

**Informing the design of voice-based technology for
capturing runners' subjective and affective experience
while running**

Submitted by: Tao Bi

Primary Supervisor: Prof. Nadia Berthouze

Co-Supervisors: Prof. Enrico Costanza, Dr. Aneesha Singh

Internal Examiner: Professor Yvonne Rogers

External Examiner: Professor Effie Lai-Chong Law

A PhD Thesis submitted for the Degree of Doctor of Philosophy

UCL Interaction Centre

Division of Psychology and Language Sciences

University College London

2023

Declaration

I, Tao Bi, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Despite significant advances in running technologies for objective running metrics (e.g., pace, heart rate, cadence), current technology is unable to fully capture the subjective and affective running experience (SARE) such as perceived exertion, pain, mood and their complex and dynamic interplay. In addition, there has been limited research on how SARE can be effectively collected, considering its complexity and dynamics in naturalistic situations. This thesis aims to address this gap through four contributions.

Firstly, we present an in-depth understanding of what components of the running experience matter to runners and their supporters, and why. This is addressed through qualitative approaches in the context of long-distance running events. Secondly, we present findings on using ubiquitous ESM based on voice and wearables to gather in-situ self-reports and running data for creating automatic recognition systems for subjective and affective running experience (SARE-ARS). The mixed-method analysis shows the feasibility of such an approach to SARE self-report gathering and suggests design implications for creating a chatbot for interactivity, companionship, and self-reflection while running. Thirdly, we demonstrate the design of a chatbot's conversation flow to facilitate voice-based SARE self-reports while running. Based on the results from user studies with the chatbot, we propose the novel Modality-Autonomy-Prompting (MAP) framework for designing chatbots for SARE self-reports. Finally, we contribute to the understanding of the dynamic and reflective elements of SARE self-reports in two steps: 1) by evaluating the effect of using a MAP-based chatbot on runners' reflective self-reporting processes; and 2) by proposing the Description-Appraisal-Management (DAM) framework that goes beyond just simple affect labels. This framework helps unpack the experiential and reflective components of SARE and guide the chatbot design to support runners to continuously self-reflect and self-report their SARE while running. The thesis concludes by discussing some ethical considerations as well as opportunities for the use of such frameworks.

UCL Research Paper Declaration Form

1. For a research manuscript that has already been published

a) **What is the title of the manuscript?**

Understanding the shared experience of runners and spectators in long-distance running events

b) **Please include a link to or doi for the work**

<https://doi.org/10.1145/3290605.3300691>

c) **Where was the work published?**

CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems

d) **Who published the work?** (e.g. OUP)

ACM

e) **When was the work published?**

2019 02 May

f) **List the manuscript's authors in the order they appear on the publication**

Tao Bi, Nadia Bianchi-Berthouze, Aneesha Singh, Enrico Costanza

g) **Was the work peer reviewed?**

Yes

h) **Have you retained the copyright?**

Yes

i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

<https://discovery.ucl.ac.uk/id/eprint/10065679/>

Publications

Papers in Preparation

Bi, Tao, Nadia Bianchi-Berthouze, Aneesha Singh, and Enrico Costanza. *“Designing a Voice-based ESM to Collect Runners’ Self-reports for Training Automatic Affect Recognition Systems.”* In preparation for submitting to the 2024CHI Conference on Human Factors in Computing Systems (Related Chapters: 5&6)

Bi, Tao, Nadia Bianchi-Berthouze, Aneesha Singh, and Enrico Costanza. *“A Chatbot for Runners’ Experience Ground Truth Gathering.”* In preparation for TOCHI. (Related Chapters: 6, 7,8,9)

Conference Paper

Bi, Tao, Nadia Bianchi-Berthouze, Aneesha Singh, and Enrico Costanza. *“Understanding the shared experience of runners and spectators in long-distance running events.”* In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-13. 2019. (Related Chapters: 4)

Workshop Papers

Bi, Tao, Raffaele Andrea Buono, Temitayo Olugbade, Aneesha Singh, Catherine Holloway, Enrico Costanza, Amanda C de C Williams, Nicolas E. Gold, and Nadia Berthouze. *“Towards Chatbot-Supported Self-Reporting for Increased Reliability and Richness of Ground Truth for Automatic Pain Recognition: Reflections on Long-Distance Runners and People with Chronic Pain.”* In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pp. 43-53. 2021. (Related Chapters: 6)

Bi, T., Olugbade, T., Mathur, A., Holloway, C., Singh, A., & Costanza, E. (2022, September). A Taxonomy of Noise in Voice Self-reports while Running. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and*

Ubiquitous Computing (UbiComp/ISWC'22 Adjunct), September 11--15, 2022, Cambridge, United Kingdom. ACM: Association for Computing Machinery. (Related Chapters: Appendix G)

Doctoral Consortium Paper

Bi, Tao. *“Wearable Sensing Technology for Capturing and Sharing Emotional Experience of Running.”* In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 70-74. IEEE, 2019. (Related Chapters: 5)

Impact Statement

The thesis proposes a unique approach that involves using conversational agents to gather real-time self-reports of SARE (Subjective and Affective Running Experience) from runners, resulting in a more thorough and ecologically valid data collection of their affective experiences during their daily routines. To do so, it proposes two frameworks called MAP and DAM that inform the design of a chatbot conversation flow to support in-depth runners' self-reflection and self-report. This approach has significant impacts in several fields.

Firstly, in the HCI community, it suggests a novel approach to structuring chatbot conversations for self-reflection, providing users with an exploratory space to think and speak iteratively, followed by structured interaction to facilitate further introspection. Such space was reported by runners to be useful for their own running experience and self-understanding.

Secondly, in the sports psychology and training community, it introduces a pervasive approach to assessing SARE that captures the dynamic and multifaceted nature of affective experiences in real-time, as opposed to traditional laboratory or retrospective evaluations. Finally, in the affective computing community, the thesis proposes to reconsider what an automatic affect recognition system should recognise: moving away from the recognition of simple affective states to a more comprehensive understanding of the experience (an in-depth specification of a state of SARE, the processes that underlie it and the action the runner aims to take to act upon it) as gathered through the in-depth self-reflection and self-reported gathered through chatbot conversations structured according to the two frameworks.

The research presented in this thesis has practical implications beyond academia. The proposed chatbot design can serve as a useful journaling tool for sports coaching and training, enabling athletes to record their subjective and affective running experiences through pervasive and conversational means. Additionally, for the sports

industry, the findings of this thesis highlight what aspects of runners' experience could be tracked, going beyond simple performance metrics to help both runners and their supporters. It also provides a new framework for the development of systems that can capture such experiences. The proposed framework for chatbot could be in future also explored as a novel method for rich ground truth data collection tools for training AI systems continuously in real-time and for lifelong purposes, contributing to the development of a personal AI assistant. The proposed design can also be extended to facilitate self-reflection and journaling in other contexts, such as mental health and sustainable personal behaviours; however, possible emotional impact on the person should be considered and carefully addressed to avoid negative and distressing effects on people.

Acknowledgement

I want to thank everyone who supported me throughout my PhD journey.

Working together with my supervisor, Nadia Berthouze, has been an incredible learning experience, as she has taught me the importance of being mindful about why I do research and for what purpose, a lifelong habit of being rational and analytical in my decision-making process. Her passion for research has been contagious. I cannot thank her enough for her kindness, time, energy, and dedication to my academic and personal growth. I would also like to express my gratitude to my other supervisors, Enrico Costanza and Aneesha Singh, who have supported me with constructive feedback and emotional support as well as constant encouragement whenever I needed it. Their guidance, wisdom, and patience have been instrumental in helping me not only learn how to do research but also become a better person. I extend my appreciation to Cathy Holloway for her valuable research feedback and scholarship support. I am forever grateful for their contribution to my PhD journey.

My PhD was partially funded by a studentship deriving from a grant awarded by the London Legacy Development Corporation (C02955) to the Global Disability Innovation Hub, University College London. The research activities were also partially supported by the EU Future and Emerging Technologies Proactive Programme H2020 (Grant No. 824160: EnTimeMent -<https://entiment.dibris.unige.it/>). The eSense earbuds used in this work were supported by Nokia Bell Lab, Cambridge, UK. I am grateful to my research participants for their invaluable contribution to my study.

I would like to express my heartfelt gratitude to my parents and family for their numerous forms of support and love throughout my life. Without their love and support, I would not have been able to pursue my dreams and achieve my goals. I would like to extend a special thank you to my grandmother for being the first person to encourage me to pursue a PhD when I was just a child. Although she did not wait

to see me begin my PhD journey, I now want to let her know that I am determined to finish it strong and well.

To my friends at UCLIC, Amid Ayobi, Albert Higgins, Akhil Mathurs, Chongyang Wang, Temitayo Olugbade, Dilisha Patel, Maryam Bandukda, Maura Bellio, Muna Alebri, Roxana Ramirez Herrera, Martin Dechant, Julia Manning, Louise Gayer, and many others who were at UCLIC but have left, thank you for your companionship and support throughout my PhD journey. I also thank my friends in Newquay, particularly John Knox and Gustav Sennton, for their technical advice in the development and programming aspects of my thesis. I am grateful to those from my close neighbourhood and those we met through surfing in the Ocean for their companionship. I particularly appreciate the Newquay Road Runners Club, especially our running coach Mr Paul Wright, for their support in my research and personal running training.

Lastly, I would like to express my gratitude to Running, who has been a constant source of inspiration, motivation, companionship, and support for me throughout my PhD journey.

Glossary

Abbreviation	Term	Definitions
ARS	Automatic Recognition Systems	Systems that use algorithms and machine learning to automatically recognize and interpret various targeted classes or numerical values.
DAM	Description-Appraisal-Management	The Description-Appraisal-Management (DAM) Framework derived from this thesis' findings
ESM	Experience Sampling Method	A research method that collects real-time data about individuals' experiences in their natural environment by asking them to report on their thoughts, feelings, and behaviours at multiple points throughout the day using electronic devices (e.g., mobile phone).
LDRE	Long Distance Running Events	Organized running events where runners cover long distances, ranging from 5KM to a marathon distance. They can be competitive or non-competitive events that involve both runners and spectators/supporters.
MAP	Modality-Autonomy-Prompting	The Modality-Autonomy-Prompting (MAP) Framework derived from this thesis' findings.
ML	Machine Learning	A subfield of artificial intelligence that involves the development of algorithms and models that enable machines to automatically learn patterns and relationships from data and make predictions.
R-ESM	A voice- and wearable-based Runner ESM System	A system that implements the ESM to collect real-time data on runners' experiences, including their thoughts, feelings, and running behaviours, during long-distance runs. The system is operated through voice commands and is supported by wearable devices, such as smartwatches, smart insoles, headset.
SARE	Subjective and Affective	The subjective and affective aspects of running experience a long-distance run. This thesis

	Running Experience	particularly focuses on the experienced related to four states, i.e., perceived exertion, perceived pain and discomfort, perceived desire to stop, and emotional valence.
SARE-ARS	SARE Automatic Recognition Systems	A specific type of ARS designed for recognising the SARE.
N/A	Think aloud	(A verb phrase) A research method that involves the action of verbalizing one's thoughts as they occur.
N/A	Think-alouds	(A noun phrase) The data instances resulted from using the Think aloud technique.

Table of Contents

Declaration	2
Abstract	3
UCL Research Paper Declaration Form.....	4
Publications	5
Impact Statement.....	7
Acknowledgement.....	9
Glossary	11
Table of Contents	13
Table of Figures	18
Table of Tables	22
Chapter 1 Introduction.....	26
1.1 Motivation	26
1.2 Research Questions	27
1.3 Contributions.....	28
1.4 Thesis structure	29
Chapter 2 Related Work.....	32
2.1 Subjective and affective running experience (SARE).....	32

2.2	Automatically capturing runners' subjective and affective running experience	52
2.3	Voice-based Chatbots for collecting self-reports	68
Chapter 3	Methodology	79
3.1	Research Question and Methods	79
3.2	Participants	83
3.3	Ethics	83
3.4	Analysis Methods	84
Chapter 4	Understanding the need for capturing Subjective and Affective Running Experience (SARE)	87
4.1	Introduction	87
4.2	Method	87
4.3	Results	90
4.4	Discussions	102
4.5	Summary	107
Chapter 5	Developing R-ESM to collect data for building SARE-ARS	108
5.1	Introduction	108
5.2	Prototyping a Voice & Wearable-based Runner ESM System (R-ESM) for Data Collection	109
Chapter 6	Understanding the efficacy of R-ESM: a user study	129

6.1	Introduction.....	129
6.2	Method.....	129
6.3	Results from Self-reports and Think-alouds data analysis.....	135
6.4	Results of constructing SARE-ARS.....	146
6.5	Results of Qualitative Interview	150
6.6	Discussion	157
6.7	Summary	165
Chapter 7 Developing a voice-based Chatbot for SARE self-reporting		167
7.1	Introduction.....	167
7.2	Iteration 1: Proposing Three Initial Design of Chatbot’s Dialogue	167
7.3	Iteration 2: Refining Chatbot’s Dialogues.....	172
7.4	Iteration 3: Implementing the Chatbot	188
Chapter 8 Understanding runners' experience of using Chatbot for SARE self-reflecting and reporting		196
8.1	Introduction.....	196
8.2	Methods	196
8.3	Quantitative Results: Chatbot and Conversation Quality.....	201
8.4	Qualitative Results.....	210
8.5	Discussion	229

8.6	Summary	239
Chapter 9 Understanding the runners' self-reported SARE		241
9.1	Evaluating the impact of a mixed-structure chatbot on SARE self-reporting 242	
9.2	Understanding the nature of runners' exploratory and unguided SARE think-alouds.....	252
9.3	Discussion	260
9.4	Summary	266
Chapter 10 General discussion and conclusions		267
	Summary of Research Questions, Methods and	267
10.1	Discussion of Findings and Contributions.....	267
10.2	Ethical considerations.....	274
10.3	Limitations and future directions	276
10.4	Conclusion	277
Bibliography		279
Chapter 11 Appendices		300
11.1	Appendix A-Chapter 3	300
11.2	Appendix B-Chapter 4.....	309
11.3	Appendix C-Chapter 5.....	310

11.4	Appendix E-Chapter 7	316
11.5	Appendix F-Chapter 8	327
11.6	Appendix G-Chapter 9	330

Table of Figures

Figure 2-1: Left - Borg RPE (6-20) [4, 41]; Right – adapted RPE (1-10)	33
Figure 2-2: Attentional focus and in Sports (a replicated diagram based on understanding from [190, 191])	36
Figure 2-3: Coping strategies ([179], Morgan and Pollock [183]).....	38
Figure 2-4 Ultra-marathoners’ Mood changes (an image from [144]).....	40
Figure 2-5: three-stage procedure for inducing and collecting reports of fatigue	59
Figure 3-1: Overview research rationale	79
Figure 4-1. DyPECS framework: the outer-layer (OL) Runner and Spectator journey with their 4-Cs and the inner-layer (IL) micro-encounter stages	91
Figure 4-2. Runner’s Experience Map (left), Spectator’s Experience Map (right)	92
Figure 5-1: Wearable & Voice-based Runner-ESM (R-ESM) system.....	110
Figure 5-2 Stridalyzer Insight Smart Insole Sensor (Left: picture of the real product; Right: picture of the concept product).....	112
Figure 5-3: visualization of foot anatomy.....	112
Figure 5-4: Box Chart of P1’s Ground Contact Time (GCT) on the Left Foot under Four Acted Emotional Conditions: Feeling Tired, Painful, Enjoyable and Motivated (Labelled as Tired_running, Painful_running, Enjoyable_running and Motivational_running in the Figure)	116
Figure 5-5: Box Chart of P2’s Ground Contact Time (GCT) on the Left Foot under Four Acted Emotional Conditions: Feeling Tired, Painful, Enjoyable and Motivated	

(Labelled as Tired_running, Painful_running, Enjoyable_running and Motivational_running respectively)	118
Figure 5-6: Box Chart of P3's Ground Contact Time (GCT) on the Left Foot under Four Acted Emotional Conditions: Feeling Tired, Painful, Enjoyable and Motivated (Labelled as Tired_running, Painful_running, Enjoyable_running and Motivational_running respectively)	119
Figure 5-7: Empatica E4 Bracelet [1]	121
Figure 5-8: Overview of eSense [127]	121
Figure 5-9: R-ESM Mobile App Architecture	126
Figure 5-10: Screenshots of the Android based R-ESM App.....	127
Figure 6-1 Study Protocol	132
Figure 6-2 Fluctuating Dynamics of SARE (using P4's data as a demonstration).....	136
Figure 6-3 Self-reports distribution of each participant: for the four variables exertion, pain, desire-to-stop, and emotion (as indicated in the respective legend) for 10 participants. In each graph, the X-axis represents participant ID, while the Y-axis indicates corresponding self-report values. Box plots displaying median levels with boxes representing the first quartile, median, and third quartile. Outlier dots represent minor outliers. The upper whisker extends the top of the box to the maximum value in the dataset while the lower whisker indicates minimum values in the dataset within 1.5 times the interquartile range (IQR)	138
Figure 6-4 Overall Self-reports distribution:.....	140
Figure 6-5 Normalized self-report values for exertion, pain, desire-to-stop, and emotion, ranging from 0 to 1 (as represented on the X-axis), to account for individual	

variability. The Y-axes represents the counts of instances and the x-axes the normalized ranges of self-report.....141

Figure 6-6: Correlation Heatmap: Exertion, Pain, Desire-to-stop, and Emotion (Spearman Method); The correlation coefficient value ranges from -1 to +1, where -1 represents a perfect negative correlation, +1 represents a perfect positive correlation, and 0 represents no correlation. Colour gradients in the heatmap indicate the strength and direction of the monotonic relationship between two variables, with darker colours indicating positive correlations and lighter colours indicating negative correlations. 143

Figure 6-7 Frequency of think-alouds for each participant (X-axis = Participant ID; Y-axis = Count of Labels).....144

Figure 6-8 Ground Truth Labels Distribution.....146

Figure 7-1: Overview of Final Conversation Flow.....176

Figure 7-2 Five components in the Chatbot prototype's architecture 189

Figure 7-3: Drop-down menu (left) and navigation button (right).....195

Figure 7-4: Using cardboard to disable accidental touch. 195

Figure 8-1: Overview of Conversation Quality 209

Figure 9-1: Overview of the Mixed-Structure Chatbot (Blue indicates the Focused Thinking aloud; Red indicates the Directed Chatbot Conversations).....243

Figure 9-2: The Description-Appraisal-Management (DAM) Framework about how Runners Naturally Reflect and Report Subjective Affective States 262

