science*, vol. 7, no. 3, pp. 285–297, 2006.

[271] T. Marek, W. B. Schaufeli, and C. Maslach, *Professional burnout: Recent developments in theory and research*. Routledge, 2017.

[272] S. Brown, *Speed: facing our addiction to fast and faster–and overcoming our fear of slowing down*. Berkley, 2014.

[273] B. Akula and J. Cusick, "Impact of overtime and stress on software quality," in *4th International Symposium on Management, Engineering, and Informatics (MEI 2008), Orlando, Florida, USA*, 2008.

[274] A. Amin, S. Basri, M. F. Hassan, and M. Rehman, "Software engineering occupational stress and knowledge sharing in the context of global software development," in *2011 National Postgraduate Conference*. IEEE, 2011, pp. 1–4.

[275] O. M. Buxton, S. W. Cain, S. P. O'Connor, J. H. Porter, J. F. Duffy, W. Wang, C. A. Czeisler, and S. A. Shea, "Adverse metabolic consequences in humans of prolonged sleep restriction combined with circadian disruption," *Science translational medicine*, vol. 4, no. 129, pp. 129ra43–129ra43, 2012.

[276] E. L. Haus and M. H. Smolensky, "Shift work and cancer risk: potential mechanistic roles of circadian disruption, light at night, and sleep deprivation," *Sleep medicine reviews*, vol. 17, no. 4, pp. 273–284, 2013.

[277] M. Hafner, M. Stepanek, J. Taylor, W. M. Troxel, and C. Van Stolk, "Why sleep matters—the economic costs of insufficient sleep: a cross-country comparative analysis," *Rand health quarterly*, vol. 6, no. 4, 2017.

[278] M. Lavallée and P. N. Robillard, "Why good developers write bad code: An observational case study of the impacts of organizational factors on software quality," in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, vol. 1. IEEE, 2015, pp. 677–687.

[279] R. Panikkar, *The Vedic experience: Mantramañjarī: an anthology of the Vedas for modern man and contemporary celebration*. Motilal Banarsidass Publ., 1994.

[280] D. Elliot, *The Reluctant Healer*. Hawk Press, 2005.

[281] D. Brulé, *Just Breathe: Mastering Breathwork*. Simon and Schuster, 2017.

[282] L. Squire, D. Berg, F. E. Bloom, S. Du Lac, A. Ghosh, and N. C. Spitzer, *Fundamental neuroscience*. Academic press, 2012.

[283] E. M. Seppälä, C. Bradley, J. Moeller, L. Harouni, D. Nandamudi, and M. A. Brackett, "Promoting mental health and psychological thriving in university students: a randomized controlled trial of three well-being interventions," *Frontiers in psychiatry*, vol. 11, p. 590, 2020.

[284] P. Sharma, A. Thapliyal, T. Chandra, S. Singh, H. Baduni, and S. M. Waheed, "Rhythmic breathing: immunological, biochemical, and physiological effects on health," *Adv Mind Body Med*, vol. 29, no. 1, pp. 18–25, 2015.

[285] J. Walker III and D. Pacik, "Controlled rhythmic yogic breathing as complementary treatment for post-traumatic stress disorder in military veterans: a case series," *Medical acupuncture*, vol. 29, no. 4, pp. 232–238, 2017.

[286] R. P. Brown and P. L. Gerbarg, "Sudarshan kriya yogic breathing in the treatment of stress, anxiety, and depression: part i—neurophysiologic model," *Journal of Alternative & Complementary Medicine*, vol. 11, no. 1, pp. 189–201, 2005.

[287] W. James, *Memories and studies*. New York, Longmans, 1911 (republished in 1924).

[288] M. Samuelson, J. Carmody, J. Kabat-Zinn, and M. A. Bratt, "Mindfulness-based stress reduction in massachusetts correctional facilities," *The Prison Journal*, vol. 87, no. 2, pp. 254–268, 2007.

[289] J. Kabat-Zinn, "Meditation is not what you think," *Mindfulness*, vol. 12, no. 3, pp. 784–787, 2021.

[290] D. Westen, *Psychology: Mind, brain, & culture*. John Wiley & Sons, 1996.

[291] B. Bernárdez, A. Durán, J. A. Parejo, and A. Ruiz-Cortés, "A controlled experiment to evaluate the effects of mindfulness in software engineering," in *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 2014, pp. 1–10.

[292] ——, "An experimental replication on the effect of the practice of mindfulness in conceptual modeling performance," *Journal of Systems and Software*, vol. 136, pp. 153–172, 2018.

[293] L. F. Capretz, "Personality types in software engineering," *International Journal of Human-Computer Studies*, vol. 58, no. 2, pp. 207–214, 2003.

[294] B. Rieken, S. Shapiro, S. Gilmartin, and S. Sheppard, "How mindfulness can help engineers solve problems. harvard business review." *Harvard business review*, 2019, https://hbr.org/2019/01/how-mindfulness-can-help-engineers-solve-problems.

[295] S. Sheppard, S. Gilmartin, H. L. Chen, K. Donaldson, G. Lichtenstein, O. Eris, M. Lande, and G. Toye, "Exploring the engineering student experience: Findings from the academic pathways of people learning engineering survey (apples). tr-10-01." *Center for the Advancement of Engineering Education (NJ1)*, 2010.

[296] W. James, *The principles of psychology.* Cosimo, Inc., 2007, vol. 1.

[297] D. Kahneman, *Thinking, fast and slow.* Macmillan, 2011.

[298] D. J. Chalmers, *The conscious mind: In search of a fundamental theory.* Oxford Paperbacks, 1996.

[299] H. Selye, "What is stress," *Metabolism*, vol. 5, no. 5, pp. 525–530, 1956.

[300] R. Crisp, "Well-being," *Stanford Encyclopedia of Philosophy*, 2001.

[301] A. Bandura and S. Wessels, "Self-efficacy," 1994.

[302] S. Evans, J. C. Tsao, B. Sternlieb, and L. K. Zeltzer, "Using the biopsychosocial model to understand the health benefits of yoga," *Journal of complementary and integrative medicine*, vol. 6, no. 1, 2009.

[303] J. D. Creswell, L. E. Pacilio, E. K. Lindsay, and K. W. Brown, "Brief mindfulness meditation training alters psychological and neuroendocrine responses to social evaluative stress," *Psychoneuroendocrinology*, vol. 44, pp. 1–12, 2014.

[304] S. Kotler, *The Art of Impossible.* Harper Wave, 2021.

[305] L. Feldmann-Barrett, *Seven and a Half Lessons About the Brain.* Picador, UK, 2020.

[306] C.-M. Tan, D. Goleman, and J. Kabat-Zinn, *Search Inside Yourself: The Unexpected Path to Achieving Success, Happiness (and World Peace).* HarperCollins, 2012.

[307] C. M. Chlebak, S. James, M. J. Westwood, A. Gockel, B. D. Zumbo, and S. L. Shapiro, "Mindfulness meditation & gratitude journalling," *Counseling et spiritualité/Counselling and Spirituality*, vol. 32, no. 2, pp. 79–103, 2013.

[308] R. L. Wasserstein and N. A. Lazar, "The asa statement on p-values: context, process, and purpose," 2016.

[309] S. Baltes and P. Ralph, "Sampling in software engineering research: A critical review and guidelines," *CoRR*, vol. abs/2002.07764, 2020. [Online]. Available: https://arxiv.org/abs/2002.07764

[310] R. Feldt, R. Torkar, L. Angelis, and M. Samuelsson, "Towards individualized software engineering: empirical studies should collect psychometrics," in *Proceedings of the 2008 international workshop on Cooperative and human aspects of software engineering*, 2008, pp. 49–52.

[311] P. Lenberg, R. Feldt, and L. G. Wallgren, "Behavioral software engineering: A definition and systematic literature review," *Journal of Systems and software*, vol. 107, pp. 15–37, 2015.

[312] L. Gren, "Standards of validity and the validity of standards in behavioral software engineering research: the perspective of psychological test theory," in *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 2018, pp. 1–4.

[313] S. Wagner, D. Mendez, M. Felderer, D. Graziotin, and M. Kalinowski, "Challenges in survey research," in *Contemporary Empirical Methods in Software Engineering*. Springer, 2020, pp. 93–125.

[314] S. L. Shapiro, D. Oman, C. E. Thoresen, T. G. Plante, and T. Flinders, "Cultivating mindfulness: effects on well-being," *Journal of clinical psychology*, vol. 64, no. 7, pp. 840–862, 2008.

[315] R. A. Baer, G. T. Smith, J. Hopkins, J. Krietemeyer, and L. Toney, "Using self-report assessment methods to explore facets of mindfulness," *Assessment*, vol. 13, no. 1, pp. 27–45, 2006.

[316] J. MacKillop and E. J. Anderson, "Further psychometric validation of the mindful attention awareness scale (maas)," *Journal of Psychopathology and Behavioral Assessment*, vol. 29, no. 4, pp. 289–293, 2007.

[317] S. Evans, S. Ferrando, M. Findler, C. Stowell, C. Smart, and D. Haglin, "Mindfulness-based cognitive therapy for generalized anxiety disorder," *Journal of anxiety disorders*, vol. 22, no. 4, pp. 716–721, 2008.

[318] J. Carmody, G. Reed, J. Kristeller, and P. Merriam, "Mindfulness, spirituality, and health-related symptoms," *Journal of psychosomatic research*, vol. 64, no. 4, pp. 393–403, 2008.

[319] S. Barajas and L. Garra, "Mindfulness and psychopathology: Adaptation of the mindful attention awareness scale (maas) in a spanish sample," *Clínica y Salud*, vol. 25, no. 1, pp. 49–56, 2014.

[320] Y.-Q. Deng, S. Li, Y.-Y. Tang, L.-H. Zhu, R. Ryan, and K. Brown, "Psychometric properties of the chinese translation of the mindful attention awareness scale (maas)," *Mindfulness*, vol. 3, no. 1, pp. 10–14, 2012.

[321] V. Höfling, H. Moosbrugger, K. Schermelleh-Engel, and T. Heidenreich, "Mindfulness or mindlessness?" *European Journal of Psychological Assessment*, 2011.

[322] M. A. Busseri, "Examining the structure of subjective well-being through meta-analysis of the associations among positive affect, negative affect, and life satisfaction," *Personality and Individual Differences*, vol. 122, pp. 68–71, 2018.

[323] V. Jovanović, "Beyond the panas: Incremental validity of the scale of positive and negative experience (spane) in relation to well-being," *Personality and Individual Differences*, vol. 86, pp. 487–491, 2015.

[324] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[325] P. A. Scuffham, N. Vecchio, and H. A. Whiteford, "Exploring the validity of hpq-based presenteeism measures to estimate productivity losses in the health and education sectors," *Medical Decision Making*, vol. 34, no. 1, pp. 127–137, 2014.

[326] M. Jerusalem and R. Schwarzer, "Skala zur allgemeinen selbstwirksamkeitserwartung," *Skalen zur Erfassung von Lehrer-und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen. Berlin: Freie Universität Berlin*, 1999.

[327] A. Bandura, W. Freeman, and R. Lightsey, "Self-efficacy: The exercise of control," 1999.

[328] M. Jahoda, *Current concepts of positive mental health.* New York, NY, US: Basic Books, 1958. [Online]. Available: https://doi.org/10.1037/11258-000

[329] P. Bech, "Health-related quality of life measurements in the assessment of pain clinic results," *Acta Anaesthesiologica Scandinavica*, vol. 43, no. 9, pp. 893–896, 1999.

[330] C. W. Topp, S. D. Østergaard, S. Søndergaard, and P. Bech, "The who-5 well-being index: a systematic review of the literature," *Psychotherapy and psychosomatics*, vol. 84, no. 3, pp. 167–176, 2015.

[331] B. Penzenstadler, "Rise 2 flow replication package," Zenodo, 2021. [Online]. Available: https://zenodo.org/record/5082388

[332] M. Houben, W. Van Den Noortgate, and P. Kuppens, "The relation between short-term emotion dynamics and psychological well-being: A meta-analysis." *Psychological bulletin*, vol. 141, no. 4, p. 901, 2015.

[333] R. McElreath, *Statistical rethinking: A Bayesian course with examples in R and Stan.* Chapman and Hall/CRC, 2018.

[334] L. Cheng, S. Ramchandran, T. Vatanen, N. Lietzén, R. Lahesmaa, A. Vehtari, and H. Lähdesmäki, "An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data," *Nature Communications*, vol. 10, no. 1, p. 1798, 2019.

[335] T. Paananen, J. Piironen, M. R. Andersen, and A. Vehtari, "Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution," in *Proceedings of Machine Learning Research*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 1743–1752. [Online]. Available: http://proceedings.mlr.press/v89/paananen19a.html

[336] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari, "GPstuff: Bayesian modeling with Gaussian Processes," *Journal of Machine Learning Research*, vol. 14, pp. 1175–1179, Apr. 2013.

[337] V. Braun and V. Clarke, "To saturate or not to saturate? questioning data saturation as a useful concept for thematic analysis and sample-size rationales," *Qualitative research in sport, exercise and health*, vol. 13, no. 2, pp. 201–216, 2021.

[338] N. T. Van Dam, M. Earleywine, and A. Borders, "Measuring mindfulness? an item response theory analysis of the mindful attention awareness scale," *Personality and Individual Differences*, vol. 49, no. 7, pp. 805–810, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191886910003727

[339] A. Sutton, "Measuring the effects of self-awareness: Construction of the self-awareness outcomes questionnaire," *Europe's journal of psychology*, vol. 12, no. 4, p. 645, 2016.

[340] C. Fletcher and C. Bailey, "Assessing self-awareness: some issues and methods," *Journal of managerial psychology*, 2003.

[341] W. Pavot and E. Diener, "The subjective evaluation of well-being in adulthood: Findings and implications," *Ageing International*, vol. 29, no. 2, pp. 113–135, 2004.

[342] D. A. Dillman, J. D. Smyth, and L. M. Christian, *Internet, phone, mail, and mixed-mode surveys: the tailored design method*. John Wiley & Sons, 2014.

[343] P. Krishnan, "A review of the non-equivalent control group post-test-only design," *Nurse researcher*, vol. 29, no. 2, 2021.

[344] S. G. Hofmann, G. J. Asmundson, and A. T. Beck, "The science of cognitive therapy," *Behavior therapy*, vol. 44, no. 2, pp. 199–212, 2013.

[345] M. Fisher, "A their of public well-being," *BMC public health*, vol. 19, 2019.

[346] V. Dagenais-Desmarais and A. Savoie, "What is psychological well-being, really? a grassroots approach from the organizational sciences," *Journal of Happiness Studies*, vol. 13, no. 4, pp. 659–684, 2012.

[347] F. S. Barrett, M. W. Johnson, and R. R. Griffiths, "Validation of the revised mystical experience questionnaire in experimental sessions with psilocybin," *Journal of Psychopharmacology*, vol. 29, no. 11, pp. 1182–1190, 2015.

[348] J. J. Lucas, "Mindful energy and information flow: A reflective account of self connection during covid-19," *Qualitative Social Work*, vol. 20, no. 1-2, pp. 214–221, 2021.

[349] R. Dass, *Be here now*. Three Rivers Press (CA), 1971.

[350] D. J. Good, C. J. Lyddy, T. M. Glomb, J. E. Bono, K. W. Brown, M. K. Duffy, R. A. Baer, J. A. Brewer, and S. W. Lazar, "Contemplating mindfulness at work: An integrative review," *Journal of management*, vol. 42, no. 1, pp. 114–142, 2016.

[351] P. P. Schultz, R. M. Ryan, C. P. Niemiec, N. Legate, and G. C. Williams, "Mindfulness, work climate, and psychological need satisfaction in employee well-being," *Mindfulness*, vol. 6, pp. 971–985, 2015.

[352] K. Everson, "Sap's sold on self-awareness," *Chief Learning Officer*, 2015.

[353] E. F. Bryant, *The yoga sutras of Patanjali: A new edition, translation, and commentary*. North Point Press, 2015.

[354] N. Markil, C. A. Geithner, and T. M. Penhollow, "Hatha yoga: Benefits and principles for a more meaningful practice," *ACSM's Health & Fitness Journal*, vol. 14, no. 5, pp. 19–24, 2010.

[355] K. Luu and P. A. Hall, "Hatha yoga and executive function: a systematic review," *The Journal of Alternative and Complementary Medicine*, vol. 22, no. 2, pp. 125–133, 2016.

[356] H. Cramer, R. Lauche, J. Langhorst, and G. Dobos, "Yoga for depression: A systematic review and meta-analysis," *Depression and anxiety*, vol. 30, no. 11, pp. 1068–1083, 2013.

[357] H. Cramer, R. Lauche, D. Anheyer, K. Pilkington, M. de Manincor, G. Dobos, and L. Ward, "Yoga for anxiety: A systematic review and meta-analysis of randomized controlled trials," *Depression and anxiety*, vol. 35, no. 9, pp. 830–843, 2018.

[358] S. G. Hofmann, G. Andreoli, J. K. Carpenter, and J. Curtiss, "Effect of hatha yoga on anxiety: a meta-analysis," *Journal of Evidence-Based Medicine*, vol. 9, no. 3, pp. 116–124, 2016.

[359] F.-J. Huang, D.-K. Chien, and U.-L. Chung, "Effects of hatha yoga on stress in middle-aged women," *Journal of Nursing Research*, vol. 21, no. 1, pp. 59–66, 2013.

[360] L. Puerto Valencia, A. Weber, H. Spegel, R. Bögle, A. Selmani, S. Heinze, and C. Herr, "Yoga in the workplace and health outcomes: a systematic review," *Occupational Medicine*, vol. 69, no. 3, pp. 195–203, 2019.

[361] K. C. Daane, "Yoga as a means of increasing job satisfaction in the workplace," Ph.D. dissertation, University of Wisconsin–Stout, 2018.

[362] A. C. Hafenbrack, "Mindfulness meditation as an on-the-spot workplace intervention," *Journal of Business Research*, vol. 75, pp. 118–129, 2017.

[363] B. Izydorczyk, K. Sitnik-Warchulska, A. Kühn-Dymecka, and S. Lizińczyk, "Resilience, sense of coherence, and coping with stress as predictors of psychological well-being in the course of schizophrenia. the study design," *International journal of environmental research and public health*, vol. 16, no. 7, p. 1266, 2019.

[364] S. E. Taylor and A. L. Stanton, "Coping resources, coping processes, and mental health," *Annu. Rev. Clin. Psychol.*, vol. 3, no. 1, pp. 377–401, 2007.

[365] A. C. Logan, B. M. Berman, and S. L. Prescott, "Vitality revisited: the evolving concept of flourishing and its relevance to personal and public health," *International journal of environmental research and public health*, vol. 20, no. 6, p. 5065, 2023.

[366] J. Hofer, H. Busch, and J. Kärtner, "Self–regulation and well–being: The influence of identity and motives," *European Journal of Personality*, vol. 25, no. 3, pp. 211–224, 2011.

[367] P. G. Reed, "Self-transcendence: Scale and theory," *STS-2018*, 2018.

[368] B. J. Aiena, B. J. Baczwaski, S. E. Schulenberg, and E. M. Buchanan, "Measuring resilience with the rs–14: A tale of two samples," *Journal of Personality Assessment*, vol. 97, no. 3, pp. 291–300, 2015.

[369] J. M. Brown, W. R. Miller, and L. A. Lawendowski, "The self-regulation questionnaire," in *Innovations in Clinical Practice: A Source Book*, L. VandeCreek and T. L. Jackson, Eds. Sarasota, FL: Professional Resource Press, 1999, vol. 17, pp. 281–289.