Figure 9-3: Example of in-depth structured Chatbot conversation questions for the topics of sensation, discomfort, or pain.....263

Figure 9-4: Example of Example of in-depth structured Chatbot conversation for topics of Emotion, Affect, Motivation, Psychological State, etc.....	264
Figure 11-1: P1's Hallux's Pressure Changes under Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series).....	313
Figure 11-2: Line Chart of OtherToes' Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series) ..	313
Figure 11-3: Line Chart of Front's Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series) ..	314
Figure 11-4: Line Chart of Front_2's Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series) ..	314
Figure 11-5: Line Chart of Arch's Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series) ..	315
Figure 11-6: Line Chart of Mid's Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series)	315
Figure 11-7: Line Chart of Time Series of Heel2's Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series)	316
Figure 11-8: General Chatbot Conversation Flowchart	330
Figure 11-9: An overview of a runner's voice self-report while running.....	336
Figure 11-10: An excerpt of a runner's voice self-report	337

Table of Tables

Table 2-1 Recent work on using runners' self-reports as ground truth for building SARE-ARS	56
Table 2-2 Correlation between subjective fatigue and objective running features ..	66
Table 2-3: Voice-based Chatbot for Collecting Survey Self-reports	72
Table 5-1 Participants Information for Initial Insole Exploration	113
Table 5-2 Prompts for the four SARE components.....	123
Table 5-3: Think-aloud Task for Training [222]	125
Table 5-4 Data Type Captured in the R-ESM	128
Table 6-1: Participants' Profile (Indoor Treadmill (IT), Outdoor Track (OT), Outdoor Park (OP), M: Male, F: Female).....	130
Table 6-2 Correlation between each type of SARE (using Spearman correlation coefficient).	143
Table 6-3 Think-aloud statistics.....	145
Table 6-4 Classification results for the 4 SARE states (Number in brackets after the F1 scores indicate the number of instances for that level)	149
Table 7-1: Chatbot's Dialogue for Perceived Exertion, Desire-to-Stop and Emotion	169
Table 7-2 Chatbot's Dialogue for Perceived Pain, Desire to Stop and Emotion	170
Table 7-3 Initial Chatbot's Dialogue for Mindfulness and Awareness of SARE	172

Table 7-4 Overview of Seven Types of Conversation	175
Table 7-5 Chatbot welcome users and asks users' running plans and goals.....	177
Table 7-6 Chatbot proactively invites users to chat	178
Table 7-7: Chatbot asks user's perceived exertion	178
Table 7-8 Chatbot asks user's affective state.....	179
Table 7-9 Chatbot provides personalized and empathic responses based on users' responses regarding exertion, emotion and motivation.....	180
Table 7-10 Chatbot asks the user's perceived pain or discomfort	182
Table 7-11 Chatbot provides personalized and empathic responses based on users' responses regarding pain, emotion and motivation	183
Table 7-12 Chatbot's mindful scanning guidance	186
Table 8-1 Participants Profile (Indoor Treadmill (IT), Outdoor Track (OT), Outdoor Park (OP), M: Male, F: Female).....	197
Table 8-2: Summary of Fallback Conversation Interactions occurred. Green = <i>"FallbackSuccessPertinent"</i> ; Blue = <i>"FallbackSuccessNotPertinent"</i> i.e., participant's response correctly understood but not pertinent to the chatbot's question; Red = <i>"Fallback Interactions"</i> i.e. response not recognized triggers further fallback response. Yellow = <i>"FallbackFailure"</i> i.e., three subsequent fallback response attempts lead chatbot ending the conversation.	206
Table 8-3: Analysis of Conversation's Fallback Rate, Conversation Smoothness, Conversation Pertinent, and participants' self-reporting Task Engagement.	208
Table 8-4 Verbalising facilitates mindful reflection: sub-themes overview	210

Table 8-5 :Four types of prompts facilitate self-reflection: sub-themes overview .	215
Table 8-6: Self-reporting to a chatbot as a socially collaborative and empathetic experience: sub-themes overview	218
Table 8-7: The role of unstructured vs structured conversation during self-reporting: sub-themes overview	220
Table 8-8: Factors or Barriers to self-reporting: sub-themes overview	224
Table 8-9: The Modality-Autonomy-Prompting (MAP) Framework: a guidance to design a chatbot to facilitate voice-based SARE self-reports while running	230
Table 8-10: Conceptual Extension to Chatbot’s structured & unstructured conversation.....	233
Table 9-1: Participants' Profile (Indoor Treadmill (IT), Outdoor Track (OT), Outdoor Park (OP), M: Male, F: Female).....	244
Table 9-2: Examples of P2’s Conversation Data	245
Table 9-3: Example structured conversation for self-report personalisation (P # refers to Participants ID).....	251
Table 9-4: Temporality	255
Table 9-5: Examples of Runners’ intentions	257
Table 9-6: Example of Runners’ Actions	259
Table 11-1: Demonstration of the conversation flow following the architecture...	316
Table 11-2 Documentation for Chatbot’s Intent, Entity, Context, and Fulfilment Implementation.....	321

Table 11-3: A Selection of Example Phrases for Chatbot Training	326
Table 11-4: Taxonomy of noise in runners' voice self-reports	334
Table 11-5: Taxonomy of noise in runners' voice self-reports (cont.).....	335

Chapter 1 Introduction

1.1 Motivation

Long-distance running, also known as endurance running (above 3 km), is an affordable and easily accessible physical activity that can improve fitness and health. In recent years, technological advancements such as Garmin¹ and Apple Watch² have enabled high-quality tracking of objective running metrics such as GPS (distance, pace, location), heart rate, cadence, and running posture, as well as providing virtual coaching for fitness and performance improvement. These technologies have also grown interests in the creation of social and affective experiences for runners, such as allowing for connecting with other runners, giving and receiving kudos (e.g., Strava³), and providing motivation via activity goal tracking [e.g., Fitbit⁴]. However, these designs still primarily rely on displaying and sharing objective running metrics and a simplistic understanding of what support runners need.

Through a review of sports psychology literature, this thesis has found that the subjective and affective running experience (SARE), such as runners' perceived exertion, pain, mood, emotional state and the interplay among them, is a crucial element in runners' experiences. However, they are not yet directly able to be sensed by current technology.

Recent research in the field of affective computing suggests that it is possible to develop automatic recognition systems (ARS) that can directly sense cues related to and possibly infer a person's subjective and emotional state. This is obtained by using

¹ <https://www.garmin.com>

² <https://www.apple.com/uk/shop/buy-watch/apple-watch>

³ <https://www.strava.com/>

⁴ <https://www.fitbit.com/>

machine learning (ML) models to build a mapping between people's self-reports of their emotional state, also known as ground truth (GT), and objective signals like physiological and movement data tracked by sensors [131, 200, 201]. This suggests the potential for the development of SARE automatic recognition systems (SARE-ARS).

After conducting a thorough literature review, this thesis has uncovered that the majority of studies examining and collecting self-report data about runners' SARE use a simplistic understanding of what an affective experience is and limited methods, such as Likert scale-based questions in a laboratory setting [46] to capture it. Such a simple understanding limits its uses in operationalising support. However, the sports psychology literature has highlighted that SARE is multifaceted and complex and can fluctuate in real-world scenarios [23, 29, 204, 217]. As a result, there is a crucial need to develop a more nuanced understanding of what runners' real-world and in-the-moment SARE are, their way of processing them and how self-reflection can be supported and recorded.

To address this gap, this thesis adopts a human-centred design approach from the field of Human-Computer Interaction [218], exploring the potential of voice-based experience sampling methods (ESM)[102, 264] and chatbot-based conversational interactions to gather in-depth and more actionable upon SARE self-reports. These approaches aim to capture a more ubiquitous and continuous stream of runners' SARE self-reports, taking into account the multifaceted and dynamic nature of SARE in real-world scenarios. Such understanding will contribute to the development of more in depth and effective SARE-ARS.

1.2 Research Questions

Thus, the main research question in this thesis is:

RQ: How can technology be designed to support runners in self-reporting their subjective and affective running experience (SARE) while running?

Four sub-research questions have been formulated to address the main research question, including

RQ1: What are the components of SARE that matter when sharing, and why?

RQ2: How can a ubiquitous ESM based on voice and wearables be designed to gather in-situ self-reports and running data for the purpose of creating SARE-ARS?

RQ 3: How can chatbot conversation flow be designed to facilitate voice-based SARE self-reports while running?

RQ 4: What are the dynamic, reflective elements of SARE self-reports, and what should SARE-ARS model?

1.3 Contributions

The novel approach presented in this thesis utilises conversational agents to collect SARE self-reports in real-time, providing a more naturalistic, richer and ecologically valid assessment of runners' affective experiences throughout their daily routines. More specifically, this contribution is three-fold:

Firstly, the study provided a deep understanding of the complex and dynamic nature of runners' subjective and affective running experiences (SARE) during their runs. The findings highlighted the importance of considering multiple factors when designing technology that supports and motivates runners.

Secondly, it contributed a new method using chatbots for supporting structured SARE self-reflection and self-report. Two new frameworks, Modality-Autonomy-Prompting (MAP) and Description-Appraisal-Management (DAM) are proposed to guide the design of such chatbots. The frameworks inform conversation structures and prompt types to facilitate a fine-grained and reflective self-reporting process during running while providing value to runners' experience.

The findings also contribute to sports psychology literature by introducing a ubiquitous approach to assess runners' subjective and affective running experience (SARE), which captures the dynamic and multifaceted nature of affective experiences in real-time, in contrast to traditional laboratory or retrospective assessments.

Furthermore, the proposed frameworks extend the design of chatbots beyond closed-ended and open-ended questions, emphasising the usefulness of a mixed-structured conversation approach in the context of sport. This approach allows users to explore, iterate, and generate thoughts in an unstructured space while also providing structured prompts and chatbot-led conversations to facilitate focused and deeper reflection.

1.4 Thesis structure

This thesis is structured as follows:

Related work: Chapter 2 presents related work to this thesis. It provides a detailed understanding of the nature of SARE, drawing from research in sports psychology and human-computer interaction (HCI). The chapter then delves into the literature on affective computing and computer science to understand how SARE-ARS can be constructed. Finally, the chapter explores the HCI approaches to understand how to collect SARE self-reports.

Methodology: Chapter 3 explains the four sub-research questions following the learning from the literature. It also introduces the general methods used to address each question.

Studies for RQ1: Chapter 4 presents a secondary analysis of pre-collected qualitative data and an analysis of newly collected qualitative data about the SARE in the context of long-distance running events (LDRE). Such context is chosen because its competition and social space lead to both runners and spectators' desire for sharing SARE, providing a rich space for informing what elements of SARE are important to

collect as GT for creating SARE-ARS and why it is important for technology to capture them.

Studies for RQ2: Chapter 5 describes the design of a system called R-ESM, which assists in collecting suggested states of SARE based on the findings from Chapter 4. The system is designed for in-situ ESM purposes and wirelessly delivers speech-based ESM prompts to record runners' voice-based self-reports (including both Likert scale-based answers and open think-alouds) via earbuds. It simultaneously collects physiological and behavioural data related to runners' SARE.

Chapter 6 presents user studies to explore the feasibility of using R-ESM to collect SARE self-reports for building SARE-ARS. A quantitative evaluation was conducted using machine learning modelling as a proof of concept. Qualitative studies, using semi-structured interviews were used to understand runners' self-reporting experience and how the design of R-ESM could be improved for better SARE self-report collection.

Studies for RQ3: Chapter 7, based on the design implications from Chapter 6, presents a design for a voice-based chatbot that supports runners in self-reporting their SARE while running. The chatbot proactively asks runners questions about several elements of SARE, responds with follow-up questions, and provides personalized feedback based on runners' answers. Additionally, it guides runners with a mindful scanning exercise to increase self-awareness.

Chapter 8 reports a user study where the chatbot was used as a probe to understand runners' experience of self-reporting to the chatbot compared to self-reporting to R-ESM. A quantitative approach was used to analyse the chatbot's performance and reliability, while a qualitative approach was used to understand runners' feedback in semi-structured interviews and inform design implications for such a chatbot. The findings led to the (Modality-Autonomy-Prompting) MAP framework, which suggests

opportunities for designing the chatbot for runners' experiences and collecting self-reports as ground truth.

Studies for RQ4: Chapter 9 presents a revised version of the chatbot and the user studies that evaluate the effectiveness of the MAP framework. Furthermore, the chapter also analyses runners' think-alouds based self-reports. This analysis aims to understand the different experiential and reflective components that form SARE in order to provide better support to runners. The research findings led to the development of the DAM framework, which identifies the various components of SARE that a chatbot conversation should assist runners in self-reporting.

General Discussions and Conclusions: Finally, in Chapter 10, the thesis concludes by summarising the studies and findings related to the research questions, highlighting the main contributions. It also presents a general discussion about limitations and future work on applications to affective computing and ethical considerations in relation to the proposed approach.

Chapter 2 Related Work

This chapter of the literature review serves three main purposes. Firstly, it (Section 2.1) provides a detailed understanding of what SARE entails, drawing from research in both sports psychology and science as well as the HCI community. Secondly, it (Section 2.2) explores how SARE can be automatically recognised, drawing on knowledge from the fields of affective computing and computer science. Finally, it (Section 2.3) explores approaches to collecting SARE from runners' self-reports from an HCI perspective.

2.1 Subjective and affective running experience (SARE)

This section presents an extensive review of SARE, following these themes: perceived exertion and fatigue, psychophysiological experience, attentional and cognitive experience, mood and emotions, motivation for long-distance runs, the interaction among different types of experiences, sharing behaviours, and how technology can support such experiences and sharing.

2.1.1 Perceived exertion and fatigue

Perceived exertion is a subjective experience that describes an individual's perception of the intensity of physical activity. It is commonly perceived during long-distance running due to the extremely intense and sustained physical and mental effort. Research has shown that perceived exertion is correlated with a variety of factors. These include physiological responses that can be measured as heart rate [41], ventilator minute volume [181, 184], and the accumulation of metabolites, such as lactate, in the working muscles [195]. Morgan (1977) found that lactate accumulation was the best predictor of the runners' subjective estimation of the perceived effort, followed by ventilator minute volume and then by their heart rate, respectively Morgan and Pollock [183]. Psychological factors (e.g., emotion, motivation) can influence perceived exertion[40]. For example, Morgan (1973)

reported that runners who have neurotic, anxious, or depressed traits experience difficulty in accurately perceiving the intensity of muscular work Morgan [180].

Assessment of runners' perceived exertion has mainly utilized *The Rating of Perceived Exertion* (RPE) developed by Borg (1973; Borg Scale)[4, 41]. Runners taking part in studies or training programmes are often asked to rate the running intensity from “very light” to “very hard” based on a psychophysical category scale ranging from 6 to 20 (See Figure 2-1, Left). Recently, the Borg Scale has gained increasing popularity within the running community, with many runners and coaches⁵ preferring the slightly adapted version of 1-10 to the traditional 6-20 scale (See Figure 2-1, Right).

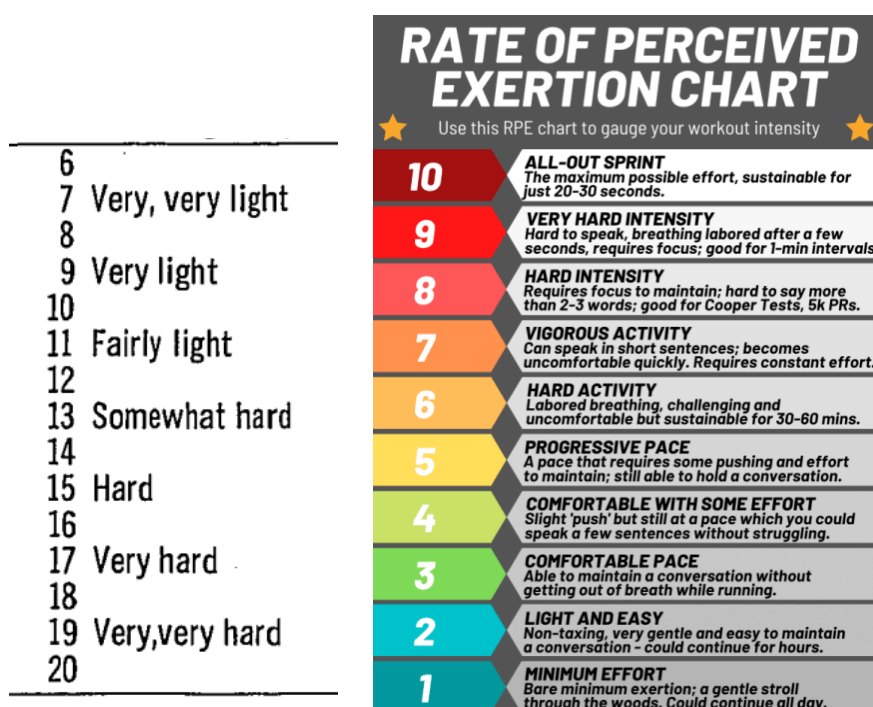


Figure 2-1: Left - Borg RPE (6-20) [4, 41]; Right – adapted RPE (1-10)⁶

⁵ <https://www.runnersworld.com/training/a29706103/what-is-rpe/>
<https://my.clevelandclinic.org/health/articles/17450-rated-perceived-exertion-rpe-scale%20>

⁶ <https://marathonhandbook.com/rate-of-perceived-exertion/>

2.1.2 Psychophysiological experience

Researchers have explored the psychophysiological aspect of affective and subjective running experiences, placing particular emphasis on runners' perceptual-sensory perception and embodied running sensations, such as those associated with auditory, smelling, and seeing, as well as tactile feelings (a series of relevant papers can be seen in [11-16, 97, 104, 105, 240]).

Runners actively engage with subjective auditory sensations while running, using breathing patterns to adjust their pace and perceive the sounds of their environment for safety. Additionally, they employ auditory feedback to distinguish between different surfaces, based on the sounds their footsteps produce. For instance, as described by Hockey (2006, p10), on soft tarmac, a *“low sssssssssssshhhh”* sound can be heard, while concrete pavement produces a *“higher and more abrasive slaaaaaaaaap or thwaaaaaaaaaaaaack sound”* and slushy snow a *“sluuuuuuuuuck sound”*. Additionally, upon firmer surfaces a *“cruuuuuuuuuuunch of crisp or firm snow”* can be discerned [104].