[370] M. E. Teixeira, "Self-transcendence: A concept analysis for nursing praxis," *Holistic nursing practice*, vol. 22, no. 1, pp. 25–31, 2008.

[371] C. M. Montes, "Evaluating the impact of a yoga-based intervention on software engineers' well-being," Jan. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.14721592

[372] S. Dasanayake, S. Aaramaa, J. Markkula, and M. Oivo, "Impact of requirements volatility on software architecture: How do software teams keep up with ever-changing requirements?" *Journal of software: evolution and process*, vol. 31, no. 6, p. e2160, 2019.

[373] J. Singer, T. Lethbridge, N. Vinson, and N. Anquetil, "An examination of software engineering work practices," in *CASCON First Decade High Impact Papers*, 2010, pp. 174–188.

[374] C. Maslach, W. B. Schaufeli, and M. P. Leiter, "Job burnout," *Annual review of psychology*, vol. 52, no. 1, pp. 397–422, 2001.

[375] D. Graziotin and F. Fagerholm, "Happiness and the productivity of software engineers," in *Rethinking Productivity in Software Engineering*. Springer, 2019, pp. 109–124.

[376] R. S. Kreitchmann, F. J. Abad, V. Ponsoda, M. D. Nieto, and D. Morillo, "Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of likert items," *Frontiers in psychology*, vol. 10, p. 2309, 2019.

[377] N. R. Kuncel and A. Tellegen, "A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development," *Personnel Psychology*, vol. 62, no. 2, pp. 201–228, 2009.

[378] B. Weijters, H. Baumgartner, and N. Schillewaert, "Reversed item bias: an integrative model." *Psychological methods*, vol. 18, no. 3, p. 320, 2013.

[379] D. Grassi, F. Lanubile, A. Motca-Schnabel, and N. Novielli, "A cluster-based approach for emotion recognition in software development," in *Proceedings of the 18th International Conference on Cooperative and Human Aspects of Software Engineering (CHASE 2025)*, 2025, pp. 1–13.

[380] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[381] J. H. Westerink, R. J. Rajae-Joordens, M. Ouwerkerk, M. van Dooren, S. Jelfs, A. J. Denissen, E. Penning de Vries, and R. van Ee, "Deriving a cortisol-related stress indicator from wearable skin conductance measurements: Quantitative model & experimental validation," *Frontiers in Computer Science*, vol. 2, p. 39, 2020.

[382] R. Kocielnik, N. Sidorova, F. M. Maggi, M. Ouwerkerk, and J. H. D. M. Westerink, "Smart technologies for long-term stress monitoring at work," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, 2013, pp. 53–58.

[383] L.-B. Fan, J. A. Blumenthal, L. L. Watkins, and A. Sherwood, "Work and home stress: associations with anxiety and depression symptoms," *Occupational Medicine*, vol. 65, no. 2, pp. 110–116, 01 2015. [Online]. Available: https://doi.org/10.1093/occmed/kqu181

[384] J. Posner, J. Russell, and B. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology." *Dev Psychopathol.*, vol. 17(3), pp. 715–34, 2005.

[385] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[386] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and biological engineering and computing*, vol. 42, pp. 419–427, 2004.

[387] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2015.

[388] S. Koelstra, C. Mühl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. on Affective Comp.*, vol. 3, no. 1, pp. 18–31, 2012.

[389] I. A. Khan, W.-P. Brinkman, and R. M. Hierons, "Do moods affect programmers' debug performance?" *Cogn. Technol. Work*, vol. 13, no. 4, p. 245–258, Nov. 2011. [Online]. Available: https://doi.org/10.1007/s10111-010-0164-1

[390] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[391] R. N. Goodman, J. C. Rietschel, L.-C. Lo, M. E. Costanzo, and B. D. Hatfield, "Stress, emotion regulation and cognitive performance: The predictive contributions of trait and state relative frontal eeg alpha asymmetry," *International journal of psychophysiology*, vol. 87, no. 2, pp. 115–123, 2013.

[392] M. M. Bradley and P. J. Lang, "Measuring emotion: Behavior, feeling, and physiology." 2000.

[393] F. Canento, A. Fred, H. Silva, H. Gamboa, and A. Lourenço, "Multimodal biosignal sensor data handling for emotion recognition," in *SENSORS, 2011 IEEE*, 2011, pp. 647–650.

[394] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, "Acute mental stress assessment via short term hrv analysis in healthy adults: A systematic review with meta-analysis," *Biomedical Signal Processing and Control*, vol. 18, pp. 370–377, 2015.

[395] S. M. U. Saeed, S. M. Anwar, H. Khalid, M. Majid, and U. Bagci, "Eeg based classification of long-term stress using psychological labeling," *Sensors*, vol. 20, no. 7, p. 1886, 2020.

[396] J. Chae, S. Hwang, W. Seo, and Y. Kang, "Relationship between rework of engineering drawing tasks and stress level measured from physiological signals," *Automation in Construction*, vol. 124, p. 103560, 2021.

[397] K. Mohanavelu, S. Poonguzhali, K. Adalarasu, D. Ravi, V. Chinnadurai, S. Vinutha, K. Ramachandran, and S. Jayaraman, "Dynamic cognitive workload assessment for fighter pilots in simulated fighter aircraft environment using eeg," *Biomedical Signal Processing and Control*, vol. 61, p. 102018, 2020.

[398] D. A. Martínez Vásquez, H. F. Posada-Quintero, and D. M. Rivera Pinzón, "Mutual information between eda and eeg in multiple cognitive tasks and sleep deprivation conditions," *Behavioral Sciences*, vol. 13, no. 9, p. 707, 2023.

[399] D. Fucci, D. Girardi, N. Novielli, L. Quaranta, and F. Lanubile, "A replication study on code comprehension and expertise using lightweight biometric sensors," in *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)*, 2019, pp. 311–322.

[400] ——, "A replication study on code comprehension and expertise using lightweight biometric sensors," in *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)*. IEEE, 2019, pp. 311–322.

[401] S. Radevski, H. Hata, and K. Matsumoto, "Real-time Monitoring of Neural State in Assessing and Improving Software Developers' Productivity," in *Proceedings of the Eighth International Workshop on Cooperative and Human Aspects of Software Engineering*. IEEE Press, 2015, pp. 93–96.

[402] M. Züger, C. Corley, A. N. Meyer, B. Li, T. Fritz, D. Shepherd, V. Augustine, P. Francis, N. Kraft, and W. Snipes, "Reducing Interruptions at Work: A Large-scale Field Study of FlowLight," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 61–72.

[403] A. Calcagno, S. Coelli, R. Couceiro, J. Durães, C. Amendola, I. Pirovano, R. Re, and A. M. Bianchi, "Eeg monitoring during software development," in *2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON)*. IEEE, 2020, pp. 325–329.

[404] J. Medeiros, R. Couceiro, G. Duarte, J. Durães, J. Castelhano, C. Duarte, M. Castelo-Branco, H. Madeira, P. De Carvalho, and C. Teixeira, "Can eeg be adopted as a neuroscience reference for assessing software programmers' cognitive load?" *Sensors*, vol. 21, no. 7, p. 2338, 2021.

[405] J. A. Coan and J. J. Allen, *Handbook of emotion elicitation and assessment*. Oxford university press, 2007.

[406] C. Martinez Montes, D. Grassi, N. Novielli, and B. Penzenstadler, "Replication package for stress multimodal study," May 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15497559

[407] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard, "Automatic identification of artifacts in electrodermal activity data," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 1934–1937.

[408] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxeda: A convex optimization approach to electrodermal activity processing," *IEEE transactions on biomedical engineering*, vol. 63, no. 4, pp. 797–804, 2015.

[409] M. Nardelli, A. Greco, L. Sebastiani, and E. P. Scilingo, "Comeda: A new tool for stress assessment based on electrodermal activity," *Computers in Biology and Medicine*, vol. 150, p. 106144, 2022.

[410] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[411] M. Gulati, L. J. Shaw, R. A. Thisted, H. R. Black, C. N. Bairey Merz, and M. F. Arnsdorf, "Heart rate response to exercise stress testing in asymptomatic women: the st. james women take heart project," *Circulation*, vol. 122, no. 2, pp. 130–137, 2010.

[412] M. Malik and A. J. Camm, "Heart rate variability," *Clinical cardiology*, vol. 13, no. 8, pp. 570–576, 1990.

[413] X. Yu, J. Lu, W. Liu, Z. Cheng, and G. Xiao, "Exploring physiological stress response evoked by passive translational acceleration in healthy adults: a pilot study utilizing electrodermal activity and heart rate variability measurements," *Scientific Reports*, vol. 14, no. 1, p. 11349, 2024.

[414] T. Reinhardt, C. Schmahl, S. Wüst, and M. Bohus, "Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the mannheim multicomponent stress test (mmst)," *Psychiatry research*, vol. 198, no. 1, pp. 106–111, 2012.

[415] S. Delliaux, A. Delaforge, J.-C. Deharo, and G. Chaumet, "Mental workload alters heart rate variability, lowering non-linear dynamics," *Frontiers in physiology*, vol. 10, p. 565, 2019.

[416] M. M. Adamson, A. Phillips, S. Seenivasan, J. Martinez, H. Grewal, X. Kang, J. Coetzee, I. Luttenbacher, A. Jester, O. A. Harris *et al.*, "International prevalence and correlates of psychological stress during the global covid-19 pandemic," *International Journal of Environmental Research and Public Health*, vol. 17, no. 24, p. 9248, 2020.

[417] D. H. Hellhammer, S. Wüst, and B. M. Kudielka, "Salivary cortisol as a biomarker in stress research," *Psychoneuroendocrinology*, vol. 34, no. 2, pp. 163–171, 2009.

[418] J. F. Thayer, F. Åhs, M. Fredrikson, J. J. Sollers III, and T. D. Wager, "A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 2, pp. 747–756, 2012.

[419] M. Sandner, G. Lois, F. Streit, P. Zeier, P. Kirsch, S. Wüst, and M. Wessa, "Investigating individual stress reactivity: high hair cortisol predicts lower acute stress responses," *Psychoneuroendocrinology*, vol. 118, p. 104660, 2020.

[420] X. Hou, Y. Liu, O. Sourina, Y. R. E. Tan, L. Wang, and W. Mueller-Wittig, "Eeg based stress monitoring," in *2015 IEEE international conference on systems, man, and cybernetics*. IEEE, 2015, pp. 3110–3115.

[421] S.-H. Seo, J.-T. Lee, and M. Crisan, "Stress and eeg," *Convergence and hybrid information technologies*, vol. 27, 2010.

[422] J. Friedrich, A. Bareis, M. Bross, Z. Bürger, Á. Cortés Rodríguez, N. Effenberger, M. Kleinhansl, F. Kremer, and C. Schröder, ""how is your thesis going?"–ph. d. students' perspectives on mental health and stress in academia," *Plos one*, vol. 18, no. 7, p. e0288103, 2023.

[423] A. W. Gaillard, "Comparing the concepts of mental load and stress," *Ergonomics*, vol. 36, no. 9, pp. 991–1005, 1993.

[424] X. Zhang, Z. Zhao, J. Sun, and J. Ren, "Good stress or bad stress? an empirical study on the impact of time pressure on doctoral students' innovative behavior," *Frontiers in Psychology*, vol. 15, p. 1460037, 2024.

[425] D. Pérez-Jorge, M. Boutaba-Alehyan, A. I. González-Contreras, and I. Pérez-Pérez, "Examining the effects of academic stress on student well-being in higher education," *Humanities and Social Sciences Communications*, vol. 12, no. 1, pp. 1–13, 2025.

[426] M. Csikszentmihalyi, M. Csikszentmihalyi, S. Abuhamdeh, and J. Nakamura, "Flow," *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi*, pp. 227–238, 2014.

[427] C. Peifer, A. Schulz, H. Schächinger, N. Baumann, and C. H. Antoni, "The relation of flow-experience and physiological arousal under stress—can u shape it?" *Journal of Experimental Social Psychology*, vol. 53, pp. 62–69, 2014.

[428] S. O. Ferreira, "Emotional activation in human beings: Procedures for experimental stress induction," *Psicologia USP*, vol. 30, p. e180176, 2019.

[429] K. M. Fahey, S. S. Dermody, and A. Cservenka, "The importance of community engagement in experimental stress and substance use research with marginalized groups: lessons from research with sexual and gender minority populations," *Drug and alcohol dependence*, vol. 260, p. 111349, 2024.

[430] American Psychological Association, "Ethical principles of psychologists and code of conduct," 2017, accessed: 2025-05-16. [Online]. Available: https://www.apa.org/ethics/code/index

[431] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "Quality measures in biometric systems," *IEEE Security & Privacy*, vol. 10, no. 6, pp. 52–62, 2011.

[432] J. Roberts, M. Baker, and J. Andrew, "Artificial intelligence and qualitative research: The promise and perils of large language model (llm)'assistance'," *Critical Perspectives on Accounting*, vol. 99, p. 102722, 2024.

[433] R. H. Tai, L. R. Bentley, X. Xia, J. M. Sitt, S. C. Fankhauser, A. M. Chicas-Mosier, and B. G. Monteith, "An examination of the use of large language models to aid analysis of textual data," *International Journal of Qualitative Methods*, vol. 23, p. 16094069241231168, 2024.

[434] M. Bano, D. Zowghi, and J. Whittle, "Ai and human reasoning: Qualitative research in the age of large language models," *The AI Ethics Journal*, vol. 3, no. 1, 2023.

[435] V. R. Basili and M. V. Zelkowitz, "Empirical studies to build a science of computer science," *Communications of the ACM*, vol. 50, no. 11, pp. 33–37, 2007.

[436] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *IEEE Transactions on software engineering*, vol. 28, no. 8, pp. 721–734, 2002.

[437] D. E. Perry, A. A. Porter, and L. G. Votta, "Empirical studies of software engineering: a roadmap," in *Proceedings of the conference on The future of Software engineering*, 2000, pp. 345–355.

[438] R. Harrison, N. Badoo, E. Barry, S. Biffl, A. Parra, B. Winter, and J. Wuest, "Directions and methodologies for empirical software engineering research," *Empirical Software Engineering*, vol. 4, no. 4, pp. 405–410, 1999.

[439] K.-J. Stol and B. Fitzgerald, "The abc of software engineering research," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 27, no. 3, pp. 1–51, 2018.

[440] B. Meyer, H. Gall, M. Harman, and G. Succi, "Empirical answers to fundamental software engineering problems (panel)," in *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, 2013, pp. 14–18.

[441] T. Dybå, R. Prikladnicki, K. Rönkkö, C. Seaman, and J. Sillito, "Qualitative research in software engineering," *Empirical Software Engineering*, vol. 16, no. 4, pp. 425–429, 2011.

[442] M. Bano, R. Hoda, D. Zowghi, and C. Treude, "Large language models for qualitative research in software engineering: exploring opportunities and challenges," *Automated Software Engineering*, vol. 31, no. 1, p. 8, 2024.

[443] I. Bennis and S. Mouwafaq, "Advancing ai-driven thematic analysis in qualitative research: a comparative study of nine generative models on cutaneous leishmaniasis data," *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, pp. 1–14, 2025.

[444] D. Byrne, "A worked example of braun and clarke's approach to reflexive thematic analysis," *Quality & quantity*, vol. 56, no. 3, pp. 1391–1412, 2022.

[445] S. De Paoli, "Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach," *Social Science Computer Review*, vol. 42, no. 4, pp. 997–1019, 2024.

[446] J. S. Y. Liew, N. McCracken, S. Zhou, and K. Crowston, "Optimizing features in active machine learning for complex qualitative content analysis," in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 2014, pp. 44–48.

[447] K. Crowston, E. E. Allen, and R. Heckman, "Using natural language processing technology for qualitative data analysis," *International Journal of Social Research Methodology*, vol. 15, no. 6, pp. 523–543, 2012.

[448] K. Crowston, X. Liu, and E. E. Allen, "Machine learning and rule-based automated coding of qualitative data," *proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–2, 2010.

[449] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political analysis*, vol. 21, no. 3, pp. 267–297, 2013.

[450] S. C. Lewis, R. Zamith, and A. Hermida, "Content analysis in an era of big data: A hybrid approach to computational and manual methods," *Journal of broadcasting & electronic media*, vol. 57, no. 1, pp. 34–52, 2013.

[451] N.-C. Chen, M. Drouhard, R. Kocielnik, J. Suh, and C. R. Aragon, "Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 2, pp. 1–20, 2018.

[452] L. Towler, P. Bondaronek, T. Papakonstantinou, R. Amlôt, T. Chadborn, B. Ainsworth, and L. Yardley, "Applying machine-learning to rapidly analyze large qualitative text datasets to inform the covid-19 pandemic response: comparing human and machine-assisted topic analysis techniques," *Frontiers in Public Health*, vol. 11, p. 1268223, 2023.

[453] V. Bogachenkova, E. C. Martins, J. Jansen, A.-M. Olteniceanu, B. Henkemans, C. Lavin, L. Nguyen, T. Bradley, V. Fürst, H. M. Muctadir *et al.*, "Lama: a thematic labelling web application," *Journal of Open Source Software*, vol. 8, no. 85, p. 5135, 2023.

[454] S. M. Renz, J. M. Carrington, and T. A. Badger, "Two strategies for qualitative content analysis: An intramethod approach to triangulation," *Qualitative health research*, vol. 28, no. 5, pp. 824–831, 2018.

[455] L. K. Nelson, "Computational grounded theory: A methodological framework," *Sociological methods & research*, vol. 49, no. 1, pp. 3–42, 2020.

[456] R. P. Gauthier and J. R. Wallace, "The computational thematic analysis toolkit," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. GROUP, pp. 1–15, 2022.

[457] T. Qiao, C. Walker, C. Cunningham, and Y. S. Koh, "Thematic-lm: a llm-based multi-agent system for large-scale thematic analysis," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 649–658.

[458] R. Sinha, I. Solola, H. Nguyen, H. Swanson, and L. Lawrence, "The role of generative ai in qualitative research: Gpt-4's contributions to a grounded theory analysis," in *Proceedings of the 2024 Symposium on Learning, Design and Technology*, 2024, pp. 17–25.

[459] L. Yan, V. Echeverria, G. M. Fernandez-Nieto, Y. Jin, Z. Swiecki, L. Zhao, D. Gašević, and R. Martinez-Maldonado, "Human-ai collaboration in thematic analysis using chatgpt: A user study and design recommendations," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–7.

[460] K. Nguyen-Trung, "Chatgpt in thematic analysis: Can ai become a research assistant in qualitative research?" *Quality & Quantity*, pp. 1–34, 2025.

[461] W. Tabone and J. De Winter, "Using chatgpt for human–computer interaction research: a primer," *Royal Society open science*, vol. 10, no. 9, p. 231053, 2023.

[462] J. Drápal, H. Westermann, and J. Savelka, "Using large language models to support thematic analysis in empirical legal studies," in *Legal Knowledge and Information Systems*. IOS Press, 2023, pp. 197–206.

[463] C. Martinez Montes and R. Feldt, "Online appendix of paper: Large language models in thematic analysis: Prompt engineering, evaluation, and guidelines for qualitative software engineering research," Oct. 2025, available at: https://doi.org/10.5281/zenodo.17401526.

[464] S. De Paoli *et al.*, "Further explorations on the use of large language models for thematic analysis. open-ended prompts, better terminologies and thematic maps," in *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, vol. 25, no. 3, 2024.