Smell sensations plays an important role in the psychophysiological experience of running, influencing runners' perception and facilitating decision makings. As Hockey (2006, p.10) noted, *“The strong, ripe smell is a stink I am used to; it is me when I am hard at work.”* Moreover, external smells, such as newly-cut grass, a hint of dog excrement, the scent of new training shoes, lilac tree, and pine, can trigger memories of past training sessions in similar contexts. Additionally, smells may serve as markers along the running route, such as *“the smell of curry indicating that the finish line is near, vehicle pollution at a busy traffic interchange foreshadowing the clean air in a park, or the smell of algae on a lake suggesting that the halfway point of the route has been reached”*. Thus, smells are an integral part of the running experience [104].

The subjective haptic sensations experienced by runners involve reciprocal haptic interactions between their bodies and the external environment they touch. For

example, when running, the feet interact with the ground, providing pressure and kinaesthetic awareness of their leg and feet movements. Even with running shoes, runners can still perceive the shape, size, texture, and temperature of the terrain through the soles. Descriptions of such tactile experiences, as referenced in [104] (p15), include *“the shape of items strode over being felt as the toes and forefoot grasp them,” “the feet swelling and becoming warm as the temperature rises,”* and *“the tendons at the bottom of the ankle stretching more in a boggy terrain.”* In addition to feet, runners can also perceive tactile sensations from their skin, particularly in various weather conditions, such as heat, wetness, cold, and air. For instance, *“as wind drives the snow directly at us, it may feel like gravel being rubbed into the skin, as if it was burning.”* The touch and haptic sensations experienced by runners can be used as running strategies and also contribute to the enjoyable, immersive experience of running.

In terms of visual experience, runners have been found to frequently utilize their visual capacities to modify their running pace (e.g., distinguishing the suitability of a surface for a speedy cadence) and to perceive potential safety hazards (e.g., recognizing objects that could cause harm or accidents). More example descriptions of such psychophysiological running experiences (auditory, visual, smelling, haptic) have also been documented in [76, 107, 108, 268].

2.1.3 Attentional and cognitive experience

Long-distance running is subject to dynamic changes in attentional focus over an extended period of time. Sport psychology suggests that attention can be conceptualized in two dimensions: width (broad-narrow) and direction (internal-external). Broad attention encompasses the consideration of multiple stimuli, while a narrow attention indicates a focus on one or two elements [190, 191]. Internal attentional states involve concentration on one's own thoughts, body sensations, and imagined scenarios, as well as conscious monitoring of motor performance, such as movement execution and breathing [247]. External attentional states, on the other

hand, involve focusing on external stimuli, including task-relevant external focus such as external strategy, mile markers, water stations, split times, route, and conditions, or task-irrelevant external focus such as the environment, other athletes, spectators, and coaches [247]. This combination of these two dimensions creates four distinct quadrants of attentional focus, namely the Broad-External zone, which focuses on perceiving and assessing many different cues from the environment; the Broad-Internal zone, which focuses on analysing the information; the Narrow-Internal zone, which focuses on preparing one thought; and the Narrow-External zone, which focuses on one environmental cue to take action [190, 191], as shown in the Figure 2-2 (a replicated diagram based on such understanding).

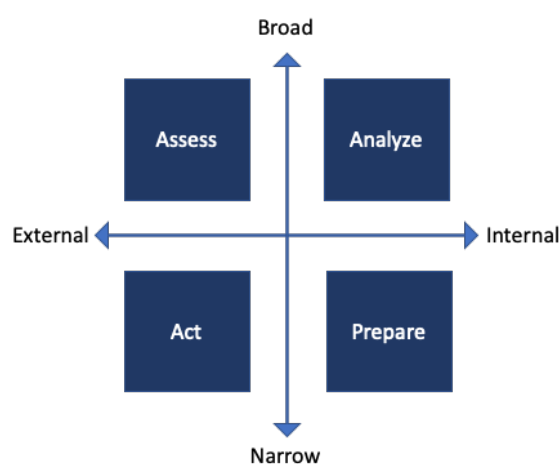


Figure 2-2: Attentional focus and in Sports (a replicated diagram based on understanding from [190, 191])

The direction of a runner's attention can affect their subjective feelings such as affective relaxation, tolerance to discomfort and exertion. For instance, having the flexibility to shift between internal and external attention can contribute to feelings of relaxation for recreational runners, as they are not required to remain in an internal, narrow state all the time, and may instead enjoy the beauty of their surroundings [255]. In fact, runners who are instructed with external focuses perceive increased tolerance to discomfort and exertion, which enables them to endure running for a longer duration than those with internal focuses, due to the

distraction that external cues provide [182]. Conversely, the perceived workload also influences a runner's attentional state, prompting them to dynamically and voluntarily shift between internal and external focus in a relaxing situation with lower perceived workload efforts, and an internal state during a high workload situation to maximise an effective use of resources or overcome challenges [255].

Attentional focus has been in fact shown to have an effect on running speed and motor movement economy. Internal focus has been linked to faster running speed, whereas external focus has been associated with slower running speed [143]. This could be caused by increased awareness of activating and using oxygen and muscles, leading to higher performance and faster pace, but consuming more energy and reducing endurance. In contrast, external focus (e.g., on the surrounding environment) has a positive impact on running economy with lower oxygen consumption, meaning fewer requirements for physiological resources [230]. Internal attentional focus can also help to monitor body sensations and cues, thus allowing adjustment of physiological resources to reduce the risk of injury. Moreover, inexperienced runners have been observed to run more economically and efficiently when their attentional focus is external[231]. Lastly, different running goals may dictate the direction of runners' attentional focus; for instance, those running for mental wellbeing or socialization may not wish to focus on internal aspects.

Proactively manipulating attentional focus is often utilized as a cognitive coping strategy during long-distance running, such as through employing association and dissociation strategies [253] [246] [183]. An associative strategy involves reading bodily signals, such as sensations and feelings arising from different body parts, perceiving effort and time, and then understanding and modulating accordingly in order to achieve a relatively steady state. Dissociative strategies, on the other hand, involve taking one's mind off the run, including any effort perception, in order to reduce boredom, monotony, and pain, unless a strong cue or stimuli, such as heat exhaustion or a bone fracture, is present. Examples of such strategies that can be employed in this regard include *“reconstructing images of past events”*,

“reconstructing previous educational experiences”, “building a house”, “writing letters to people”, and “performing complex mathematical exercises” (Page 9) [183].

The dissociative runner may experience energy deficits throughout the run, prompting them to adjust their pace (e.g., slow down) in order to maintain their dissociative state. However, this strategy is inefficient in terms of fuel consumption, which can lead to the runner hitting the proverbial “wall” and ultimately stopping the race or affecting their performance. Conversely, elite runners who employ associative strategies are able to consume less oxygen than amateur distance runners.

Cognitive association and dissociation strategies can also lead to a change in runners' subjective perceptions of pain and exertion, especially for experienced runners, who have the cognitive ability or emotional intelligence to switch off such subjective feelings, as shown in Figure 2-3

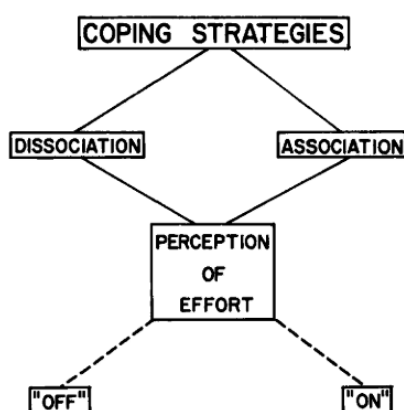


Figure 2-3: Coping strategies ([179], Morgan and Pollock [183])

The findings related to attentional and cognitive states suggest that relying solely on simple measures of exertion, such as using RPE, may not provide a complete picture of a runner's perceived state. As runners often adopt various strategies to improve their energy expenditure and manage their attentional and cognitive states. Therefore, gaining insight into these states and their developing effective strategies

to manage them can provide a deeper understanding of runners' subjective experiences.

2.1.4 Mood and emotions during long distance running

Given the important of affective experience in our studies, we review here related literature from both the sport psychology and affective science fields. In doing so, we introduce the main emotion frameworks and theories and related debates to inform our work on how runners' emotion and mood are processed and what components of affect can be measured.

2.1.4.1 Affective experience in sport psychology

Long distance running is characterised by dynamically evolving affective states. For instance, [144] showed that ultra-long-distance runners experienced significant fluctuations in mood throughout the Marathon of Britain (282 km over six days), marked by a decline in vigour and an increase in fatigue from the start to the finish, as shown in Figure 2-4. Additionally, levels of runners' perceived calmness and happiness appeared to decrease in the later stages of the race, but rose again at the final stage, close to the finishing line. Fatigue and confusion significantly increased and reached a peak at the last second stage before the finish. Furthermore, runners tested with high emotional intelligence displayed higher levels of calmness and happiness, and lower levels of anger, confusion, depression, fatigue, and tension than those with low trait emotional intelligence.

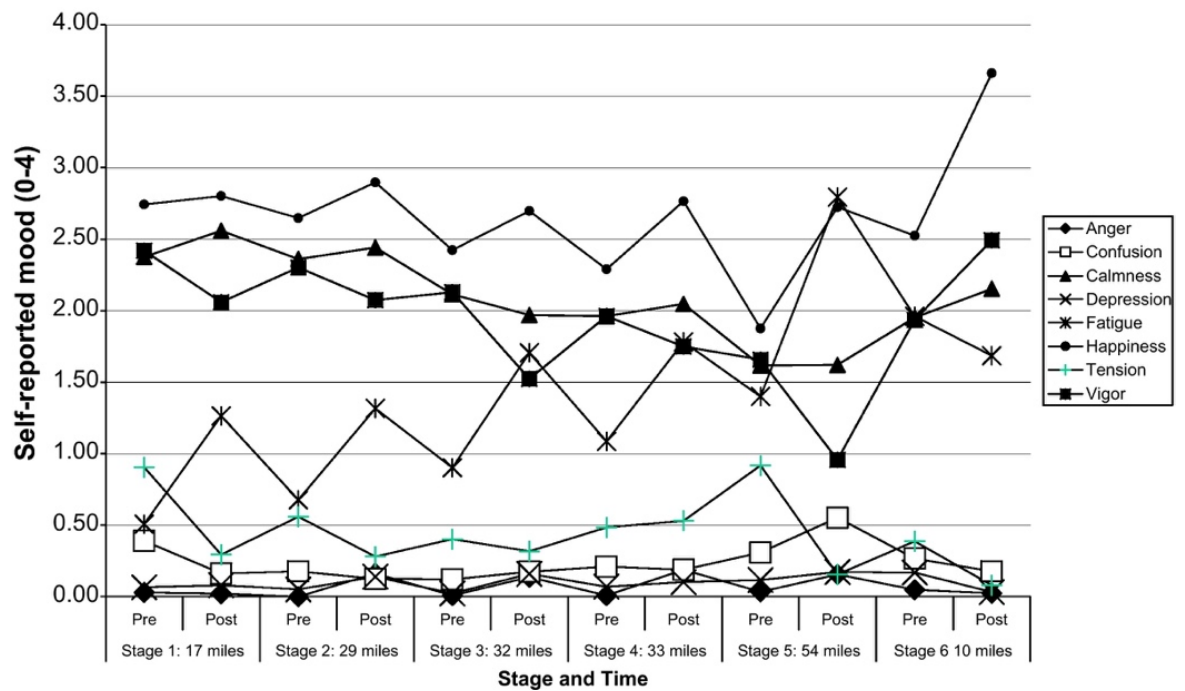


Figure 2-4 Ultra-marathoners' Mood changes (an image from [144])

In order to assess runners' mood changes, the most commonly employed approach is the use of Likert scale-type questionnaires, such as the Profile of Mood States (POMS) - 65 items rated on a 5-point scale [170], the Brunel Mood Scale (BRUMS) - a shortened version of the POMS for assessing unpleasant emotions such as anger, confusion, depression, fatigue, tension and vigour [256], and the UWIST for assessment of pleasant emotions such as happiness and calmness [168]. Examples of studies using such scales for assessing runners' emotions can be found in [149, 209]. According to affective and cognitive science research, mood is a more chronic and persistent affective state that is comparatively stable when compared to transient emotions, which are typically experienced as more short-lived psychological events [211]. Emotional states can vary in intensity over time and typically have *"a distinct cause, a specific cognitive and experiential content, and a rapid decay following elicitation"* [211] (page2). One of the limitations of studies [149, 168, 209, 256], aiming to assess mood and affect is that they have mainly been conducted before and/or after the run, with a focus on the overall evaluation of the sustained mood, lacking the measurement of the detailed progress of how emotional states transit

while in-the-moment running. Thus, this thesis seeks to address this gap by exploring the capture of the transitory emotional state during the running process.

2.1.4.2 Main emotion framework and related debates on measuring affect

To enhance our understanding of the possibility of measuring components of emotions and affect, we review the main emotion frameworks from the affective science literature. The definition of emotion or affect has been a topic of discussion for several decades but has yet to reach a consensus.

According to the appraisal theory [178], emotions are not merely reactive responses but rather a result of a multilevel sequential evaluation of different aspects of a situation. These evaluations include considerations such as the direct impact of an event on an individual or their social group, the novelty or familiarity of the event, the pleasantness or unpleasantness of the event, the level of urgency or importance attributed to it, the assessment of consequences and their impact on personal goals, the causal attributions and reasons behind the event, the potential outcomes, and the coping strategies employed to deal with the consequences. It suggests that emotions are influenced by these diverse appraisal checks, thus highlighting the intricate nature of emotional experiences.

On the other hand, a constructionist view of emotion [31, 33] focuses on the physiological determinants of an affective experience. It suggests that each specific emotion category is characterized by distinct responses of the autonomic nervous system (ANS), which can serve as diagnostic markers of the specific state. However, recent constructivist theories (e.g., Siegel (2018) [238]; Feldman-Barrett (2014) [212]) provide evidence that ANS patterns vary within and across categories, and there may be overlaps among different categories. Such variability can even occur in the same situation across different instances. The consistency of ANS patterns may be relatively low, meaning that statistically generalized patterns may not be observed in each individual instance. Levenson [150] has also proposed that the identified distinct

patterns may be a result of emotions often being induced and studied in controlled laboratory environments, suggesting that real-world scenarios may not exhibit the same clear-cut ANS distinctions across different emotion categories. Context, instead is seen as critical information that brings more uniformity to such affective responses (both physiological and behavioural), creating a common language to classify them and facilitate their interpretation.

Traditional models of emotion have been predominantly categorized into two types: category-based and dimensional models. Basic Emotion theory [72] is one of the most impactful category-based emotion theories, which identifies a set of universally recognized emotions such as fear, anger, joy, sadness, disgust, and surprise. However, there is no consensus among researchers regarding the specific types of basic emotions, with different perspectives being put forth. For example, Jack et al. (2014) suggest four rather than six latent and culturally common facial expressions of emotions: fear, anger, joy, and sadness [111]. Other emotional states are often considered as complex combinations of those basic emotions [208].

On the other hand, dimensional models represent emotions along multiple dimensions based on core features or underlying factors. One of such models is the circumplex model [220], which proposes two affective dimensions: valence and arousal. Valence represents the positive or negative nature of an emotion, where for example, a state of happiness would be considered to have a positive valence and sadness a negative one. Instead, arousal reflects the intensity or activation level of emotion, ranging from low (calmness) to high (excitement). Other typical dimensions used to characterise an emotional experience are dominance [172] and action tendency [82, 84]. These dimensions aim to capture the sense of control one perceives to have and their readiness to act in response to the situation.

These emotion models (and related ones) have influenced the development of various measurement approaches to assess emotional states. In the context of categorical theory, researchers have created instruments such as "The Multiple

Affect Adjective Check List" [283] and different types of Profile of Mood States (POMS) [35, 235]. These tools utilize item checklists to measure affective states, targeting specific emotions or mood dimensions. Often, the items included in these measures are often modified to align with the research goals of individual studies. However, according to Siegel (2018)[238], there is a lack of conceptual argument and empirical evidence to support the idea that the categorical distinct states, e.g., the six states assessed by the Profile of Mood States, can adequately represent the entire range of mood experiences. As a result, making extrapolations from assessments based on the Profile of Mood States to encompass the broader domain of mood is unjustified [238] (p108). Siegel (2018) (p.116) further proposes that the study of emotions should expand each individual instance of the emotion with respect to its context rather than relying solely on high-level categories.

Dimensional models have also led to the development of specific measurement approaches. One example is the Affect Grid [221], which consists of a horizontal axis representing valence, ranging from negative on the left to positive on the right, and a vertical axis representing arousal, ranging from low at the bottom to high at the top. This creates four quadrants within the grid, each representing different combinations of valence and arousal. For instance, the Pleasant High Arousal quadrant represents positive and highly activating emotions like excitement or enthusiasm, while the Unpleasant High Arousal quadrant represents negative and highly activating emotions such as anger or anxiety.

Another measurement approach used in dimensional models is the Self-Assessment Manikin (SAM) [42]. SAM employs cartoon characters with progressively changing facial and bodily cues to represent emotional states along specific scales. However, Siegel (2018) argues that relying solely on these two dimensions is insufficient to fully capture the diverse characteristics of different emotion categories. For instance, the affect grid might be overly simplistic and fail to capture the multidimensional nature of emotions. Additionally, it lacks the inclusion of contextual information and relies on subjective interpretation by researchers, which disregards cultural and individual

variations. Furthermore, it does not provide specific emotion labels, posing a challenge in precisely identifying and categorizing emotions. Other approaches (e.g., the Geneva Wheel [228]) have proposed tools that combine the dimensional and discrete emotion approach to capture a richer description of the states. Despite the wide use of different tools, there is still a lack of consensus among researchers regarding the validity of scales or questionnaires due to the complex and dynamic nature of emotions [71] as well as their variability in expression [212].

Furthermore, there are some concerns regarding the application of such approaches in the form of simple questionnaires. One major criticism is that many researchers merely replicate these approaches by stating that "*this measure was used because it has been used before*" (p11) without offering a rationale for their choice of measures [71]. They often assume that these measurement approaches are well-established and valid, overlooking the importance of critically evaluating and addressing the potential disagreements surrounding the psychometric properties of the instruments they employ. This is particularly evident in the field of affective computing, where researchers frequently gather discrete emotional labels through these methods and treat them as ground truth in machine learning research, without adequately considering and justifying the validity of these approaches for their specific applications. For instance, some researchers directly use the Exercise-induced Feeling Inventory questionnaire to assess a person's feelings before physical activity (p31) [71]. However, certain constructs in this questionnaire, like feeling refreshed, may not align with the context in which a person has not yet started the physical activity.