[465] J. P. Wiebe, R. Khan, S. Burns, and J. D. Slotta, "Qualitative research in the age of llms: A human-in-the-loop approach to hybrid thematic analysis," in *Proceedings of the 19th International Conference of the Learning Sciences-ICLS 2025, pp. 1123-1131*. International Society of the Learning Sciences, 2025.

[466] J. Manning and A. Kunkel, "Making meaning of meaning-making research: Using qualitative research for studies of social and personal relationships," *Journal of social and personal relationships*, vol. 31, no. 4, pp. 433–441, 2014.

[467] Z. Han, A. Tavasi, J. Lee, J. Luzuriaga, K. Suresh, M. Oppenheim, F. Battaglia, and S. R. Terlecky, "Can large language models be used to code text for thematic analysis? an explorative study," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 171, 2025.

[468] P. A. Christou, "How to use artificial intelligence (ai) as a resource, methodological and analysis tool in qualitative research?" *Qualitative Report*, vol. 28, no. 7, 2023.

[469] S. K. Ahmed, "The pillars of trustworthiness in qualitative research," *Journal of Medicine, Surgery, and Public Health*, vol. 2, p. 100051, 2024.

Appendix A

Appendix - Paper A

A.1 Data Collection Instruments

A.2 Survey Instruments

A.2.1 The MAAS instrument

The Mindfulness Attention Awareness Scale (MAAS) is replicated from [140].

Instruction MAAS:

> Below is a collection of statements about your everyday experience. Using the 1 (almost never) - 6 (almost always) scale below, please indicate how frequently or infrequently you currently have each experience. Please answer according to what *really reflects* your experience rather than what you think your experience should be. Please treat each item separately from every other item.

A.2.2 The instruments SPANE, PWB, and PTS

Diener et al. [141] proposed a set of related instruments in 'New measures of well-being' that includes the Scale of Positive And Negative Experience (SPANE), the scale of Psychological Well-being (PWB), and the scale of Positive Thinking (PTS).

Instruction SPANE:

> Please think about what you have been doing and experiencing during the past four weeks. Then report how much you experienced each of the following feelings, using the scale below. For each item, select a number from 1 (Very rarely or never) to 5 (Very often or always).

Instruction PWB:

> Below are 8 statements with which you may agree or disagree. Using the 1 (Strongly disagree) – 7 (Strongly agree) scale below, indicate your agreement with each item by indicating that response for each statement.

	1	2	3	4	5	6
I could be experiencing some emotion and not be conscious of it until some time later.						
I break or spill things because of carelessness, not paying attention, or thinking of something else.						
I find it difficult to stay focused on what's happening in the present.						
I tend to walk quickly to get where I'm going without paying attention to what I experience along the way.						
I tend not to notice feelings of physical tension or discomfort until they really grab my attention.						
I forget a person's name almost as soon as I've been told it for the first time.						
It seems I am "running on automatic," without much awareness of what I'm doing.						
I rush through activities without being really attentive to them.						
I get so focused on the goal I want to achieve that I lose touch with what I'm doing right now to get there.						
I do jobs or tasks automatically, without being aware of what I'm doing.						
I find myself listening to someone with one ear, doing something else at the same time.						
I drive places on 'automatic pilot' and then wonder why I went there.						
I find myself preoccupied with the future or the past.						
I find myself doing things without paying attention.						
I snack without being aware that I'm eating.						

Table A.1: The Mindfulness Attention Awareness Scale (MAAS) [140]

Instruction PTS:

The following items are to be answered "Yes" or "No." Write an answer next to each item to indicate your response.

	1	2	3	4	5
Positive					
Negative					
Good					
Bad					
Pleasant					
Unpleasant					
Happy					
Sad					
Afraid					
Joyful					
Angry					
Contented					

Table A.2: The Scale of Positive and Negative Experiences (SPANE) [141]

	1	2	3	4	5	6	7
I lead a purposeful and meaningful life.							
My social relationships are supportive and rewarding.							
I am engaged and interested in my daily activities							
I actively contribute to the happiness and well-being of others							
I am competent and capable in the activities that are important to me							
I am a good person and live a good life							
I am optimistic about my future							
People respect me							

Table A.3: The Psychological Well-Being (PWB) [141]

A.2.3 Self Efficacy

The instrument was developed by Jerusalem et al. [326] and based on Bandura et al.'s [327] self-efficacy model. It is used to assess the individual stress resilience of the participants and encompasses ten items that offer a positively phrased statement on change, challenges or unexpected circumstances which the participant has to rate as "Not true" (1), "Hardly true" (2), "Rather true" (3) or "Exactly true" (4).

Instruction:

> Please rate the following statements on the basis of the given scale and tick as appropriate:

	Yes	No
I see my community as a place full of problems.		
I see much beauty around me.		
I see the good in most people.		
When I think of myself, I think of many shortcomings.		
I think of myself as a person with many strengths.		
I am optimistic about my future.		
When somebody does something for me, I usually wonder if they have an ulterior motive.		
When something bad happens, I often see a "silver lining," something good in the bad event.		
I sometimes think about how fortunate I have been in life.		
When good things happen, I wonder if they might have been even better.		
I frequently compare myself to others.		
I think frequently about opportunities that I missed.		
When I think of the past, the happy times are most salient to me.		
I savor memories of pleasant past times.		
I regret many things from my past.		
When I see others prosper, even strangers, I am happy for them.		
When I think of the past, for some reason the bad things stand out.		
I know the world has problems, but it seems like a wonderful place anyway.		
When something bad happens, I ruminate on it for a long time.		
When good things happen, I wonder if they will soon turn sour.		
When I see others prosper, it makes me feel bad about myself.		
I believe in the good qualities of other people.		

Table A.4: The Positive Thinking Scale

A.2.4 Perceived Productivity

The HPQ[1] measures perceived productivity in two ways: First, it uses an eight-item scale (summative, multiple reversed indicators), that assesses overall and relative performance, and second, it uses an eleven-point list of general ratings of participants' own performance as well as typical performance of similar workers.

[1] http://www.hcp.med.harvard.edu/hpq

	1	2	3	4
When problems arise, I find ways to carry through.				
I always succeed in solving difficult problems, if I try.				
It does not give me any difficulty to realize my intentions and goals.				
In unexpected situations I always know how to behave.				
Even with surprising events, I believe that I can handle them well.				
I can easily face difficulties because I can always trust my abilities.				
Whatever happens, I'll be fine.				
For every problem I can find a solution.				
When a new thing comes to me, I know how to handle it.				
If a problem arises, I can do it on my own.				

Table A.5: Self efficacy instrument by Jerusalem et al. [326]

Instructions PP:

> The next questions are about the time you spent during your hours at work in the past 4 weeks (28 days). Select the one response for each question that comes closest to your experience from "None of the time" (1) to "All of the time" (5).

- On a scale from 0 to 10 where 0 is the worst job performance anyone could have at your job and 10 is the performance of a top worker, how would you rate the usual performance of most workers in a job similar to yours?
- Using the same 0-to-10 scale, how would you rate your usual job performance over the past year or two?
- Using the same 0-to-10 scale, how would you rate your overall job performance on the days you worked during the past 4 weeks (28 days)?
- How would you compare your overall job performance on the days you worked during the past 4 weeks (28 days) with the performance of most other workers who have a similar type of job?
  - You were a lot better than other workers
  - You were somewhat better than other workers
  - You were a little better than other workers
  - You were about average

	1	2	3	4	5
How often was your performance higher than most workers on your job?					
How often was your performance lower than most workers on your job?					
How often did you do no work at times when you were supposed to be working?					
How often did you find yourself not working as carefully as you should?					
How often was the quality of your work lower than it should have been?					
How often did you not concentrate enough on your work?					
How often did health problems limit the kind or amount of work you could do?					

Table A.6: Perceived Productivity from the HPQ

- You were a little worse than other workers
- You were somewhat worse than other workers
- You were a lot worse than other workers

A.2.5 The WHO-5 instrument

The 5-item World Health Organization Well-Being Index (WHO-5, see Tab. A.7) is a short and generic global rating scale measuring subjective well-being. Because the WHO considers positive well-being to be another term for mental health [328], the WHO-5 only contains positively phrased items, and its use is recommended by [329].

Instruction:

> Please indicate for each of the five statements which is closest to how you have been feeling over the last week from "At no time" (1) to "All of the time" (6). Over the last week:

	1	2	3	4	5	6
I have felt cheerful and in good spirits.						
I have felt calm and relaxed.						
I have felt active and vigorous.						
I woke up feeling fresh and rested.						
My daily life has been filled with things that interest me.						

Table A.7: WHO-5

A.3 Model designs

A.3.1 Gaussian Process model

Below is the model specification for modeling the weekly or daily trends using a Gaussian Process.

$$
\begin{aligned}
\begin{bmatrix} \text{Q1}_i \\ \vdots \\ \text{Q5}_i \end{bmatrix} &\sim \text{Cumulative}\left(\begin{bmatrix} \phi_{\text{Q1},i} \\ \vdots \\ \phi_{\text{Q5},i} \end{bmatrix}, \mathbf{S}\right) && \text{[likelihood]} \\
\text{logit}(\phi_{\text{Q}\{1,\ldots,5\},i}) &= \gamma_{\text{TIME}[i]} + \alpha_{\text{ID}[i]} && \text{[linear model]} \\
\begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix} &\sim \text{MVNormal}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{K}\right) && \text{[prior Gaussian process]} \\
\mathbf{K}_{ij} &= \tau^2 \exp(-T_{ij}^2/2\rho^2) && \text{[covariance matrix } \mathcal{GP}\text{]} \\
\tau &\sim \text{Weibull}(2,1) && \text{[prior std dev } \mathcal{GP}\text{]} \\
\rho &\sim \text{Inv-Gamma}(4,1) && \text{[prior length-scale } \mathcal{GP}\text{]} \\
\mathbf{S} &= \begin{pmatrix} \sigma_{\text{Q1}} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\text{Q2}} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\text{Q3}} & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\text{Q4}} & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\text{Q5}} \end{pmatrix} \mathbf{R} \begin{pmatrix} \sigma_{\text{Q1}} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\text{Q2}} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\text{Q3}} & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\text{Q4}} & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\text{Q5}} \end{pmatrix} && \text{[covariance matrix]} \\
\sigma_{\text{Q1}},\ldots,\sigma_{\text{Q5}} &\sim \text{Weibull}(2,1) && \text{[prior std dev among questions]} \\
\mathbf{R} &\sim \text{LKJ}(2) && \text{[prior correlation matrix]} \\
\alpha_{\text{ID}[i]} &\sim \text{Normal}(\bar{\alpha}, \sigma_{\text{ID}}) && \text{[adaptive prior]} \\
\bar{\alpha} &\sim \text{Normal}(0,2) && \text{[hyperprior avg ID]} \\
\sigma_{\text{ID}} &\sim \text{Weibull}(2,1) && \text{[hyperprior std dev of IDs]}
\end{aligned}
$$

For the weekly trend, on Line 1 we assume a Cumulative likelihood where we model all questions' covariance using a covariance matrix $\mathbf{S}$. The linear model on the next line uses a logit link function as is default, and then models the time, γ, with a Gaussian Process ($\mathcal{GP}$), with a varying intercept α for subjects.

Line 3 places a multivariate normal distribution as prior for the $\mathcal{GP}$, while Lines 4–6 declares a covariance matrix, a prior for the standard deviations, and a prior for the length-scale argument of the $\mathcal{GP}$.

On Line 7 a covariance matrix is declared for $\mathbf{S}$. Then priors for the standard deviations among questions and the correlation matrix $\mathbf{R}$ are declared (Lines 8–9).

Finally, Lines 10–12 declare an adaptive prior for the varying intercept among subjects, and hyperpriors for the average subject (Line 11) and the standard deviation of subjects (final line).

For the daily trend the same model can be used. However, for the daily trend there was only one question asked. This means that the covariance between questions does not need to be modeled and, hence, Lines 7–9 can be removed. Additionally, a suitable prior for the daily data concerning length-scale is Inv-Gamma$(1.6, 0.1)$.

As is evident from the reproducibility package, prior predictive checks were conducted and the combination of priors were uniform on the outcome scale.

A.3.2 Dummy variable regression model

Recall, that for the dummy variable regression models (DVRMs) each instrument (MAAS, SPANE, etc.) was modeled separately with the time (t_0 vs. t_1) used as an indicator (predictor). Four population-level effects (age, gender, occupation, and living conditions) and one group-level effect (subject) were used as predictors.

$$\begin{bmatrix} \text{Q1}_i \\ \vdots \\ \text{Q}n_i \end{bmatrix} \sim \text{Cumulative}\left(\begin{bmatrix} \phi_{\text{Q1},i} \\ \vdots \\ \phi_{\text{Q}n,i} \end{bmatrix}, \mathbf{S}\right) \quad \text{[likelihood]}$$

$$\mathbf{S} = \begin{pmatrix} \sigma_{\text{Q1}} & 0 & 0 \\ & \ddots & \\ 0 & 0 & \sigma_{\text{Q}n} \end{pmatrix} \mathbf{R} \begin{pmatrix} \sigma_{\text{Q1}} & 0 & 0 \\ & \ddots & \\ 0 & 0 & \sigma_{\text{Q}n} \end{pmatrix} \quad \text{[covariance matrix]}$$

$$\sigma_{\text{Q1}}, \ldots, \sigma_{\text{Q}n} \sim \text{Weibull}(2,1) \quad \text{[prior std dev among questions]}$$

$$\mathbf{R} \sim \text{LKJ}(2) \quad \text{[prior correlation matrix]}$$

$$\begin{aligned} \text{logit}(\phi_{\text{Q}\{1,\ldots,n\},i}) &= \alpha \cdot \text{AGE} + \gamma \cdot \text{GENDER} + \omega \cdot \text{OCCUPATION} \\ &\quad + \lambda \cdot \text{LIVING} + \tau \cdot \text{TIME} + \iota_{\text{ID}[i]} \end{aligned} \quad \text{[linear model]}$$

$$\alpha, \gamma, \omega, \lambda, \tau \sim \text{Normal}(0,3) \quad \text{[priors population-level effects]}$$

$$\iota_{\text{ID}[i]} \sim \text{Normal}(\bar{\alpha}, \sigma_{\text{ID}}) \quad \text{[adaptive prior]}$$

$$\bar{\alpha} \sim \text{Normal}(0,2) \quad \text{[hyperprior avg ID]}$$

$$\sigma_{\text{ID}} \sim \text{Weibull}(2,1) \quad \text{[hyperprior std dev of IDs]}$$

For each instrument we assumed a Cumulative likelihood where all questions' covariance was modeled by a covariance matrix $\mathbf{S}$. On Line 2 the covariance matrix is declared for $\mathbf{S}$ and priors for the standard deviations among questions and the correlation matrix $\mathbf{R}$ are declared on Lines 3–4).

The linear model on the next two lines uses a logit link function as is default, and then declares five population-level parameters and a varying intercept ι for subjects. On Line 7 priors for the population-level parameters are declared.

Finally, Lines 8–10 an adaptive prior with hyperpriors is declared for the varying intercept ι.

The only thing that differs between the instruments are the number of questions asked. This implies that the covariance matrix $\mathbf{S}$ differs in size depending on number of questions.

Additionally, for one instrument, SE, there were two questions modeled with a Bernoulli likelihood due to responses on two levels.

As is evident from the reproducibility package, prior predictive checks were conducted and the combination of priors were uniform on the outcome scale.

A.4 Detailed Findings: Significant Effects of Other Predictors

To show that the experiments of run 1 and run 2 confirm the general tendencies, we confirm the underlying latent scale in Fig. A.1. The similar curves with similar centers of the peak show that there is no threat to validity given by the two different lengths of the experiment. In addition, combining the two runs gives the model more certainty, which makes the results more reliable. Had we taken the results of both runs separately, there would be more uncertainty in both individual models, but this was not necessary given the present latency.

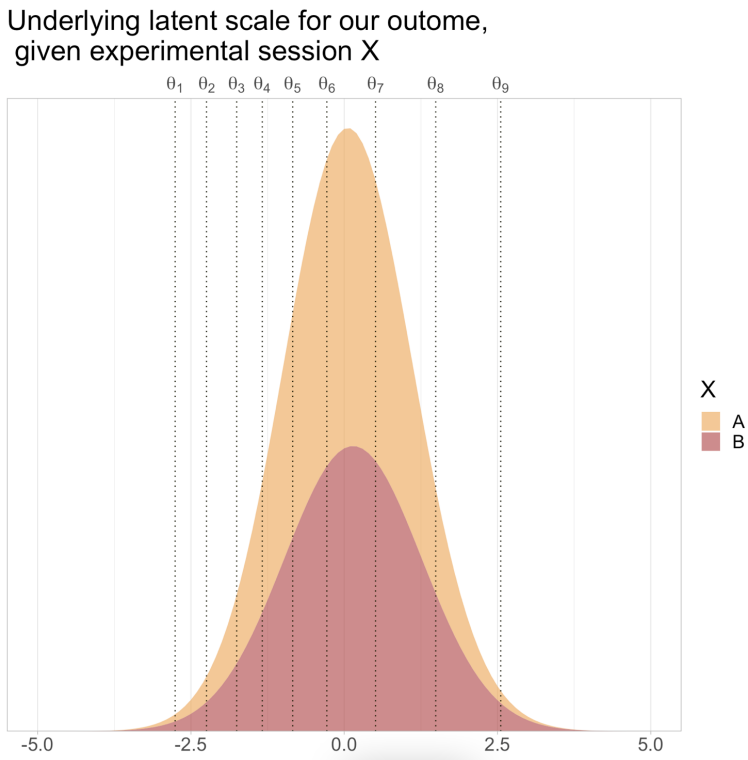


Figure A.1: Underlying latent scale for outcome, given experimental session X

A.4.1 Mindfulness Attention Awareness Scale

The MAAS instrument (App. A.2.1) consisted of 15 statements to agree or disagree with. Eleven of the ratings indicated a significant difference at t_0 vs. t_1: Q1–8, 11–12, and 14. In all the above cases the effect was negative, i.e., the responses were higher at t_0 than at t_1 (please see Fig. 4.6. If we look at the other predictors, age and gender did not have a significant effect, while occupation was significant (negative) for Q2, i.e., "I break or spill things because of carelessness, not paying attention, or thinking

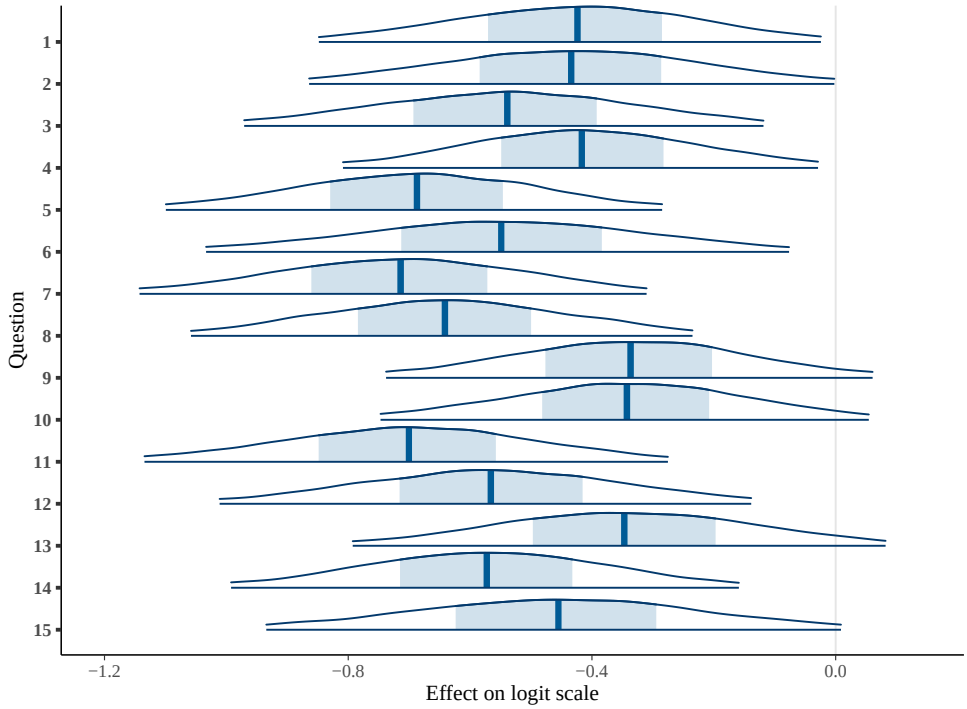


Figure A.2: MAAS Density plots computed from posterior draws. The densities are cut off at 95% and the shaded area is the 50% uncertainty interval. We can see a number of questions not crossing zero (no effect observed).

of something else."