To address these issues, Ekkekakis proposes three steps for researchers. Firstly, they need to determine whether their research focuses on core affect, mood, or emotion and establish a specific definition of emotion that suits their research agenda. This step aims to avoid misinterpretation and facilitate the selection of appropriate measurements. Secondly, researchers should explore and explain how and why they employ current emotion theory to conceptualize emotion in their research. For instance, they should justify their choice of viewing emotions as distinct categories

or as dimensions with respect to current emotion theories, considering factors that align with their research objectives. One example provided in Ekkekakis' (2013) book is to discuss whether *“pleasure and pain should be regarded as points on a continuum or as completely separate and independent dimensions”* (p98)[71]. Lastly, researchers should critically evaluate and discuss the validity of the chosen measurement before applying it to their research agenda.

The review of literature indicates that emotions and affect are highly complex and dynamic constructs. Neither a simplistic categorical approach nor a dimensional approach alone would sufficiently capture the dynamic and complex nature of runners' affective experience in real-world contexts. In order to comprehensively understand and analyse this phenomenon, it is essential to establish a solid rationale and demonstrate the applicability of current measurement approaches, while avoiding overly simplistic usage. In addition, there is variability in the expression of emotion, both physiologically and behaviourally, as well as in how the context influences the labelling of each emotion [32].

In this thesis, instead of solely relying on predetermined emotion models, our approach will involve an in-depth qualitative exploration with a focus on the real-world context of long-distance running to characterize the crucial aspects involved in SARE. While we will initially use traditional approaches to gather runners' ratings of their emotional states and employ machine learning algorithms for automatically recognising them, the remaining thesis will focus on novel approaches to understanding and capturing the complex experiences of runners, going beyond simple states or dimensions. This includes exploring dynamic and complex emotional experiences, such as how runners generate appraisals and intentions to effectively manage their emotional states in real-world scenarios characterized by continuous running and evolving emotions. We will also explore runners' thinking processes and self-reporting of these experiences, aiming to propose innovative approaches and design new tools for capturing them.

2.1.5 Motivation for long-distance distance running

Motivation for long-distance running engagement and adherence can be categorised into two types, intrinsic and extrinsic [141, 261, 262]. Intrinsically motivated running behaviour is driven by the enjoyment derived from the activity itself, such as a sense of fun and satisfaction with performance, joy of running, and sense of wellbeing; Extrinsic motivation, on the other hand, is driven by external factors, such as social approval, peer influence, parental influence, rewards, and recognition [262]. It has been found that extrinsic motivation may be the initial driver of involvement in running, while intrinsic motivation helps to maintain persistence in distance running [183]

Researchers in HCI community also presented a perspective to explain the motivational factors of amateur runners in long-distance races, such as the festival atmosphere, competition, practicalities, togetherness, and the support of spectators [134]. The intense atmosphere of the event was reported to increase the runners' excitement and euphoria. Additionally, the sense of competition, both internally with oneself and externally with others, was identified as a source of motivation. Practicalities, such as having a scheduled event as a goal, served as a motivator for racers to attend their daily training sessions. The social togetherness experienced with other runners and spectators further encouraged runners. Moreover, the support of spectators was identified as a major motivating factor for runners, as they were encouraged on by cheering and positive affirmation to help them achieve their racing goals.

2.1.6 Interaction among different types of experience

The literature on sports psychology has already demonstrated an interaction effect between the different aspects of running experience. For instance, the influence of mood states on competitive running performance is contingent on the associated physiological arousal level. Specifically, when accompanied by high levels of

physiological arousal, tension, anger, and vigour can be detrimental to performance; conversely, when accompanied by moderate levels of physiological arousal, these same moods can be advantageous. Furthermore, depression, confusion, and fatigue, which are associated with lower levels of physiological arousal, tend to lead to a decreased level of motivation for performing and competing Prapavessis [209].

Sports psychologists later argued that the level of arousal does not necessarily determine performance; rather, it is the athlete's subjective and affective interpretation of such arousal as either positive or negative that affects their performance [139]. Arousal in sports psychology is designed as a bodily energetic response caused by the nervous system when an athlete is emotionally stimulated [95]. For the same level of arousal, when interpreted positively, can be seen as excitement that can facilitate athlete performance, whereas with a negative interpretation, it can be seen as anxiety that hinders the athlete's performance. Individual differences also exist when perceiving and interpreting arousal, e.g., between an athlete and a recreational runner [139].

The notion that subjective appraisal of arousal can affect sports performance is further explained by the conscious processing theory (also called the reinvestment theory). The theory describes that when in a negative state, athletes may become more cognitively conscious of their movement control, leading to a loss of automatic response skills and a decrease in motor performance speed and efficiency [167].

Cue Utilization Theory [279] could also explain the relationship between arousal and running performance, from the perspective of the width of attentional focus. At a low arousal level, an athlete may have a broad attentional state, encompassing external surroundings, the crowd, past experiences, and other environmental factors, without sufficient focus on the most pertinent elements. At an optimal arousal level, an optimal range of attentional width is achieved, allowing for selective focus on the most important stimuli without being too wide or too narrow. For instance, a basketball player in an optimal arousal state may focus on movement techniques,

collaboration with team members, and the defence of the opposing team. Conversely, when overly aroused, an athlete may have a very narrow attentional state, wherein they focus on only one or two stimuli, potentially leading to a negative impact on performance.

The findings imply that gaining insight into how different subjective states interact with one another, such as a runner's subjective evaluation of their physical state and their attentional response to their perceived physical arousal, can be valuable in obtaining a comprehensive understanding of their subjective experiences.

2.1.7 Runners' social sharing behaviours and supporting technology

Given this thesis' focus on the sharing of subjective and affective running experiences, this section presents an overview of available knowledge concerning runners' social sharing behaviours and the ways in which current technologies support this practice.

[78] investigated outdoor runners' sharing behaviors (with a particular focus on trail running) from five dimensions: temporal, motivational, content-related, audience-related, and privacy-related. Information shared by runners can be categorized according to three temporal stages: prior to running (e.g., weather and training plans), during the run (e.g., GPS and heart rate data), and after running (e.g., pace and distance data). Runners' sharing preferences can also be divided by the individuals with whom they choose to share. Some runners only share with those they trust or members of their running group, either through online posts or direct messages, while others do not mind sharing publicly. Some may have privacy concerns, such as exposure of personal information, psychological and personality exposure, and potential spamming, which can lead to reduced sharing behaviours over time [78] [214]. This is inspiring for this thesis, as it focuses on the sharing affective running experience between LDRE runners and spectators with close relationships to them. Despite previous research into the need for sharing running-related information, this has typically focused on supporting the sharing of workout statistics and location

(Romanowski, Mayer [219]). As suggested by Woźniak (2017), there is a need to develop socially-embedded technology for outdoor sports in order to address the social and affective needs of these runners.

Sports psychology research has shown that sharing running information is an important means of motivating athletes to remain physically active, as well as fostering the creation and maintenance of social ties, promoting egalitarianism, mutuality, openness, and trust within a community [116]. Runners may share routes for seeking social companionship or increasing running diversity, running challenges for obtaining help, live location data for coordinating help with commuting and transportation, physiological and biomechanical data for technical discussion with fellow runners, tracking and feeling fatigue for training and performance optimization, and communicating running achievements to friends and family, or sharing performance for building an attractive social profile or for a sense of peer competition (e.g., [78], Ojala [196], [274]). It is noteworthy that completed running events or races are often shared and receive greater attention from audiences, as they are seen as a challenge and the experience of achievement [274]. This body of literature does demonstrate runners' interests in running-related information for affective needs; however, a comprehensive understanding of needs for sharing affective and subjective running experiences, particularly in the context of long-distance running events, is still lacking.

Runners may choose to share with only those they trust or may be concerned about the exposure of personal information, psychological, and personality traits, which can lead to reduced sharing behaviours over time [78]. This also motivates the scope of focusing on the affective experience of sharing running between long-distance runners and spectators with whom they have close relationships.

2.1.8 Long-distance running event (LDRE) and technology for sharing

Running events have been widely used as a public health strategy to address public health concerns and promote citizen well-being, social engagement, enjoyment, and entertainment [1, 2], as well as spurring local economies [3]. Long-distance running events (LDRE) have gained popularity in recent years, with approximately 807,000 events held in the UK in one year [4]. The number of participants in marathons worldwide has also increased by 13% between 2009 and 2014, with significant growth in Asia (92%), Europe (10%), and the USA (14%) [5]. Moreover, large crowds come to spectate, providing much-needed encouragement and motivation, thus enhancing engagement for the runners. For example, at the 2018 London Marathon, 800,000 spectators showed up - 20 times more than the number of runners in the race [6]. These events not only provide a platform for runners and their spectators to share experiences but also serve as an encouragement for spectators to become more physically active and participate into running [237].

Despite the popularity of LDRE, there remain certain challenges that prevent runners and spectators from having a coordinated experience. Geographically dispersed routes, high participant numbers, and prolonged durations (4-6 hours) typically restrict opportunities for interactions and connected experiences between runners and spectators, who may only have brief encounters when the runner passes by [50, 276].

Current commercial solutions focused on utilizing cutting-edge wireless communication technology to update runners' real-time location and distance. For example, the London Marathon mobile app tracked runners' progress every five kilometres, enabling spectators to plan for optimal viewing points and reach them in time to cheer on their supported runners [9][7-10]. Current technology (e.g. MYLAPS) also allows remote spectators to track runners through live position and leader board updates or social media sharing[127]. Further, larger-scale events like the Boston Marathon or the Olympic Games marathons also utilize broadcasting technology to

allow for remote spectators to livestream the event [2, 3, 34]. However, such approaches only provide limited information on location and basic performance statistics, and they still treat the spectators as passive consumers of such information.

Researchers in the HCI community have proposed shifting spectators from passive consumers of running information to active digital supporters by sending digital cheering via interactive technology [63, 276]. Some HCI studies have attempted to support runner-spectator interaction in real-time during LDRE, such as HeartLink and RUFUS. HeartLink allows a runner to broadcast heart rate data and enables online spectators to send cheering vibrations to support runners [62]. The RUFUS system allows a runner to actively request or stop cheering signals from spectators [276] as runners might want to limit external interruptions during a competitive race. Interactive tools like these can motivate runners and make spectators feel engaged, but spectators have expressed eagerness for a stronger sense of connectedness [62, 277]. These studies showed that spectators were not satisfied with simple virtual “cheers” but wanted to access more runners’ experience and have real interactions like onsite spectators [62].

While some researchers are developing technology to automatically recognize advanced running information (e.g., foot strike [101], gesture [115, 194, 234, 250-252] and rhythmic breathing[263]), others are innovating new ways of socially sharing running information. For example, TickTockRun allows runners to display their daily training information (i.e., location, speed, route) at home, enabling them to share their running experiences with family members [133]; SMASH is a handle-shaped device that enables spectators to feel haptic sensations (synchronized to runner’s heartbeat) through in their palms for a sense of closeness [106]; RunWithMe uses 360-degree video to stream a runner’s view to online spectators [18], while a computer-controllable flying Quadrotor has been developed as a running companion [96]. Additionally, technology has been developed for virtual team running experiences [185], and shared LED screens have been designed on running t-shirts to

display group running pace and distance for augmented social running experience[169].

However, the captured and shared information supported such technologies remains limited to objective running metrics. The SARE is yet to be directly understood by the technology. Developing such technology that can automatically recognize SARE and support the sharing of it will be of great importance, particularly in supporting the cheering between runners and spectators during LDRE, as the runners' SARE remains unseen by their spectators and largely unaccounted for by existing wearable tracking and sensing technologies.

Hence, the following section provides an overview of the literature on automatic recognition on runners' SARE.

2.2 Automatically capturing runners' subjective and affective running experience

This section provides an overview of various aspects related to the construction of technology designed to automatically capture runners' SARE. It introduces foundational concepts from the fields of machine learning (ML), automatic recognition systems (ARS), and ground truth (GT) in the context of runners. The limitations of current approaches in collecting runners' self-reports for building SARE-ARS are discussed, along with an exploration of the concept of ubiquitous ground truth collection using the Experience Sampling Method (ESM). The section also presents an overview of the ML methods employed and the results. Finally, potential features and sensing technologies that can contribute to the development of an effective ARS for assessing runners' SARE are explored.

2.2.1 Machine Learning (ML), Automatic Recognition System (ARS) and Ground Truth (GT) of affective states

Machine learning is a widely adopted approach for building automatic recognition systems for emotions from signals such as facial expressions, speech, behaviour, and

physiological changes. Common Machine learning approaches include supervised learning, unsupervised learning, and reinforcement learning. Supervised machine learning is often chosen due to its accuracy in predicting or classifying the user's emotional state. It requires labelled datasets to train the algorithms to learn the meaning and patterns, in order to be able to perform recognition on other datasets. In addition, it tends to require less data than unsupervised learning, which makes it the preferred choice for small-scale studies aiming for initial exploration of automatic recognition. Therefore, this section begins the exploration with supervised machine learning for building an automatic recognition system for runners' subjective and affective state.

Literature in the field of affective computing has provided evidence for the feasibility of recognizing emotional and cognitive states through physiological and behavioural signals. For instance, motion sensors have been used to automatically infer people's pain levels during simple physical rehabilitation exercises [201], and classify positive emotions (e.g. excitement and pride) and negative emotions (e.g., defeated) and cognitive state (e.g. concentration, low-self-efficacy) during physical activity [132, 198, 224]. Head gestures can be captured via motion sensors to indicate cognitive state of reflection and thinking as people (children in this case) engaged in exploring the external world [200]. EMG sensors have been used to capture muscle activity to infer fear and pain [27, 254]. Physiological sensors can be used to capture heart rate (HR) and galvanic skin response (GSR) to detect feelings of pain and track high arousal, excitement and frustration in both active and static situations [118-121, 201, 207, 270]. In addition, A combination of various signals captured with multiple types of sensors may contribute to increased machine learning recognition performances (e.g., [27, 202]).

Using machine learning approaches to build ARS often requires ground truth labels. Ground truth⁷ labels refer to the annotations assigned to data points, typically in supervised machine learning settings, where a model is trained on labelled data. These labels serve as the “truth” or gold standard against which the model's performance is then evaluated. In the context of affective computing, despite the community talk about ground truth, there is a recognition that such ground truth is simply a general subjective observation and qualification of an event that is from an external observer or self-observations. Still, such observations can provide often useful cues on how a person feels and how the person could be supported (e.g. [202]). Self-reports, in the context of affective computing data collection, refer to individuals reporting their own experiences, thoughts, or behaviours [59]. Self-reports have been widely used as a means of obtaining ground truth labels in various fields, including psychology [85] medicine [249], and sports[215]. Self-reports are most commonly based on reporting affective states labels or rating such states on a set scale. In some studies, ranking rather than rating has been also used people differ in the way they use numerical scales when judging their affective states. In addition, ranking is sometime considered more reliable thanks to the use of reference point to assess one’s affective situation [166]. Recently there is a growing recognition that such methods are still quite limited calling for new methods [161], e.g., qualitative approaches [229]. Still, their use of qualitative approaches is very limited and focused on post processing of the data and lab studies.

2.2.2 Limitations in current work related to collecting runners’ self-reports for building SARE-ARS

Table 2-1 provides a summary of the research work related to SARE-ARS available up to 2023. The interest in applying machine learning to automatically recognize runners' subjective experience only emerged in 2017, and the study presented in this

⁷ <https://www.dominodatalab.com/data-science-dictionary/ground-truth>

PhD thesis was conducted in 2018 and completed in 2019. By 2019, only two studies had explored machine learning approaches for classifying runners' fatigue, but there has been an increasing interest in this area since then, with more attention in 2021 and 2022. In this section, we first present a critical analysis of the approach for collecting self-reports as ground truth in these works. ML approaches used in these studies will be summarised in Section 2.2.4.

Refs	Year	SARE type	Induced/ Natural	Evaluation Method	Lab/Outdoor	Report- Behaviour method	Target Class	Data Used For ML	Data Ignored	Sensor	Features	Models
[46]	2017	Overall Fatigue	Induced & Controlled, 3-Stage Run	Borg RPE 6-20	400m Outdoor Track	Researcher Instructed	Non- Fatigued, Fatigued	Static Non- Fatigue or Fatigue Data Collected at Stage 1 & 3.	Dynamically Progressive Fatigue Data Collected Stage 2.	IMU	Biomechanical Features	RF,SVM, KNN, Naïve Bayes
[203]	2018	Overall Fatigue	Induced & Controlled Multi-Stage Run	Borg RPE 6-20	400m Outdoor Track	Researcher Instructed.	Non- Fatigued, Fatigued	All Progressive Fatigue Data Were Used.	N/A	IMU	Statistical & Biomechanical Features	GBRT, ANN,EN, LASSO
[128]	2019	Localised Fatigue on Leg Muscle	Induced & Controlled, 3-Stage Run	Lactate Threshold	Lab Treadmill	Researcher Instructed (30 Seconds)	Aerobic, Anaerobic, Recovery	All Data in Three Stages	N/A	sEMG; Blood Lactate Sensor	Frequency & Time-Based Domain Features	RF
[99]	2021	Localised Fatigue on Trunk	Induced & Controlled, 3-Stage Run	Borg RPE 6-20	Lab Treadmill	Researcher Instructed	Non- Fatigued, Fatigued	Static Non- Fatigue or Fatigue Data Collected at Stage 1 & 3.	Dynamically Progressive Fatigue Data Collected Stage 2.	Motion Camera & Ground Reaction Force	Biomechanical Features	RF, Bagging SVM, KNN, BDT
[165]	2021	Overall Fatigue	Induced & Controlled, 3-Stage Run	Borg RPE 6-20	400m Outdoor Track	Researcher Instructed.	Non/Mild Heavily Fatigued	Static Non/Mild/Heavi ly-Fatigue Data Collected at Stage 1 & 3.	Dynamically Progressive Fatigue Data Collected Stage 2.	Motion IMU	Biomechanical, Statistical & Spatiotemporal Features	RF
[89, 90, 125]	2022	Overall Fatigue & Wellbeing	Natural & Non- Controlled, Continuous Run	Borg RPE 6-20; Wellbeing Scale [-5, 5]	Outdoor Across Locations	Participants Self- Reported, (prompts)	Individual Ratings of RPE & Wellbeing	All Stage Data	None	IMU Sensor, Audio Sensor	Biomechanical Features, Audio of Surface Type	FFNN; Linear SVM

Table 2-1 Recent work on using runners' self-reports as ground truth for building SARE-ARS

Despite the growing interest in the development of SARE-ARS, the current literature is limited in terms of the methods for collecting self-reports as ground truth. First of all, the focus of current ARS has mainly been on their subjectively perceived fatigue or exertion during running. For instance, ARS has been built to recognize runners' overall perceived exertion or fatigue (i.e., whether a runner feels fatigued or not fatigued due to an increased workload) [46],[203], [165], [259]. Other research has also developed ARS to recognize runners' localized exertion or fatigue, such as leg muscle fatigue [128] or trunk and core fatigue [99]. These studies have primarily been aimed at avoiding injury, optimizing training efficiency, and improving running performance. However, as discussed previously, SARE is not limited to fatigue or perceived exertion. Especially in long-distance running, emotional experiences such as discomfort, pain, and emotional changes (e.g., excitement, happiness, confidence, anxiety, fear, self-motivation) are often co-existed with fatigue and exertion. Consequently, there is a gap in the literature about researching ARS for these more multifaceted subjective and affective experiences during running. By the time of this thesis' completion, the ARS literature had just started to see a similar interest in runners' other affective states, such as runners' subjectively perceived wellbeing [90]. However, wellbeing is a broad, umbrella term when being used to describe a runner's affective state, because it can encompass a variety of affective or physical experiences such as feeling happy, comfortable, fulfilled, or injury-free [257]. Nevertheless, such an emerging interest also, in a way, validates the concept of this PhD thesis.

Furthermore, the ground truth used for building SARE-ARS has been mainly rating-based evaluation of runners' subjective fatigue. The studies reviewed in the Table 2-1 commonly used a scale-type questionnaire, Borg's Rating of Perceived Exertion (Borg RPE), to ask runners to rate their perceived fatigue on a numerical scale ranging from 6 to 20. This differs from standard scales that often start from 0, as it was developed to be correlated to heart rate; specifically, 6 corresponds to a heart rate of 60 beats/minute, whereas 19 corresponds to a heart rate of 190 beats/minute. Semantic

descriptions were also provided to facilitate participants in their rating; for instance, “7.5: *Extremely light*,” “9: *Very light*,” “11: *Light*,” “13: *Somewhat hard*,” “15: *Hard (heavy)*,” “17: *Very hard*,” “19: *Extremely hard*,” and “20: *Maximal exertion*.” Although such numerical reports are easily used for machine learning modelling, they remain simple and missing the rich contextual information on other subjective states (e.g., emotional, cognitive and attentional state as stated in Section 2.1) that are often related to perceived exertion.

More importantly, the development of ARS for runners' subjective fatigue has primarily relied on experimentally induced fatigue rather than the fatigue experience that runners might encounter in naturalistic real-world situations. To collect data for constructing an ARS to measure runners' fatigue, researchers typically employ a three-stage procedure for controlling and inducing fatigue. Figure 2-5 illustrates this process, with the x-axis representing the time series of a run and the y-axis indicating the fatigue level of the runners (measured using the Borg Rating of Perceived Exertion Scale). This procedure begins with participants running at a controlled, easy pace in Stage 1, followed by a fatigue-inducing protocol that progressively and linearly increases workload in Stage 2, and ends with participants running at a controlled pace again in Stage 3. For instance, in Study [46], Stage 1 involves participants running 400 m at their natural 5 km pace, which is labelled as non-fatigued run, whereas Stage 2 involves participants running at progressively increased speed every 2 minutes until either an ($RPE \geq 18$) or an inability to continue is reached, which is the fatigue-inducing protocol. Stage 3 involves participants running another 400 m, which is labelled as a fatigued run. Similarly, in Study [128], Stage 1: Non-Fatigued State involves participants running 6 min at a constant speed (below 10 km tempo speed). Stage 2: Fatiguing-Inducing Protocol involves participants running at a progressively increasing speed (workload) and reporting their RPE every 2 minutes until they reach a lactate level of 5 mmol/L or an inability to continue. They then maintain the workload until an inability to continue (for a minimum of 6 min). Finally, Stage 3:

Fatigued State involves participants running for 6 min at the same speed as in Stage 1.

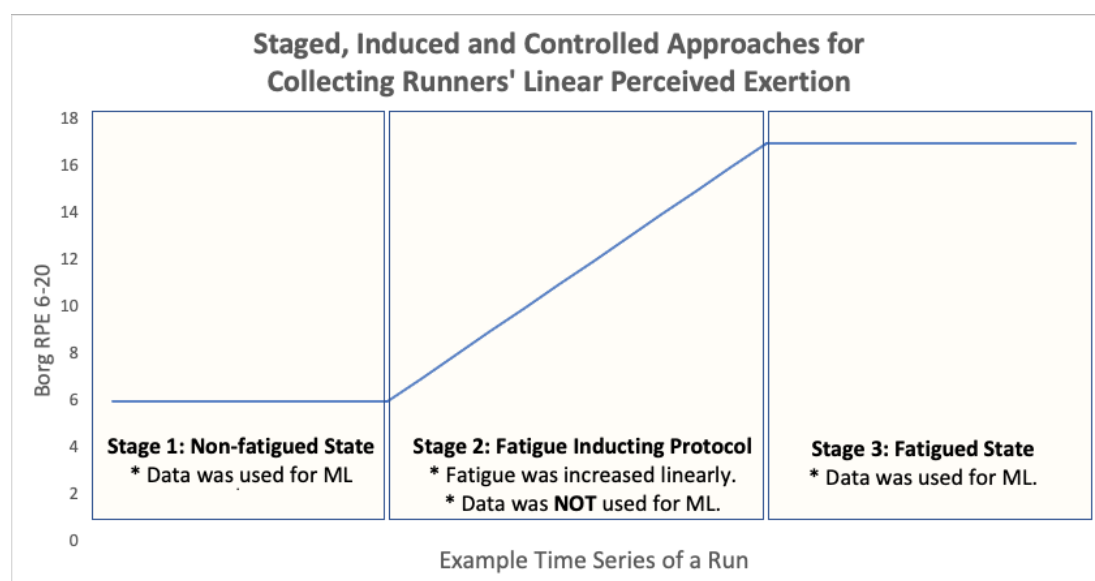


Figure 2-5: three-stage procedure for inducing and collecting reports of fatigue

Although such a three-stage approach can demonstrate the rigor of scientific research design, it may not fully capture the ubiquity of subjective experiences of runners in their natural environments or the complex interactions among various affective, social, and environmental factors that occur during uncontrolled running sessions outside the laboratory. As a result, the potential limitations of machine learning performance when applying this approach to real-world scenarios must be carefully considered. For instance, in these studies, the data used for machine learning training were based on artificially induced and controlled affective running data, mainly derived from running at stages 1 and 3, where the target subjective fatigue states are either non-fatigued or fatigued. These two states only represented the two extremes of runners' fatigue spectrum, and they are relatively static. However, the participants' subjective fatigue reports gathered during Stage 2 were at intermediate levels and, more importantly, are variably and dynamically changing, although they were artificially manipulated into a progressive and linear increase pattern by researchers. Unfortunately, data from Stage 2 that correspond to

dynamically fatigued states were not of interest to researchers and thus not incorporated into the machine learning model. In fact, literature has shown that exertion and fatigue are not only dynamic but also fluctuating in reality, as demonstrated in [23, 29, 204, 217] Section 2.1. Consequently, this controlled focus may pose a challenge when applying such ARS models to address runners' affective state in real-life scenarios, as their training process does not accurately reflect the runners' affective experience in realistic contexts. Therefore, this thesis argues that there is a need to collect more natural affective running data and pay attention to the fluctuating dynamics of runners' SARE while collecting data and building ARS.

Additionally, these studies' fatigue reporting procedures were often researcher-assisted and involved some level of researchers' interference to participants' self-reporting while running. For example, in laboratory settings, such as study [46], researchers consistently reminded or encouraged participants to report their Rating of Perceived Exertion (RPE) while running on a treadmill. In study [128], participants had to step off the treadmill to stop running for 30 seconds to report their RPE to the researchers. Although some studies shifted their focus to outdoor settings, as in [203], researchers had to wait for participants to complete each lap before instructing them to report their RPE. Similarly, in [165], during outdoor runs on a track, a researcher cycled in front of the participants to control the fatigue protocol and self-reporting procedure. Such experimental arrangement and interference from researchers may have affected the ubiquity of collecting runners' affect related data. In some instances, these instructions disrupted the natural flow of running, leading to potential inaccuracies in the reported data.

In summary, this review highlights that the building of SARE-ARS in current literature is limited in terms of its ubiquity and granularity. Firstly, the research focus has been primarily on collecting fatigue self-reports as GT, neglecting other affective and emotional states. Moreover, the self-reported GT are typically limited to numerical or rating-based inputs, which may not capture comprehensive forms of information. More importantly, these GT labels are often sourced from induced, rather than

60

naturalistic, experiences, potentially leading to ARS that do not reflect the complex interaction between affective, social, and environmental factors that occur during uncontrolled running sessions outside of the laboratory setting. Additionally, collecting such self-reports in current research has been non-ubiquitous due to researcher assistance, involvement, and potential interference. Therefore, there is a need for further investigation to identify methods that can enhance the ubiquity and granularity of affective GT collection in a running context.

2.2.3 Ubiquitous Ground Truth collection using Experience Sampling Method (ESM)

Experience Sampling Methodology (ESM) is a widely used approach for collecting longitudinal self-reported subjective feelings of users in the wild, in conjunction with naturalistic behavioural data. This approach involves repeatedly asking participants questions about their emotions, feelings, or other types of experience within a specific temporal period; typically, these questions are sent out via smartphone notifications or text messages, enabling users to report their experience outside of laboratory environments[264]. This allows for the gathering of naturalistic behavioural and physiological responses that accurately reflect real-life scenarios, as opposed to acted or elicited expressions within lab contexts, ensuring that the developed automatic affect recognition systems are functional once deployed.

However, ESM researchers have also identified challenges of using ESM to gather self-report. First, due to the querying nature of ESM systems, users quite often need to physically interact and manipulate such systems to input the feedback or self-report. For instance, when a user receives the notifications from the ESM system on their mobile phone, they often need to stop performing the current activity, and take their phone out of the pocket, and then type the answer on their phone [91]. Such extra manipulation interferes with users' original activities, which not only has impact on the natural experience of the task they are performing, but also affects the data quality (e.g., task-irrelevant movement cause extra acceleration signals that might complicate the original sensor data [100]). Such physical efforts and burden caused

by the needs for manual operation of ESM system may affect the retention rate of using of such system, and therefore reduce the logging of self-report or ground-truth data [49]. Thus, such issues above could be more serious when collecting self-report in on-the-move contexts such as sports and physical activities (running, walking, cycling). This requires further research on designing ESM systems with consideration of contexts such as indoor and outdoor settings, social and personal environments, and activities (sitting, working, or on the move) [58, 102, 110, 280].

It is true that mobility and hand occupation can have an impact on people's experience of interacting with ESM systems/interfaces [280]. This could be more serious in the context of running. Runners need to pay attention to their surroundings and road conditions, which could make self-reporting to an ESM system more difficult and require additional cognitive effort. Therefore, using speech as a means of interaction could potentially be less obtrusive and allow for easier control of data collection while running. Additionally, speech interaction might help to reduce movement artefacts [100] that may arise from manipulating ESM systems while running, which is important for reducing noise in running movement signals. However, there is little understanding of the self-reporting experience using voice-based ESM during running, as well as how this influences the overall running experience.

2.2.4 Machine learning methods, features exploration and sensing technology for runners' fatigue

This section provides an overview of the machine learning models and features utilized to develop automatic recognition systems (ARS) for the subjective perception of fatigue in runners. Additionally, it reviews relevant literature in the sports domain to identify other potential features identifying fatigue and opportunities for sensing technology to tracking the relevant signals.

2.2.4.1 Overview of used ML methods and ARS results for runners' fatigue

This section outlines how machine learning models were used by researchers to predict runners' fatigue perception. As outlined in Table 2-1, Study [46] used an inertial measurement unit (IMU) sensor to capture motion signals and trained various models such as Random Forests (RF), Support Vector Machines (SVM), k-Nearest Neighbor (kNN), Naive Bayes on biomechanical features. Among them, the RF model yielded the best classification results based on the data collected from the runners' body parts such as the lumbar, right shank, and left shank with a global accuracy of 75%, 70%, and 67%, respectively, and a personalized accuracy of 89%, 100%, and 99%. Study [203] trained Gradient Boosted Regression Trees (GBRT), Artificial Neural Networks (ANN), Linear Regression with Elastic Net regularization (EN), Linear Regression with Least Absolute Shrinkage and Selection Operator regularization (LASSO) models on both statistical and biomechanical features derived from IMU signals. Among them, GBRT achieved the best performance using data from the wrist and it was found that using both Statistical Features (e.g., minimum, maximum, skew, kurtosis, and average absolute difference (AAD)) and Sport Science Features (e.g., Sample Entropy, Detrended Fluctuation Analysis (DFA), and Stride Regularity). Additionally, the study in [128] constructed both frequency and time-based domain features using a surface electromyography (sEMG) sensor to track muscle activity and a blood lactate sensor to monitor the blood lactate status, and trained a RF model. Study [99] used motion camera and ground reaction force to construct biomechanical features and trained RF, Bagging SVM, KNN and BDT models. Finally, study in [165]] used a motion IMU to construct biomechanical, statistical and spatiotemporal features and trained an RF model.

Such work also suggested that a decision tree based supervised machine learning approach such as RF was a feasible method to be used to explore the automatic recognition system of runners subjective and affective state, and with other literature suggesting that RF is more reliable than GBRT when training the data with a lot of

noise. Therefore, RF will be considered to be the priority to use this in this thesis when exploring the building of ARS.

2.2.4.2 Potential features exploration and sensing technology

This section review literature in sports science have shown the correlations between runners' subjective fatigue experience and the objective signals captured by sensing technology (See Table 2-2). This can further inform what information can be captured by sensing technology and used to as features for the machine learning process.

Sports science literature provides evidence that exhaustion and fatigue can affect runners' posture. For instance, Maas (2018) [157] reported that fatigue increases peak trunk flexion, trunk rotation range of motion (ROM), peak pelvic anterior tilt, pelvic rotation ROM, peak hip flexion, peak knee extension, and decreases peak ankle abduction in both novice and competitive runners. Additionally, Schütte (2015) Schütte, Maas [232] also demonstrated that running-induced fatigue causes deviations in dynamic centre of mass motion, specifically increasing mediolateral and anteroposterior ratios of acceleration Root Mean Square (RMS), the anteroposterior sample entropy of trunk accelerometry patterns, and decreasing the anteroposterior step regularity. Furthermore, Anbarian (2016) [21] showed that fatigue leads to an increase in the maximum pressure underneath the first, second, and third metatarsals and to a decrease in the maximum pressure on the fourth and fifth metatarsals, as well as a decrease in the maximum force and impulse underneath the small toes and fifth metatarsal maximum force. Moreover, their study further explored the difference between the effects on people with low and high arches, with fatigue shifting the load to the lateral side of the foot for high-arch runners and to the medial side for low-arch runners. An early study from Nardone (1997) Nardone, Tarantola [186] also proposed that fatigue increases body sway area and sway path, indicating that the whole body becomes unstable when fatigued. These findings serve as both evidence and inspiration for utilizing running kinematics features to classify exhaustion and fatigue. In this thesis, we will explore the

incorporation of such features while considering the availability of ubiquitous wearables for use during real-world running sessions. For example, we may explore the use of insoles equipped with pressure sensors to capture feet pressure signals. Insoles are less cumbersome since runners do not have to carry them. Additionally, we can investigate the use of smart bracelets to capture movement signals, as many runners already incorporate such wearables into their daily lives.

Table 2-2 Correlation between subjective fatigue and objective running features

REF, Year	Assessment of fatigue	Focus	Sensors	Features	Subjects	Lab & Wild	Results
[146], 2006	BORG score	Stride Pattern	Triaxial Accelerometer on the lower back (lumbar spine)	Stride frequency; Stride symmetry and regularity; Signal energy and impulse; Integral of total acceleration vector	Athlete distance runners	Track running	Acceleration Vector increases, A significant increase of medial- lateral impulse and the medial-lateral signal energy; A decrease in regularity (temporal- spatial periodicity of running cycle)
[21], 2016	BORG Scale	Feet pressure distribution	A foot scan pressure plate & software captures ten anatomical zones of the foot: medial heel (HM), lateral heel (HL), midfoot, metatarsal first to fifth (M1-5), and the hallux (T1) and lesser toes (T2-5).	maximum pressure, peak forces (body weight %), impulses (absolute force time-integral)	novice runners (running less than 2-3 times per week for <45 min or <10 km), but they had the ability to run at a self- selected speed for approximately 30 min at a time	running- induced fatigue protocol	increase in the maximum pressure underneath the 1st 2nd 3rd metatarsals ; and to a decrease in the maximum pressure on the 4th and 5th metatarsals, and to a decrease in the maximum force and impulse underneath the small toes and 5th metatarsal maximum force.
[157], 2018	voluntary exhaustion defined by reaching to at least 17/20 of the BORG score	Joint angles and global trunk and pelvis angles	10 camera VICON motion capture system with markers are placed on the trunk and lower limb, thighs and shanks, and shoe	peak pelvic anterior tilt, pelvic rotation range of motion, ankle plantar flexion,	Novice runners (weekly running distance was less than 10 km/week) VS competitive runners (at least 3 years of running experience; weekly training distance of 50- 70 km/week)	Controlled lab environment	Fatigue increased the peak trunk flexion, trunk rotation ROM, peak pelvic anterior tilt, pelvic rotation ROM, peak hip flexion, peak knee extension, and decreased peak ankle abduction. NOVICE runners showed larger kinematic adjustments when exhausted than COMP runners.

[232], 2015	voluntary exhaustion (at least 17/20 of the BORG score)	Centre of Mass (CoM) Motion	10 camera VICON motion capture system with markers; The sacral marker displacement method to represent the body CoM displacement	the tri-axial ratio of acceleration root mean square (RMS) to the resultant vector RMS, step and stride regularity, sample entropy.	novice or competitive runners who run at least 10 km per week	indoor treadmill with fatigue induces protocol	Running induced-fatigue increased mediolateral and anteroposterior ratios of acceleration RMS, the anteroposterior sample entropy ($p < .05$) of trunk accelerometry patterns, and decreased the anteroposterior step regularity
[22], 1997	More than 60% of the maximum heart rate Borg 0 – 10 scale of perceived exertion	Body sway variables (sway area and sway path)	force plates scanning	The position of the instantaneous centre of foot pressure (CFP) the mean position of CFP the sway path (SP) the sway area (SA)	young healthy subjects	Indoor lab setting	Fatigue increased body sway area and sway path and the Romberg quotient (the ratio of EC/EO of sway area, or sway path). Fatigue induced an increase in the median frequency of oscillation of the centre of foot pressure, independent of the amplitude of sway. The mean position of the centre of foot pressure (CFP) was unchanged after the exercise.