Additionally, the predictor living condition was significant (negative) in Q1–3, 8, and 12 (items listed in App. A.2.1).

A.4.2 Scale of Positive And Negative Experiences

For the SPANE items, see App. A.2.2. The results for the predictor time are in Fig. 4.9.

Below we summarize the significant effects of the other predictors. In all the following tables for predictors, a + means that the item was rated higher for that variable, and a − means that the item was rated lower for that variable.
For gender, a − means that females rated themselves more negatively than males, and a + means that females rated themselves more positively. This is not visible directly from the table below, but requires to know how the data was coded inside the model. For this specific reason, we moved these tables into the appendix, as they are not relevant to understand the narrative of the article, but can be considered interesting observations.

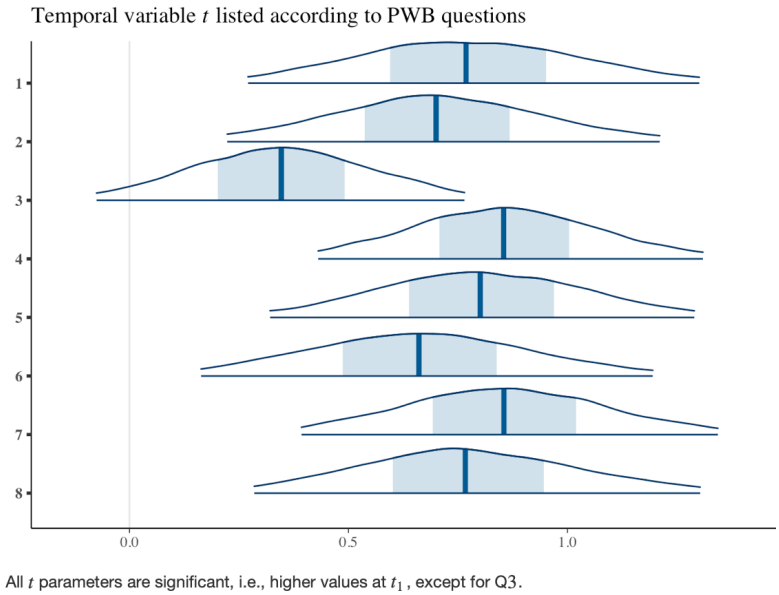


Figure A.3: The effects of t for the PWB instrument. The temporal variable t clearly has an effect (positive) in all questions except Q3.

Question	Age	Gender	Occupation	Living conditions
Q3		−		
Q6		−		
Q7		−		
Q9	+			

In summary for this table, the higher the age, the higher the response in Q9. Concerning gender, males answered with higher values in Q3, Q6, and Q7.

A.4.3 Psychological Well-Being

Figure A.3 shows the effects for the predictor time. The temporal variable t clearly has an effect (positive) in all questions except Q3.

Below we summarize the significant effects of the other predictors for PWB (for the items, see App. A.2.2). The same logic applies here as in the previous table; however, one new effect is present, i.e, occupation. In Q3 (*I am engaged and interested in my daily activities.*), participants with occupation student replied with *higher* responses compared to others.

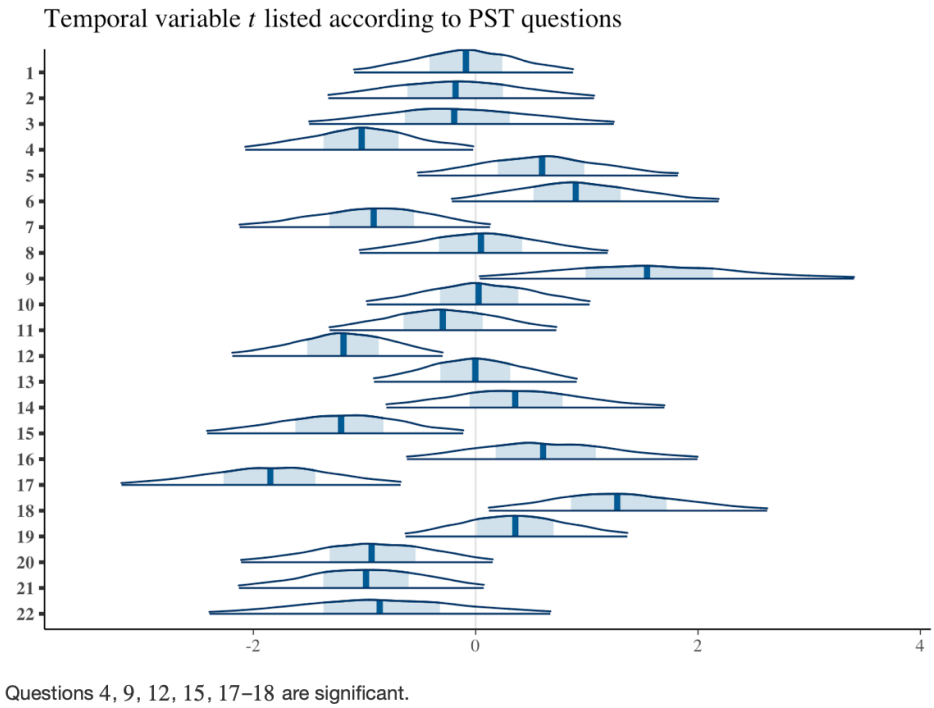


Figure A.4: The PTS results for the predictor time.

Question	Age	Gender	Occupation	Living conditions
Q1				+
Q2	−	−		
Q3	+		−	
Q4		−		
Q7				−

A.4.4 Positive Thinking Scale

For the PTS items, see App. A.2.2. The results for the predictor time are given below in Fig. A.4.

Below we summarize the significant effects of the other predictors. Please refer to the appendix for the respective survey items.

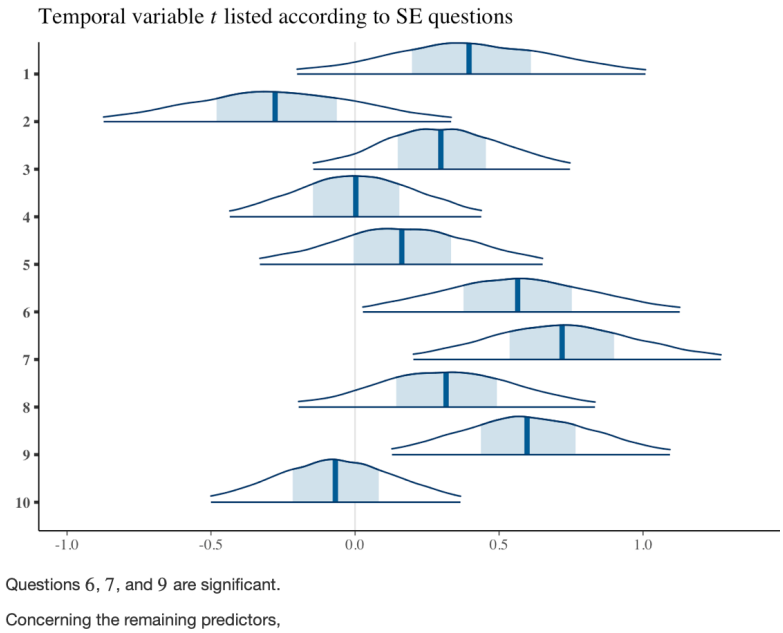


Figure A.5: SE effects for predictor time.

Question	Age	Gender	Occupation	Living conditions
Q1			−	−
Q3			+	
Q11	−	−		
Q16				+
Q17			−	−
Q19				−

A.4.5 Self Efficacy

The SE instrument (App. A.2.3) consisted of ten questions (Likert 1–4). Questions 6, 7, and 9 showed a significant effect (positive), i.e., higher responses at t_1, see Fig. A.5.

Q6 I can easily face difficulties because I can always trust my abilities.

Q7 Whatever happens, I'll be fine.

Q9 When a new thing comes to me, I know how to handle it.

Concerning the other predictors, no significant effects were present, i.e., it is not clear which predictors drove the significant difference between t_0 and t_1.