2.3 Voice-based Chatbots for collecting self-reports

The chatbot technology has garnered considerable interest and popularity in various research areas [81] because of its ability to engage in conversational interactions with users. In this section, we will examine how current research collect people's thoughts via thinking aloud and how chatbots have been employed for collecting self-reports.

2.3.1 Thinking aloud for collecting data on peoples' thoughts

Cognitive science studies suggest that using thinking aloud as a data collection method might affect individuals' experience and performance of their primary tasks [74]. As discussed in [83], during the process of thinking aloud, when individuals simply express thoughts that they can readily articulate or when they convert non-linguistic information into verbal expressions, the cognitive demand is limited. Hence, the impact on individuals' experience and performance of their primary tasks is minimal. In this case, data obtained from thinking aloud may closely reflect individuals' thought processes during the activity. However, when individuals are expected to incorporate explanations, elaborations, reasoning, assessment, and interpretation of their thoughts and actions during thinking aloud, it has the potential to introduce additional cognitive demands and divert focus from the task at hand, thereby impacting task performance and influencing task-related collected through thinking aloud.

HCI literature also shows evidence that thinking aloud might affect people's perception of time estimation possibly due to the cognitive efforts required by thinking aloud itself [103]. Consequently, data collected through concurrent think-aloud regarding people's time perception may not be entirely accurate. Ericsson and Simon (1987) [75] identified two types of think-aloud methods. Type 1 involves verbalizing thoughts concurrently as participants perform a task, while Type 2 involves verbalizing thoughts immediately after completing a task. Hence, if the

research were to be about runners' perception of time (e.g., estimating the duration of their run), retrospective thinking aloud may be a more suitable approach to more accurately capture runners' perception of time than concurrently thinking aloud.

Despite the issues highlighted above, some sport psychologists have explored the use of thinking aloud as a method for gathering data on athletes' thought processes during sports activities. For instance, [222] used think-aloud to gather runners' thoughts during distance running, and [197] used think-aloud to gather golfer's thoughts. Both studies reported that participants experienced no significant cognitive effort that negatively disrupts their sporting experience. In fact, think-aloud is often used as a cognitive strategy by professional competitive runners to monitor their internal bodily experiences, goals and managing running information to enhance their performance during real-world running races [192].

In addition, when employing thinking aloud as a research method, it is essential to consider two other factors regarding user behaviours or user experience. First of all, [223] revealed that even when individuals engage in thinking aloud alone, they tend to imagine a listener, whether it be a person or an object. Consequently, participants' perceptions of the addressee may influence their thinking aloud. Furthermore, Individuals may selectively choose what to report during the think aloud process, omitting certain thoughts or experiences. This selectivity introduces biases and limitations to the collected data, as not all aspects of their thinking process may be fully captured or verbalized [223].

Given the use and benefits observed in sport psychology and specifically in running in the use of think-aloud, in this thesis, we employ think-aloud as a method to gather self-reports from runners regarding their subjective and affective experiences. We will also consider its potential effects and the listener aspect when analysing such self-report data. In the following sections, we will delve into a literature review on chatbots as potential conversational agents and listeners during thinking aloud and reflective processes.

2.3.2 Chatbot for collecting self-report from runner

Thus far, there has been a lack of literature on utilising chatbot technology to collect runners' self-reports while running, both indoors and outdoors. An exception to this is an emerging proposal for using chatbots to enable athletes to self-report injuries in [20]. Although no actual chatbot was presented in this paper, this paper proposed that the Subjective, Objective, Assessment, and Plan (SOAP) method, a clinical approach used to assess injuries or musculoskeletal disorder. The 'S' component refers to the athlete's subjective description, such as impressions of the injury, the subjective perception of injury mechanism (e.g., the direction of force, sound, and sensation), previous injury history, provocation of symptoms (recreating the pain by performing movement), pain quality (burning, sharp, tingling, etc.), the region of the injury (localised or spreading out), severity (on a 0-10 scale), and timing (macro trauma or microtrauma). The 'O' component refers to the doctor's objective observations or inspection of postural alignment, signs of trauma such as *“gross deformity, swelling, bleeding, signs of infection, and atrophy”*. Furthermore, the 'A' component is used to make a diagnosis and the 'P' component is used to propose a training plan for the athlete. While this framework may serve as a useful guide for designing the chatbot's conversational flow for post-run injury reporting, its applicability for facilitating runners' in-the-moment self-reporting of their feelings during running.

The literature review in this section has shown the limitations in terms of collecting simple rating-based self-reports as ground truth for building SARE-ARS. In addition, studies heavily rely on experimentally induced fatigue, which does not fully encompass the multifaceted subjective and affective experiences encountered during long-distance running in natural environments. In the next sections, we will review relevant knowledge on how chatbots can be used as a more ubiquitous approach and serve as potential listeners and supporters to collect self-reports in more naturalistic contexts.

2.3.3 Chatbot for collecting survey responses

In the HCI community, Chatbots have emerged as a novel and appealing approach for completing surveys or questionnaires for self-reporting purposes, as opposed to traditional web-based questionnaires. Kim (2019) had already investigated the reliability and validity of chatbot to collect survey data. By comparing the questionnaire data collected via a chatbot and a webpage, they concluded that the questionnaire implemented on the chatbot with casual conversational style could improve user engagement as well as the data quality with greater diversity [129]. Draxler (2022) conducted a longitudinal study in which a messenger-based chatbot app was used to send out well-being questionnaires to gather students' wellbeing information during two years of the COVID-19 pandemic, demonstrating the chatbot to be a flexible and reliable platform for collecting survey data [70]. Inkster (2018) conducted a study in which they used a text-based chatbot app (called as Wysa) to deliver the Patient Health Questionnaire-9 (PHQ-9) over four months. Their findings showed that greater interaction with the chatbot-assisted self-reporting process led to an improved mood, promoting mental wellbeing and resilience [109].

Furthermore, research has explored how a chatbot can aid participants in self-disclosure while responding to the chatbot. It has been suggested a chatbot's prompts that includes information on its own feelings, information and thoughts could in turn encourage users to respond longer and more in-depth information with exposure of their moods, experiences, gratitude, and stress [147, 148]. This suggests that designing a chatbot to mimic human self-exposure behaviour may be an effective approach to promote richer self-report data related to mental and emotional state. Although, the chatbot's conversation structure in this study remains as simple conversation turns.

2.3.4 Voice-based chatbot for collect survey data

In addition to text-based chatbots, there is an emerging interest in designing voice-based chatbots as ubiquitous tools to deliver surveys and collect self-reports in indoor settings, with a focus on health and wellbeing topics. It has been argued that the utilization of smart speakers to administer voice-based surveys offers a hands-free and type-free self-reporting experience for the collection of survey data in an uncontrolled environment, resulting in greater survey engagement [272]. Next, we first present a quick summary and review of such work (see Table 2-3), followed by the critical analysis with respect to the gap the thesis aims to address.

Table 2-3: Voice-based Chatbot for Collecting Survey Self-reports

Study	Year	Chatbot Type	Conversation Interaction	Proactive Prompts	Required User Activation	Use Scenarios
[158]	2022	Voice-based chatbot	WHO-5 questionnaire with daily open-ended questions and bi-weekly Likert-scale questions	No	Invocation phrase <i>"Hey Google, talk to Sofia"</i>	At-home settings
[17]	2022	Voice-based chatbot	Fluid intake and sleep surveys with 3 and 11 questions respectively	No	Users initiate conversation by saying <i>"Hey Google, talk to Fluid Monitor"</i>	At-home settings
[135].	2018	Voice-based chatbot	10 direct reflection questions (one question daily, for two 5-day workweeks)	No	Press and hold the Dash's button and uttering the command <i>"Start Work Reflection"</i>	Work settings
[273]	2022	Voice-based chatbot	Four questions every hour between 9am and 10pm	Yes	Raspberry Pi used to trigger the chatbot to proactively invite participants to report	At-home settings

Maharjan (2022)[158] developed a voice-based chatbot, named Sofia, based on the Google Nest Mini 2 smart speaker, to enable participants with depression and bipolar disorders to self-report their wellbeing from home. The chatbot's dialogues were

constructed using the WHO-5 questionnaire, which included daily open-ended questions, such as *“How are you feeling today?”*, *“Please share more about your feelings,”* and *“Do you want to add anything else?”*, as well as bi-weekly Likert-scale questions, such as: *“On a scale of 0 to 5, 0 being 'at no time' and 5 being 'all of the time', over the last 2 weeks, how often have you felt 1) cheerful and in good spirits? 2) calm and relaxed 3) active and vigorous, 4) fresh and rested after waking up, 5) fulfilled with interesting things in life?”*. The chatbot was not designed to proactively prompt users to self-report. To interact with the chatbot and answer the survey questions, participants were required to initiate the conversation with the invocation phrase *“Hey Google, talk to Sofia”*. The chatbot then delivered the survey questions one by one, with each answer time restricted to a maximum of 8-12 seconds due to the technical limitation of using Google Nest. To ensure the flexibility and autonomy of the conversation flow when delivering the open-ended questions, the researchers adopted the Wizard of Oz technique to handle unexpected user questions group [159], [158] and [160].

Similar, Almzayyen (2022) constructed a voice-based chatbot for use on Google Home smart speakers to collect users' self-reports on fluid intake and sleep. This chatbot asked three questions (three times per day) for fluid intake: *“What is your participant ID?”*, *“On a scale of 1 to 5, how do you feel right now?”*, and *“How much have you drunk since the last survey?”*. Participants answered each question without additional conversational interactions, typically taking less than one minute. A similar format for conversation about sleep was used, directly incorporating eleven questions from the SHUTi sleep diary questionnaire (once per day). However, the chatbot was unable to proactively invite participants to report, requiring them to remember to initiate the conversation by saying, *“Hey Google, talk to Fluid Monitor.”* The evaluation of the interaction experience focused on metrics such as *“ease of use, likeability, and burden level”* [17].

Kocielnik (2018) developed a voice-chatbot, integrated with the *“Alexa Dash Wand”*, to facilitate users' reflection on their work, with 10 direct reflection questions (one

question daily, for two 5-day workweeks) in the conversation. Activation of the chatbot required the user to press and hold the Dash's button and uttering the command *"Start Work Reflection"*. Upon activation, the chatbot delivers structured questions to which users provide spoken reports [135].

Wei (2022) [273] employed a smart speaker to deliver a voice-based survey to participants in their homes, every hour between 9am and 10pm. The survey consisted of four questions: (1) How available are you on a scale of 1 to 5; (2) How bored are you on a scale of 1 to 5; (3) How would you rate your current mood on a scale of 1 to 5; and (4) What are you currently doing? Participants had 90 seconds to answer all questions. Unlike previous studies, this voice agent was designed to reduce the requirement for participant initiation and provide self-reports. A Raspberry Pi (placed close to the smart speaker) was used to play audio to trigger the smart speaker, allowing it to proactively invite participants to report.

After reviewing the related work on the use of chatbots for collecting self-reports, it became apparent that there were several gaps. Firstly, the scenarios of using voice-based or GUI-based chatbots for delivering surveys and collecting self-reports mainly focused on indoor or at-home settings at a fixed location, such as the voice-chatbot developed on the Google Nest Mini 2 smart for collecting wellbeing self-reports at home [158], the voice-chatbot based on Google Home smart speakers for fluid and sleep reports [17], the voice-chatbot designed for workspaces [135], the Google smart speaker at home [272] [273], and the GUI-based chatbot for stock investment at workspaces [216]. Collecting self-reports in outdoor and mobile contexts is necessary not only for the purposes of this PhD thesis, which seeks to collect runners' subjective and affective experiences during running, but more importantly, because human activities often take place on a large scale in outdoor environments and involve movement rather than remaining stationary. Consequently, this poses the question of whether a chatbot could be designed to serve the goal of collecting self-reports in such scenarios. This thesis contributes to this question by focusing on the running context.

In addition, these designed chatbots were still limited in their ability to proactively prompt to deploy experience sampling ubiquitously, despite being claimed to be able to collect self-reports in in-the-wild settings. Most of them were unable to proactively prompt users to give self-reports, instead requiring participants to expend extra effort to initiate the conversation. For example, participants in [158] were required to set up a reminder to remind themselves to speak and self-report to the chatbot, with a mixture of varied scheduling frequencies (e.g., daily and bi-weekly). Similarly, participants in [17] were instructed to talk to the chatbot at different times (e.g., 9am, 3pm, and 9pm). In addition, when give self-reports, participants had to first send a verbal request to the chatbot (e.g., *“Hey Google, talk to Sleepy”* or *“Hey Google, talk to Fluid Monitor”*). Researchers in [273] acknowledged the need to be proactive when conducting experience sampling, yet their proposed solution relied on the use of additional devices (e.g., a Rsperry Pie), which cannot be easily replicated in moving and outdoor contexts. Moreover, these studies did not try to break through the existing technical limitations of current smart speaker or conversational agent platforms, which are designed with a fundamentally response-based mechanism, where users have to send an initiation to express their intention, while the chatbot is designed to only be reactive. This is in stark contrast to the core principle of experience sampling, which stipulates that users should not be required to expend effort, but should instead be prompted to give self-reports without intrusion and with maximum ubiquity and immersion into their daily life. In the context of collecting self-reports from runners, it is especially important that participants do not need to remember the reporting task, as their primary goal is running. Therefore, the chatbot should be designed with the ability for proactive interaction.

Last but not least, despite the increasing interest in utilizing a conversational style for gathering self-reports, the design of most chatbots’ conversations remains identical to the original questionnaire questions, often with a formal, written language style that may sound too stiff and inappropriate when delivered through a voice-based interface. Literature indicates that, compared to web-based surveys, a chatbot with

a more casual and diverse conversation style may result in increased data diversity, with no additional benefit from a chatbot featuring a formal language style [129]. Furthermore, the conversation flow is limited to a basic form of interactivity, consisting of the successive execution of a series of questionnaire questions with little natural language comprehension of users' responses to create content-relevant answers or personalized follow-up dialogues. Therefore, the chatbot for collecting runners' self-reports in this study aims to achieve more natural and interactive conversations with improved procedural intelligence for context understanding, technical domain skills or knowledge, and managing turns [284].

2.3.5 Chatbot for reflection

The feasibility of leveraging chatbots to encourage self-reflection through the posing of reflective inquiries and questions has been being investigated. Kocielnik (2018) [136] investigated the efficacy of a text-based chatbot in promoting deeper reflection upon physical activity. The chatbot was designed with an initial set of questions that prompt reflection and a two-step mini-dialogue structure, which enabled dynamically tailored follow-up questions based on users' responses. For instance, the chatbot would refer to what users said to provide personally relevant responses. Results indicated that, following their interactions with the chatbot to answer reflective questions, participants reported increased awareness of their past activities, gained novel insights about their activity, and triggered future behaviours of adjustment (e.g., becoming more realistic and objective-minded when creating activity plans for the future).

Kocielnik (2018) also investigated the potential of a chatbot to facilitate reflection in work scenarios. Domain knowledge of work-related reflection from existing literature was used to design a set of dialogue scripts for the chatbot questions, which included *"task-related questions"*, *"short-term and long-term activities and goals"*, *"motivation and satisfaction at work"* and *"personalised questions"*. Results showed that, after interacting with the chatbot-guided reflection, participants had increased

awareness of their work progress and enhanced their memory of their work achievements [135]. However, the designed chatbot's questions were limited to simple interactivity, as it lacked natural language understanding to have follow-up questions or continue dialogues within each topic or provide more relevant responses.

In [284], four types of reflection were suggested when designing a chatbot, including (1) revisiting: the recollection of past events; (2) transformative: a positive experience; (3) explanation: memory triggers which can also solicit further explanation; and (4) exploration. They proposed that the chatbot's intelligence should incorporate emotional intelligence to demonstrate empathy, social skills, the protection of user privacy, and procedural intelligence (context understanding, technical domain skills or knowledge, managing turns) as well as accurate empathy (e.g., affirmation, non-judgmental, context aware). Their findings suggested that the chatbot's evoking questions (prompts and probing) provided motivation and triggered ideas for further thinking. More advanced interactivity (e.g., follow-up questions) could offer the potential for rethinking and help users to process their feelings (p.121). They proposed that guided reflection could facilitate reflection and exposure in more difficult situations, with a mechanical, highly structured exploration of an often complex emotional event. They suggested that the chatbot should facilitate meaning-making by scaffolding the thinking process, decision-making, and logic [284].

Reichert et al. (2022) [216] present an insightful example of designing a text-based and proactive chatbot to facilitate ongoing reflection during real-world stock investment tasks. Their qualitative findings revealed that interacting with the chatbot slowed down decision-making processes, providing users with more time for reflection and increased awareness of their own thought processes. The authors proposed three design strategies for the chatbot: 1) proactive engagement without leading the user; 2) revisiting previous stock investment decisions; and 3) enabling users to externalize their inner thoughts. During the investment tasks, participants were prompted with questions from the chatbot, including open-ended, scale-based,

and multiple-choice questions, in specific scenarios. The authors further emphasized the need to investigate how users would engage in reflective interactions in more naturalistic and realistic scenarios, where researchers are not present, and users may face real financial losses or investment failures. Additionally, the study suggests the benefits of incorporating multimodal feedback, such as visual interfaces with stock data visualizations, to complement the chatbot's dialogue and enhance the reflection and sense-making benefits. Overall, this study provides valuable insights into the potential of chatbot-mediated reflection during ongoing cognitive tasks. It underscores the benefits of slowed decision-making and the importance of exploring more realistic user scenarios while considering multimodal feedback to foster effective reflection and decision-making processes.

The studies above indicate that gaining a comprehensive understanding of the concept of reflection within a particular domain is essential when designing a chatbot to aid in reflection. Consequently, our research aimed to gain domain knowledge of mindful running experiences among runners, based on existing online resources and expert review.

Chapter 3 Methodology

3.1 Research Question and Methods

The literature review presented in Chapter 2 illustrated the need and opportunity to create Automatic Recognition Systems (ARS) that can sense and interpret runners' subjective and affective running experiences (SARE-ARS) to not only benefit individual runners but also support sharing. Constructing SARE-ARS through machine learning approaches requires the collection of runners' real-world SARE and running data. SARE is usually collected through self-report, which is the gold-standard approach for obtaining subjective experiences to be used as ground truth for such ARS. Such rationale is illustrated in Figure 3-1.

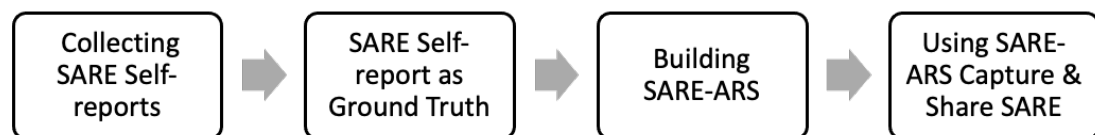


Figure 3-1: Overview research rationale

Research on voice-based experience sampling methods (ESM) systems and chatbots has been promising in supporting self-reports data collection, yet it has yet to be examined within the context of collecting SARE from runners. Thus, the primary research question of this thesis is:

RQ: How can technology be designed to support runners in self-reporting their subjective and affective running experience (SARE) while running?

The literature review in Chapter 2 has demonstrated the complexity of subjective and affective running exertion (SARE), illustrating that runners' subjectively perceived exertion and pain are not only affected by physiological factors, but also by social, emotional, and cognitive factors. Previous HCI research has suggested a transition from merely tracking running-related data to utilizing interactive technology that

supports communication channels which facilitate emotional cheering and social support. This has opened a space for further exploration in order to gain a more comprehensive understanding of the aspects of SARE that should be captured and shared, taking into account the complexity of the SARE. Thus, this thesis aims to address the first sub research question,

RQ1: What are the components of SARE that matter when sharing, and why?

In Chapter 4, a qualitative study was conducted to answer this question, involving a re-analysis of pre-collected semi-structured interview data, observations about long-distance running events (LDRE). LDRE was chosen as a specific context as it offers an opportunity to apply technology to sense and share the runner's SARE with their spectators as discussed in Chapter 2's literature review. In addition, this study was supplemented with further ethnographical studies that comprised the analysis of comments and reviews of digital apps developed for LDRE, as well as onsite observations of the London Marathon. Thematic analysis was employed to formulate overarching themes to address the research question. The results of this study highlighted the information that runners and their supporters want to be aware of, as well as the great importance of designing technology that can sense such information to enable its use and potential sharing. Given the results of the first study, the subsequent research question is:

RQ2: How can a ubiquitous ESM based on voice and wearables be designed to gather in-situ self-reports and running data for the purpose of creating SARE-ARS?

Chapters 5&6 build upon the findings from Chapter 4 and the understanding from literature to understand how data for building SARE-ARS could be collected in ecological setting. For this thesis, we focused on the self-reports of runners' perceptions of exertion, pain, and affective desire to stop or maintain running, and their emotional valence as they emerged to be very important from chapter 4. However, our affective states that could be also considered if SARE-ARS technology

should be deployed. To answer RQ 2, a mixed method approach was employed. Firstly, Chapter 5 presents the design of a voice and wearable-based experience sampling method system for runners (R-ESM) which aims to (1) expose people to the experience of self-reporting while running, and (2) explore the feasibility of using ARS infer affective and physical perceived states from sensed physiological and behavioural data.

In Chapter 6, a user study is reported where participants are asked to use the designed prototypes in their running sessions. Quantitative machine learning approaches were then used to validate the feasibility to build a SARE-ARS using such data. In addition, qualitative semi-structured interview was employed to investigate people's experiences of and openness to the use of such a system, and in particular, how they could actively contribute to the data collection (in particular, the self-reported ground truth) and passively to the sensed physiological and behavioural information during running. The results of this study suggested that, a fine-grained data collection of self-reported ground truth contributes to an ARS with accuracy rate well-above chance level. The qualitative results showed that it was acceptable and even useful for runners to use technology for self-report during running as it could lead to a better understanding of themselves to improve their running experience. However, it was suggested that a more conversational approach rather than a simple rating-based ESM would increase the quality of experience and effectiveness of the process. Hence, the next sub research question was:

RQ 3: How can chatbot conversation flow be designed to facilitate voice-based SARE self-reports while running?

RQ3 was addressed through a mixed-method approach. In Chapter 7, a prototype chatbot was created as a research probe to (1) expose people to the experience of having conversational interactions for self-reporting while running and (2) explore the feasibility of using a chatbot to collect self-reported subjective and affective states sensed while running. The chatbot's design and development followed an

iterative process, beginning with an initial thematic analysis of public available data (online articles about mindful running) to inform the design of the chatbot's conversational interaction. This conversation design was further refined through a design workshop with HCI experts passionate about running. The chatbot was iteratively refined and tested before it was used as a probe with runners for in-the-wild evaluation.

In Chapter 8, quantitative analysis of the conversation data between runners and the chatbot was employed to study the quality of the designed chatbot and the conversations that occurred during self-reporting tasks. Qualitative approaches were used to investigate the design opportunities of such a chatbot, including thematic analysis of semi-structured interview data of runners' experiences and needs. Participants acknowledged the benefits of being guided by the chatbot to engage in structured conversation for self-reporting, but also desired a space for flexible and relaxing self-discovery before conversing and reporting. This led to the research question,

RQ 4: What are the dynamic, reflective elements of SARE self-reports, and what should SARE-ARS model?

The aim was to address the gap in the literature on what should form the ground truth for useful SARE-ARS. RQ 4 was addressed through qualitative approaches by analysing think aloud reflections of runners. The chatbot prototype was refined based on previous results to integrate a mix structure of focused think-aloud moment triggered by the chatbot and structured conversation with the chatbot. Participants were recruited to experience the last chatbot prototype during their runs, and the conversation data obtained with the mixed-structure chatbot were analysed using qualitative approaches. The results revealed that the mixed-structure enhances self-reporting, with a learning, tuning and personalization effect. This research question was additionally addressed through supplementary analysis of other self-report data collected in previous chapters, resulting in a framework to guide the design of the

chatbot's conversation structure to provide richer descriptions of SARE together with runners' appraisal elements that clarify their SARE as well as their action intentions to address their SARE.

3.2 Participants

The study recruited individuals aged 18 years or older who engaged in running from the general population. The research did not limit participation based on running experience level. Participants were not asked to modify their typical physical activity routine but were asked to document and reflect on their experience during their usual exercise routine using a novel research prototype designed in this thesis. Participants were recruited from university, social media, and local running clubs.

All levels of runners were welcome to participate in the study. Participants were categorized into three types of runners based on their training amount and level of competitive experience using the standard in [188], including novice/beginner, intermediate, advanced and elite. Novice/beginner runners train for less than 10 hours/week and compete in fewer than three events per year. They may be new to running or have limited experience and require guidance on proper form, training schedules, and injury prevention. Intermediate runners train for less than 10 hours/week and compete in four to six events per year. They have some experience and are looking to improve their performance. Advanced runners train over 10–15 hours/week and compete in more than three events per year. They have been running for a while and have a solid training base, specific goals, and multiple race experiences. Elite runners have achieved a high level of performance and often compete at a professional level.

3.3 Ethics

This thesis has been granted ethical approval by UCL to conduct human-centred research aimed at designing technological tools that support engagement in physical activity for individuals with chronic pain and fitness activities for the general

population. The study is part of a larger project, H2020 FET EnTimeMent, funded by the European community, which aims to develop new movement sensing technologies to better define movement at different time scales. Documents in Appendix 11.1 provides comprehensive details regarding study activities, what was involved, how data was collected, stored, and used, how participants were informed, and how consent was gained. Safeguards were implemented to ensure the safety of participants. Participants had the freedom to determine the extent of their exercise and take breaks when necessary and bring less changes as possible to their normal routines. Changes requested were that they would use only safe place for running. The study required participants to perform their exercises in a safe environment, such as during the day, in parks, and with other people present. The study team strictly adhered to all necessary COVID-19 guidelines and protocols. In addition, it was clarified to participants, that the study did not focus at all on performance but aimed to gather insights on how the novel research prototype supported engagement in physical activity and could help understanding the experience of running.

3.4 Analysis Methods

This thesis employs mixed-methods analysis in various chapters. Thematic analysis is the primary method for qualitative data and will be introduced here, while other methods will be discussed within each respective chapter.

Reflective note taking after the interviews and transcription of the interviews was done by the researcher. Thematic analysis was primarily followed the six stages outlined by Braun (2016) [43, 44]. The first phase involved familiarizing the researcher with the data by transcribing it using Nvivo and noting the emergent ideas. The second phase entailed generating initial codes, which could be accomplished through an inductive, deductive or hybrid approach. The inductive coding approach gives full consideration to all elements of the entire dataset, allowing for a richer and more nuanced analysis. Alternatively, the deductive coding method only focuses on

extracting the most relevant information of research interests and is normally facilitated with a pre-defined guidance, such as a codebook or set of guiding questions. For this thesis, a hybrid approach was adopted to use a set of questions for coding guidance while simultaneously retain an open-mindedness, allowing for more comprehensive and robust findings for the thesis.

A reflexive approach was employed to ensure a consistent coding process across the entire dataset. This involved developing and reviewing a codebook before coding each new data item in order to avoid forgetting past codes and making timely adjustments to prevent interchangeable or redundant codes. This reflexive approach also included multiple rounds of coding. The first round aimed to code all interesting information with both semantic/descriptive and latent labels (with a focus on semantic codes based on the initial understanding). Subsequent iterative rounds of coding entailed rereading and revising codes and extracts to generate more latent codes, thereby facilitating the development of more overarching themes.

Phase 3 attempts to uncover more comprehensive and higher-level patterns that encompass the various codes and the full dataset. This development process is reflective, beginning with the clustering of codes into groups or subgroups using visualization techniques (e.g., mind map). This is followed by iterative revising and refining of the code clusters until provisional themes are formed, with some leftover codes that cannot be included. The coded data is then analysed to determine whether it fits into these provisional themes, after which the themes are evaluated to determine their applicability to the entire dataset, until robust and nuanced themes are achieved.

In Phases 4 and 5, the provisional themes proposed in the previous phase were reviewed and refined in consultation with the team of supervisors, in order to ensure their coherence, consolidate overlapping themes, and divide complex themes into logical groups. Finally, the final themes were established and named, and the analysis results was drafted.

Phase 6 entails the author and his supervisor panel reaching consensus on the final themes, followed by the organization and presentation of the results as written findings.

In the next chapter, we present the study that addresses the RQ1.

Chapter 4 Understanding the need for capturing Subjective and Affective Running Experience (SARE)

4.1 Introduction

The primary objective of this thesis is to investigate how technology can be designed to assist runners in self-reporting their SARE. This chapter specifically addresses the first sub-research question, RQ1: ***What are the essential components of SARE for reporting, and why?*** By answering this question, we aim to gain a more comprehensive understanding of the aspects of SARE that should be captured and shared.

To achieve this objective, we have selected the Long-distance running events (LDRE) as a special context to understand the complexity of runners' SARE and the need for automatic recognition systems to recognize them. As highlighted in the literature, LDRE provides opportunities for both runners and spectators to use ARS for capturing SARE for social needs.

This chapter entails a series of qualitative analyses, including a re-analysis of pre-collected qualitative data and an analysis of newly collected qualitative data. The following sections will present the details of the methods used, the results obtained, and the subsequent discussions.

4.2 Method

4.2.1 Data Collection

Re-analysis has been established as an efficient and cost-effective method in qualitative research [51], and we utilized this approach for our study. We drew on pre-collected semi-structured interview data originally gathered by the PhD candidate during his MSc research project [36], which explored how runners and spectators can co-create a connected experience during long-distance running events (LDRE). It should be noted that this secondary analysis (supported by extra data collected for this thesis) led to a more in depth understanding of SARE as provided in

the MSc thesis and it's the opportunity offered by its' use by runners and others. The aim of this study is to provide context for and rationale for the need and opportunity that SARE-ARS could offer and as such create the space for the subsequent and main research questions of this thesis.

The specific context of long-distance running events presents unique opportunities for runners and spectators to use ARS for capturing SARE, which makes this case particularly relevant for the thesis. This pre-collected data is an ideal resource to address the first sub-research question, RQ1, as it contains valuable information about runners' SARE, including what needs to be shared, when, how, and why. The interview questions used in the original data collection process were well-suited for our study, asking participants to reflect on their experiences of memorable moments during long-distance races and their thoughts, feelings, and actions during those moments ([36], page 5). We believe that analysing these responses can offer important insights into the key components of SARE that should be reported for establishing a ground truth that can be used to build an effective SARE-ARS system.

Additionally, the in-person onsite observation approach was used to obtain an in-the-moment understanding about runners' and their spectators' experience during LDRE. Building on the understanding achieved in the MSc project, an additional field observation of the 2018 London Marathon (the largest marathon event in the United Kingdom) was conducted. The observations were accompanied by note taking, following a structure with three components: settings (date, time, subjects, location, and context of the race), observed conversations and behaviours, and the researcher's interpretation. An unsystematic documentation approach [19] was used in the observation to record whatever was emerging.

Besides, publicly available online data such as magazine articles about marathon experiences and user reviews of marathon applications were also reviewed to supplement the research question and thematic analysis in this study.

4.2.2 Data Analysis

Overall, the pre-collected semi-structured interview data comprised responses from twelve participants (four female, eight male; aged 20-45), including five spectators and seven amateur/beginner runners who had participated in either a full marathon (n = 10) or other long-distance running event (n = 2). The interview transcripts included participants' responses to questions regarding their experiences and memories of the LDRE, as well as the type of experiences they desired to share with each other. The transcripts also contained supplementary materials such as participants' annotations of their memories of the LRDE. The pre-collected qualitative data in the public domain included 10 articles of LDRE experiences (selected using criteria of 1) about marathon event experience, 2) approximately more than 500 words. Additionally, it included user reviews of eleven selected LDRE applications in the App Store⁸ (comprising 89 records, approximately more than 40 words per record). Furthermore, the qualitative analysis also built upon the observational notes about two onsite marathons (24 records, approximately more than 100 words per record).

All data were analysed using thematic analysis in accordance with the procedure in [44] with a hybrid approach of both inductive and deductive methods [43]. The inductive analysis involved with following in the literature review's result, the Subjective and Affective Running Experience (SARE) Framework, to guide the analysis of the components of runners' and spectators' experiences and their interactions during long-distance running events. The deductive approach involved with the researcher being openminded for any other potential concepts during the coding and theme forming process.

⁸ <https://www.apple.com/uk/app-store/>

The identified themes were discussed with supervisors prior to finalisation and reporting. The findings section outlines the themes, with quotes from interviewees (Ps# for spectators and Pr# for runners), online articles (B#), and app reviews (R#) represented accordingly. For the sake of data privacy protection, quotes taken from online articles and app reviews were paraphrased.

4.3 Results

In this results section, we first present the LDRE experiential structure: the Dynamic Physical, Emotional, Cognitive, Social (DyPECS) Framework. This framework provides an overall understanding of the experiences and interactions of R&S during LDRE, mainly derived from observations and online article analyses that have been supplemented with interviews and literature. Then, six themes will be presented to provide an in-depth understanding of these experiences and sharing, derived from a synthesis of all qualitative studies conducted.

4.3.1.1 The LDRE experiential structure: Dynamic Physical, Emotional, Cognitive, Social (DyPECS) Framework

The analysis identified an initial structure, termed as the DyPECS (Dynamic Physical, Emotional, Cognitive, Social) framework (Figure 4-1), to characterize the complex experience of a marathon from the perspectives of both the runner and the spectator. This framework consists of an outer layer (OL) and an inner layer (IL). The OL encompasses the phases of the journey and the four components, physical, emotional, cognitive, and social aspect, (4-C) of the runner and supporter's (R&S) experiences and interactions. The IL captures the dynamics of the R&S' physical micro-encounters during LDRE and highlights the complex dynamic nature of support and interaction between the two stakeholders. Furthermore, it facilitates an understanding of the process beyond the delivery of cheering which is currently employed in LDRE technology. These layers are further described below.

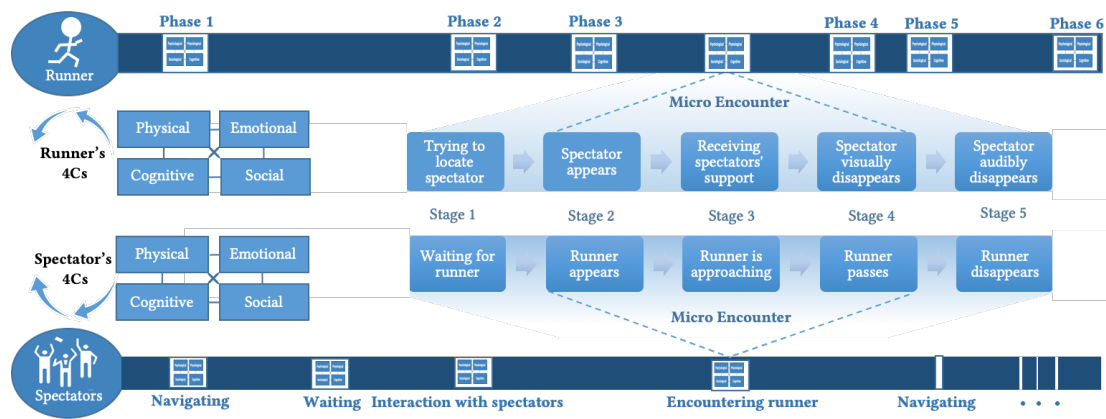


Figure 4-1. DyPECS framework: the outer-layer (OL) Runner and Spectator journey with their 4-Cs and the inner-layer (IL) micro-encounter stages

OL Runner's thread: Analysis of online articles revealed that runners often use an incremental distance structure when describing their long-distance running experiences. For example, a full marathon event could be divided into six phases based on distance milestones: 0-5KM (Phase 1), 5-10KM (Phase 2), 10-20KM (Phase 3), the half-way point (Phase 4), 30-40KM (Phase 5), and the last 2KM or finish line (Phase 6). The proposed DyPECS framework only adopted the phase concept (i.e., Phases 1-6) rather than specific incremental distances. This could accommodate running events of different distances and takes into account that the order of the phases may vary or even not occur at all, depending on the runner's capabilities and 4-C state on the day. The phase-based structure helps in the systematic analysis of long-distance running experiences, aiming to capture the main differences among the phases in terms of changes in 4-C (and their interactions) as the journey progresses, rather than every aspect of it.

This phase-based structure also facilitated the analysis of interview data by structuring and visually representing significant experiences during the marathon on a map. For instance, Figure 4-2 (left) depicted and structured Pr12's marathon experience, displaying details of their four-C experiences of emotions (e.g., joy, fear, doubt, and pride in red), physical sensations (e.g., heat and pain in blue), cognitive

strategies (e.g., adjustment of plan in brown), and social interactions with spectators (e.g., in green).