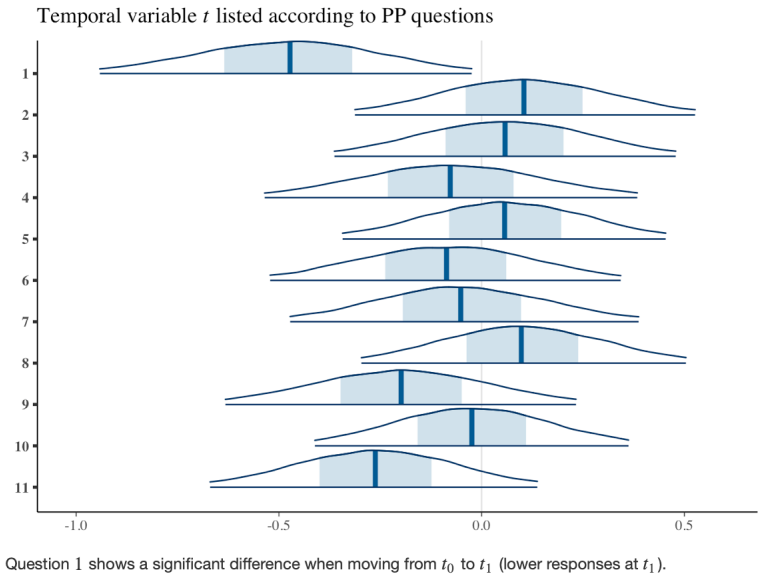


Figure A.6: The PP results for the predictor time.

A.4.6 Perceived Productivity

The HPQ part consisted of eleven questions (with Likert scales varying, going up to 5, 7, or 10, depending on the question, see App. A.2.4). The results for the predictor time are given in Fig. A.6. Only Q1 (*How often was your performance higher than most workers on your job?*) shows a significant difference when moving from t_0 to t_1 (lower responses at t_1).

Below we summarize the significant effects of the other predictors, i.e. Q3 (*How often did you do no work at times when you were supposed to be working?*) showing a higher score for gender female, and Q5 (*How often was the quality of your work lower than it should have been?*) showing a lower score when the living condition was shared with partner or family as opposed to living by oneself.

Question	Age	Gender	Occupation	Living conditions
Q3		+		
Q5				−

A.4.7 Predictor Number of Sessions

The following Table A.8 shows an overview of all significant effects for total number of sessions as predictor. The first column is an ID, the rowname indicates the variable of the instrument, e.g. MAASQ116_total_sessions refers to MAAS question 1 (Likert scale 1 -6) for total sessions attended. The next two columns indicate the estimate and the estimation error. Please note that for SPANE, the results seem to be alternating, but looking back at the instrument (see Sec. A.2.2), half of the items were scored reversely in exactly the pattern that is reflected here.

	rowname	Estimate	Est.Error	Q2.5	Q97.5
1	$MAASQ116_total_sessions$	-0.3155496	0.1216486	-0.5569970	-0.08184895
2	$MAASQ216_total_sessions$	-0.3804576	0.1273390	-0.6411205	-0.13837388
3	$MAASQ316_total_sessions$	-0.2634123	0.1231561	-0.5100075	-0.02042572
5	$MAASQ516_total_sessions$	-0.3689709	0.1167109	-0.6023460	-0.14523413
6	$MAASQ616_total_sessions$	-0.2894895	0.1305140	-0.5477286	-0.03658702
7	$MAASQ716_total_sessions$	-0.3647491	0.1231923	-0.6058283	-0.12399760
8	$MAASQ816_total_sessions$	-0.2611191	0.1214209	-0.5011597	-0.02438610
10	$MAASQ1016_total_sessions$	-0.2886498	0.1174016	-0.5226733	-0.05928175
11	$MAASQ1116_total_sessions$	-0.4540885	0.1211362	-0.6957715	-0.21564968
12	$MAASQ1216_total_sessions$	-0.2509503	0.1246984	-0.4957514	-0.01287479
14	$MAASQ1416_total_sessions$	-0.4358311	0.1179180	-0.6693166	-0.20832100
1	$SPANEQ115_total_sessions$	0.4662730	0.1511756	0.1756023	0.77767102
2	$SPANEQ215_total_sessions$	-0.5187067	0.1341723	-0.7911272	-0.25860345
3	$SPANEQ315_total_sessions$	0.4918396	0.1524530	0.2054288	0.80508272
4	$SPANEQ415_total_sessions$	-0.4509748	0.1308125	-0.7134680	-0.20059290
5	$SPANEQ515_total_sessions$	0.3955807	0.1311677	0.1416188	0.65872865
6	$SPANEQ615_total_sessions$	-0.2643148	0.1243299	-0.5096883	-0.01981721
7	$SPANEQ715_total_sessions$	0.5689896	0.1411704	0.3003263	0.84980113
8	$SPANEQ815_total_sessions$	-0.3191885	0.1221512	-0.5628583	-0.08297879
9	$SPANEQ915_total_sessions$	-0.4594716	0.1445001	-0.7530686	-0.18126877
10	$SPANEQ1015_total_sessions$	0.3753050	0.1239305	0.1374020	0.61918997
11	$SPANEQ1115_total_sessions$	-0.2855759	0.1255116	-0.5319962	-0.04022932
1	$PWBQ117_total_sessions$	0.3232594	0.1505071	0.03253444	0.6231016
2	$PWBQ217_total_sessions$	0.2971393	0.1408516	0.02316987	0.5816784
4	$PWBQ417_total_sessions$	0.3391010	0.1257622	0.09843479	0.5881914
5	$PWBQ517_total_sessions$	0.2659871	0.1345689	0.01074883	0.5332659
6	$PWBQ617_total_sessions$	0.3150326	0.1478417	0.02922867	0.6087226
7	$PWBQ717_total_sessions$	0.3061679	0.1298780	0.05548536	0.5639220
8	$PWBQ817_total_sessions$	0.3378056	0.1378315	0.07209965	0.6095512
9	$PSTQ901_total_sessions$	1.9809234	0.9627086	0.470679200	4.20554675
12	$PSTQ1201_total_sessions$	-0.5350738	0.2776467	-1.103002250	-0.01894271
17	$PSTQ1701_total_sessions$	-0.9643101	0.3736942	-1.751491250	-0.28321947
18	$PSTQ1801_total_sessions$	0.6554668	0.3499538	0.009657123	1.38352275
7	$SEQ714_total_sessions$	0.4327188	0.1503527	0.1492953	0.736231
6	$PPHQ615_total_sessions$	-0.2617817	0.1271908	-0.5070381	-0.01792041

Table A.8: Significant effects for total number of sessions as predictor

The following Table A.9 shows an overview of all significant effects for number of sessions live and recorded as predictor.

	rowname	Estimate Est.	Error	Q2.5	Q97.5
1	$MAASQ116_{l}ive_{s}essions$	-0.2939231	0.1254551	-0.5422444	-0.04608940
3	$MAASQ216_{l}ive_{s}essions$	-0.2663390	0.1292418	-0.5273328	-0.01924983
9	$MAASQ516_{l}ive_{s}essions$	-0.3714334	0.1238394	-0.6162243	-0.13673993
13	$MAASQ716_{l}ive_{s}essions$	-0.3463737	0.1278613	-0.5996803	-0.10126535
19	$MAASQ1016_{l}ive_{s}essions$	-0.2683671	0.1265510	-0.5189259	-0.02028169
21	$MAASQ1116_{l}ive_{s}essions$	-0.3519825	0.1277744	-0.6044776	-0.10675305
22	$MAASQ1116_{r}ecorded_{s}essions$	-0.2729695	0.1309911	-0.5282038	-0.01542422
27	$MAASQ1416_{l}ive_{s}essions$	-0.4692584	0.1242340	-0.7120738	-0.22517830
1	$SPANEQ115_{l}ive_{s}essions$	0.3714696	0.1564535	0.07600304	0.68959905
3	$SPANEQ215_{l}ive_{s}essions$	-0.4590504	0.1376351	-0.73415957	-0.19549265
5	$SPANEQ315_{l}ive_{s}essions$	0.3926880	0.1523264	0.10560065	0.70054970
7	$SPANEQ415_{l}ive_{s}essions$	-0.3858643	0.1326585	-0.65485220	-0.13025333
9	$SPANEQ515_{l}ive_{s}essions$	0.4090131	0.1388661	0.14721700	0.69743993
13	$SPANEQ715_{l}ive_{s}essions$	0.6621751	0.1569447	0.36908735	0.98411335
15	$SPANEQ815_{l}ive_{s}essions$	-0.2616480	0.1268337	-0.51337150	-0.01369651
17	$SPANEQ915_{l}ive_{s}essions$	-0.3424188	0.1480671	-0.63512853	-0.05596512
19	$SPANEQ1015_{l}ive_{s}essions$	0.4365976	0.1324461	0.17729748	0.69863018
7	$PWBQ417_{l}ive_{s}essions$	0.3380077	0.1354551	0.07777354	0.6092814
9	$PWBQ517_{l}ive_{s}essions$	0.2920369	0.1449304	0.01599527	0.5898144
16	$PWBQ817_{r}ecorded_{s}essions$	0.3250622	0.1521206	0.03104731	0.6273903
2	$PSTQ101_{r}ecorded_{s}essions$	-0.8114198	0.4127656	-1.70676300	-0.08103662
8	$PSTQ401_{r}ecorded_{s}essions$	-0.6589501	0.3618867	-1.41424650	-0.01133453
17	$PSTQ901_{l}ive_{s}essions$	3.1336475	1.5388280	0.76178103	6.67567475
22	$PSTQ1101_{r}ecorded_{s}essions$	-1.2048545	0.4597776	-2.20838500	-0.40500393
28	$PSTQ1401_{r}ecorded_{s}essions$	1.5137920	0.9497095	0.02391954	3.70106175
33	$PSTQ1701_{l}ive_{s}essions$	-0.7760812	0.3848592	-1.59662025	-0.07400772
36	$PSTQ1801_{r}ecorded_{s}essions$	1.0777616	0.6296321	0.02372501	2.47252725
44	$PSTQ2201_{r}ecorded_{s}essions$	2.0895201	1.2682835	0.10201318	4.99020650
13	$SEQ714_{l}ive_{s}essions$	0.4491137	0.158366	0.1545877	0.7680788
22	$PPOQ117_{r}ecorded_{s}essions$	-0.2526645	0.1271695	-0.5066494	-0.003687787

Table A.9: Significant effects for number of live and recorded sessions as predictor