Outer-Layer (OL) Spectator thread: The spectator's thread of DyPECS emerged mainly from the observations study confirming the phases identified in the literature on active spectator experience during other types of long-distance sports events [113, 114]. Spectators plan how to move from one spot to another, wait for their runner to appear, support the runner both emotionally and physically as they pass, and then plan and move to another spot. During idle moments, they interact with other spectators or appreciate the atmosphere. Thus, the 4-C model also applies to spectators: the annotated map from Ps11 (Figure 4-2 - right) that offers examples of multi-faceted spectator experiences, including emotions (in red), physical effort (in blue), cognitive planning (in brown), and social interactions (in green).

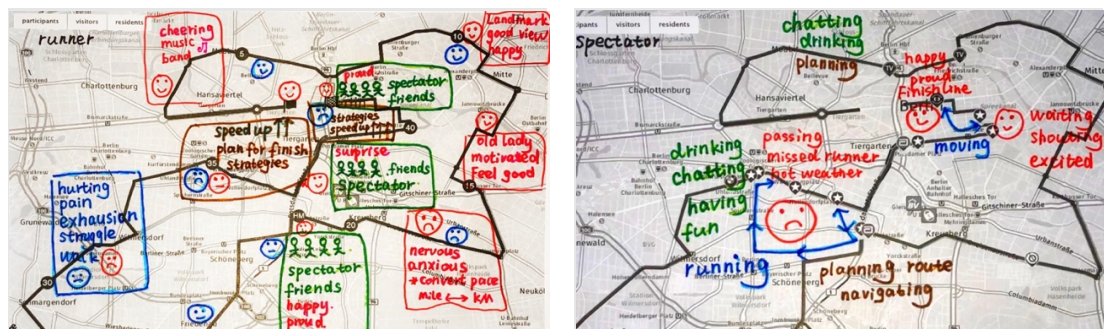


Figure 4-2. Runner's Experience Map (left), Spectator's Experience Map (right)

Inner-Layer (IL) Runner-spectator micro-encounters: The inner layer (IL) of the DyPECS captures the structure of the R&S' physical micro-encounters, which includes 5 consecutive stages emerged from both observations and interviews:

Stage 1 – Searching for Each Other: Spectators are “awaiting the runner” to appear and runners are “attempting to locate their spectator” in the crowd, with excitement mounting.

Stage 2 – Connecting: Spectators may attempt to make contact with runners by shouting or waving to gain their attention; however, the physical state of the runners may impede their ability to process external information. Moreover, the density of the crowd, particularly close to the finish line, may also obstruct the connection.

Stage 3 – Passing By: The connection is established. As the *“runner is approaching”*, anticipation increases and excitement heightens. The spectators actively strive to interact with them via talking, clapping, and hugging.

Stage 4 – Separating: The *“runner has passed by”*. Spectators still catch sight of the runner from behind and continue to cheer in order to provide motivation. For runners, the *“spectator has visually faded away”* from view.

Stage 5 – Disappearing: The *“runner fades away”* from spectators’ sight and the *“spectator audibly disappears from the runner’s mind”*. Spectators keep cheering until they do not see the runner anymore and the excitement drops.

The DyPECS structure is employed in Part 2 to construct a more comprehensive comprehension of the 4-C of R&S, and their desire to share experiences. *“DyPECS: OL-R:P#”* refers to the Outer-Layer (OL) Runner-Thread phases, *“DyPECS: OL-S:P#”* to the OL Spectator-Thread phases, and *“DyPECS: IL”* to the Inner-Layer (IL) micro-encounter stages.

Next, we will describe the six themes that emerged from the analysis characterizing the runners' SARE and spectators' experiences. Spectators' experiences are included to emphasize the importance of having automatic recognition systems to track SARE.

4.3.2 SARE: Psychophysical beauty and pain

Similarly to findings from the social phenomenology literature for athletes [12-14], but absent from the HCI literature, runners in this study described the beauty of running.

Runners often report a range of physiological sensations during long-distance running events (LDRE). These include auditory, haptic, and thermal experiences such as panting, wind crossing the face, and a warm chest, burning lungs, and dry throat when ascending hills. As Pr2 noted, *“When this (burning) sensation lessens, there is a pleasant mixture of heat and cool, as well as fresh air.”* Pr2 further suggested that these sensations can create a *“chemical mixture of adrenaline and endorphins”* that is particularly enjoyable in the second and third phases of the marathon (DyPECS: OL-R:P2-3), when the effort of running is still manageable. Moreover, runners in this study showed that often seek to share these sensations as a part of the beauty of their running experience with spectators.

The qualitative data revealed a commonly discussed subjective experience: the correlation between runners' mind and the physiological chemistry of the body when *“hitting the wall”*. B2 noted that the physical transition from glucose to fat burning can be quite uncomfortable and likened it to a hangover sensation, although the feeling does not last for long. The mental state may suffer a lack of motivation to continue running (e.g., DyPECS: OL-R:P4-5), which can have a negative impact on the remainder of the run if the runner has not trained themselves to transition between energy resources. Pr2 also confirmed that it is quite common for those that attempt to push themselves beyond their training to experience a loss of motivation and confidence.

4.3.3 SARE: Runners' cognitive strategies

Both associative and dissociative strategies (as discussed in Chapter 2) were observed throughout the running journey (OL-R:P1-6). Both strategies can be employed to maintain motivation during a marathon. Pr6 described a purposeful dissociation strategy, in which they associated *“a person or something to each of the last few miles”*. Additionally, the atmosphere of the live event can reduce the sense of boredom and fatigue, resulting in a passive switch to the dissociative state. Pr5 remarked: *“Every several hundred meters, there were some music bands, music,*

drums playing [...] it helps you ...". Furthermore, an associative strategy of actively attending to one's body in order to regulate form and manage pain (e.g. OL-R:4-6Phase) is an essential skill to maintain motivation during a marathon. As Pr6 explained, *"I started to feel a little pain on my hip and I realized I started to run a bit differently[...]. I started to operate my strategies. As soon as I focused on the run, my hip got better."* These strategies differ from those of elite athletes as suggested in literature (e.g., mental computation and motor imagery). The interaction between these strategies and other components of the experience is important to consider, as will be discussed in the following themes.

4.3.4 SARE: Runner's Emotions and 4-C interactions

Whilst the sports literature has mainly investigated emotions felt before and after a run, the analysis results show that runners experienced a full dynamic emotional journey with these emotions interacting with the other components and behaviors:

The run was an emotionally charged journey. At the beginning of the race (DyPECS: OL-R:P1-2), many runners were filled with excitement, surrounded by the event atmosphere and buoyed by the support from family and friends. They reported feeling relaxed, confident, and eager to compete (e.g., enthusiasm for overtaking other runners was documented in B1). By the end (DyPECS: OL-Runner-phase 5 or 6), the pain and difficulty were dominant, but the sense of pride and accomplishment that came from having overcome and completed the race was palpable (e.g., feeling exhausted, but with a feeling of accomplishment was shared in B2). Furthermore, the social aspects of the LDRE had a profound emotional impact on the runners. When they encountered familiar spectators (DyPECS: IL), they felt energized, as evidenced by Pr3 who adjusted his running performance in response, saying *"It was a pleasant surprise. I waved to them and began to increase my speed and energy."* In addition, runners were delighted by the sights they encountered: people dressed in costumes, breathtaking scenery, and high-fiving with children.

The results of this study highlight the interaction between physical and emotional factors as indicated reported in the literature [180], particularly during the more demanding phases (DyPECS: OL-R:P4-5) when physical resources are depleted and the finish line is still distant. Negative emotions such as exhaustion, desperation, self-doubt, and lack of confidence were observed to arise, which participants collectively referred to as *"hitting the wall"*. B2 described a downward emotional spiral where one begins to doubt themselves, question the point of what they are doing, and start to find excuses for not finishing the run. In such emotionally vulnerable states, many runners sought understanding and support. Pr4 noted, *"My legs were extremely heavy. I couldn't run anymore. I wanted someone who could talk to me, who could understand my terrible situation."* Even when not explicitly stated, some runners sought encouragement; Pr3 commented, *"If there's some technology [without the runner asking], [supporters] can shout messages out via an app to say 'Keep going' to motivate you when you are in pain."*

Social support and simple cheering are not always beneficial: Contrary to the literature [64, 219], social support did not emerge as a simple process. As opposed to runners who sought encouragement at difficult moments, some participants in our studies felt frustrated and misunderstood by the crowd or even by their friends cheering. Pr1 stated: *"There were always some people who shouted 'keep going, keep going'".* They didn't understand my real condition that I was in pain. [...] I felt very annoyed." Cheering by spectators encouraging them to run faster irritated runners who needed to adjust their pace (utilizing associative strategies) according to their physical state and pain. Pr2's friend was running with him and excessively cheering when he was struggling: *"He (Pr2's friend) attempted to be cheerful as I was faltering. He said 'Come on! We can (do it)! He was dragging me along. But I said 'No, it's not going to work. I don't want to!' I shouted at my friend. I was feeling angry."*

In some cases of extreme difficulty, runners might lack the mental resources to interact with spectators (and need to shift to associative strategies) and their support was no longer enjoyable or easy to process. Pr6 stated, *"I'm tired. [...] It was all about*

[...] going deeper into my little corner in my own mind. I was no longer enjoying the spectator crowd as much. I actually moved away from the people so that I didn't have to deal with that." In extreme cases, runners could injure themselves in trying to meet spectators' expectations. Pr1 commented: *"A runner was trying to high five with the kids ... in the crowd. And after that, he seemed more motivated and tried to speed up, but suddenly he cramped and fell over"*. Spectators' misguided cheering was likely due to their lack of awareness of the runner's (Emotional X Cognitive X Physical) state. Pr2 said of his friend, *"... he knows a lot about running, the wall stuff, but he couldn't feel this (Pr2's feeling) because he hasn't done the whole running."*

4.3.5 Spectators' emotional journey: needs for knowing about runners' SARE

According to the data, spectators have formed a strong emotional connection with their runners. They view themselves as supporters of the runners and feel as though they are running alongside them during the whole journey. Unlike the literature, our study's spectators recognize that support goes beyond simply motivating the runner. They desire to connect with the runners in various ways, such as understanding when the runner is facing difficulties or feeling stressed, as well as knowing the best time to offer mental encouragement through cheering or physical support via energy supplements. They understand that running is a complex experience and to truly support their runners and fulfill their role as supporters, they need to also consider the runners' cognitive needs. For instance, Ps7 stated, *"If she needs water, I can prepare it. It would be helpful to know when she is tired or feeling doubtful so I can encourage her."*

Knowing and sharing the runners' goal: spectators talked about knowing the physical and emotional effort put into the training by their runners: *"I really feel very emotional cos she has been working hard, running and training a lot. She said she want to finish within two hours. I was so emotional she has really done it. For me, it's important to see her (at the finishing line), the way she looks when she's gone through, ...she just achieved that ..."* (Ps7); *"Great to see him after the finishing line*

with the medal, that's the highlight [...] it's the months and months work," (Ps10). They described themselves as mentally running with their runners with the shared goal of finishing the race. For example, Ps8 said, "You stay here for a long time, you mentally run with them, you know how much work they put it in to prepare for it".

Spectators' Investment (enabling, properly supporting, effort during both training and LDRE): The spectator's investments of time and preparation both during the marathon (DyPECS:OL-S-navigating & waiting) and already started during the months of training contribute to build up the sense of involvement in the marathons (and its preparation) and lead them to appreciate the effort of their runner. Ps11 felt more involved when seeing the runner approaching the finish line by having been part of it for long: *"Great to see her in the finish line. You might be tenser and so much more involved with the experience of finishing and success as you have already emotionally invested months before you hear about it. You've already been a part of it".*

The tension of planning for the best meeting spot: The spectators' journey clearly emerged since planning for the first spot to see the runner (DyPECS: IL-stage 1). Spots are often pre-planned with runners but also decided on the moment by using tracking apps and rough estimates of the runner's speed. App review data shows spectators like the app tracking function for this. R26 commented, *"Being able to track runners and see the leader board is brilliant..."*. As they wait for the runner (DyPECS: IL-stage 2), Ps7 said, *"We have the app that can get measurement of her location every five miles. I kept looking at the runners [...], see if I could spot her. [...]I felt quite tense, quite nervous, thinking 'she's going to come through...'".* At the same time, planning was colored by nervousness also due to the lack of precision of the tracking device. Ps11 said, *"It was difficult to tell what the exact time she would come."* This was also confirmed by app reviews where most complaints were about this lack of tracking accuracy. R36 commented, *"The live tracking said he had finished the race when he hadn't, which meant I missed cheering him on. Spoiled my day."* For some, unpredictability itself was a part of the event. Ps8 enjoyed the nervousness: *"after the marathon, I feel I enjoyed this feeling because it was an element of the experience."*

Runner passing moment: The runner passing by (DyPECS: IL-stage 3) is a brief and only moment where R&S can interact. Attracting the runner's attention was considered a pleasure. Ps10 said, *"I'm excited and firstly trying to get her attention to see me. I waved to her straight away"*. Observational data of spectators' facial and verbal expressions showed that spectators were emotionally aroused even after the runner had just passed (DyPECS IL: runner passed). Spectators tried to extend the connection with runners. Ps7 said: *"I still kept shouting her name and trying to give her some boost as much as I could. Even though I [knew]she might not hear that."* As the runner disappeared (DyPECS IL-stage 5), spectators chatted and exchanged their experiences, covering topics including runners' face, gestures, and movements. Ps11 described, *"I was quite happy because she looks comfortable. She looks like she is doing well."*

Missing the Runner: However, there were situations when spectators were disappointed by missing the runner (lack of DyPECS: IL-stage 3). Ps10 missed the runner because *"there are so many spectators, it is very difficult to get a good view. I know she was coming but I couldn't get into the people to see her. So, we missed"*. R67 missed the runner because the moment is too brief, *"I watched the coloured line of my friend on the app get closer, with my camera poised to take her photo ... and then to my surprise, she passed me. I missed the chance."* Spectators were also upset when they were not recognized by the runner as the runner was focusing on running or did not manage to locate the spectator. Ps7 described, *"I did shout to her, and I waved my hands. But I don't think she saw me cos she had headphones in. She was too focused on the run [...] I was a little bit disappointed."*

Role recognition by the runner: As for the sport psychology, strengthening of the team-identity is critical to reach fulfillment of the event. Our findings suggest a different type of element critical to the strengthening of the runner-identity: recognition of supporting role. Spectators needed to have their effort and support recognized by runners. *"It would be nice for her to see me there. So, she knew I was there to support her at that point."* (Ps7).

4.3.6 Difficulty of sharing SARE VS desire to share SARE

The literature emphasizes the importance of social support; however, it underplays the sharing of running experiences with a cheering service that can be switched on and off [276]. This section's findings depict a much more nuanced process. As seen in the previous subsections, both runners and spectators are eager to share and know in their interactions. Runners want to share in order to receive support, but also to reward the spectator. As Pr12 stated, *"If your friend is travelling to Berlin because you run a marathon, it is the best way to engage them and thank them for being there..."* Additionally, spectators want to comprehend the runners' needs (4-C) so as to better support them, and to some extent, share their own 4-C in order to have their supporting role recognized. Nevertheless, sharing is a complex process, and a few barriers were identified as specific to this sharing activity.

Runners experienced difficulty in conveying their experiences to spectators during an event, despite their eagerness to do so. Expressing their complex, real-time emotions was particularly challenging. As Pr2 noted, *"It is easy to talk about things such as 'my calf is hurting' or 'I need water', but it is difficult to express the emotional aspect."* Pr2 further commented, *"I do not know how to make non-runners understand your feelings, such as the sense of accomplishment of getting to the top of the hill."* Additionally, it was difficult to communicate information while running, as Pr5 stated, *"It requires too much mental energy and you do not really want to have a conversation; you must focus on the race."* The only opportunity for runners to interact with their spectators was when they passed them (DyPECS: IL-Micro-encounter). Yet, due to the short duration of this moment, they could only utter brief comments such as *"I am doing okay"* or *"I am feeling good"* (Pr2). The observation notes also highlighted that non-verbal behaviors, such as waving, smiling and holding the head and chest up, were employed for conveying emotional states, either consciously or unconsciously.

Some runners chose not to express the extent of their pain to spectators, particularly their parents, so as to avoid worrying them. Participant 2 commented, *"If I told them it was really bad, they would be scared."* Rather than their pain, what the runners sought to share was their need for support. Participant 4 stated, *"I didn't want to let them know about my pain, but I still wanted to receive encouragement from them whilst I was in pain."*

Many runners expressed a desire for using annotated experience maps (e.g., Figure 4-2) to document their running experiences and tell their running stories. Pr2 commented that *"it is definitely amazing when you share your own running experience with friends, as it contains interesting information that does not need to be recalled by oneself. It is easy to tell the story."* Pr6 added, *"It would be great if the app could present a map like this. It is like a record of my memory and I would likely forget some details a few days later."* Some runners thought that maps were useful for comparing experiences (not just performance) as they help to understand and remember previous runs, thus aiding in improving running strategies. Pr12 noted that, *"If I had such a map, I would look back again to see where I did well or not, to adjust my strategies for my next race. The emotion or the pain can more directly show my situation when compared with basic speed data."*

Spectators also enjoyed reading the runner's experience map, delving deeper into their experiences. Ps felt more involved, *"I must confess I thought I knew everything about [the runner's] marathon, however reading the map, I've learnt so much more and feel I've been given a close-up view of the day [the runner] lived, one of torment, euphoria and celebration!"* (Ps11). Additionally, spectators expressed their dissatisfaction with existing commercial applications, and desired to have a live experience map during the race, in order to better support the runners. Ps9 commented, *"All we can know are the checkpoint times."* Ps11 added, *"[using the annotated map] I could give her some motivation if I saw her emotion on this map, especially when she was struggling."*

4.4 Discussions

In this section, we present a discussion of the emerging results and their implications. The analysis results presented in this chapter differ from the findings of the Master's project in several ways. Firstly, we developed the DyPECS framework to provide a more comprehensive understanding of the dynamics of multi-faceted SARE and sharing needs during long-distance running events (LDRE). Moreover, this section (see 4.4.1 and 4.4.2) discusses how these findings could contribute to answering RQ1 and inform the potential components of SARE that are essential to be collected as ground truth for the development of SARE-ARS.

In addition, we also explore other design implications that emerge from our results. Although they are not directly related to RQ1, they provide potential scenarios in which SARE-ARS can be incorporated and applied.

4.4.1 Essential SARE elements

This section discussed the findings present above to answer **RQ1: What are the components of SARE that matter when sharing, and why?** As evidenced by the findings, runners often seek to share the discomfort they experience while running, such as burning lungs and dry throat during ascents. Although these sensations may be not perceived comfortable, they were seen and shared as part of the beauty of the running experience. Pain and hardship are primarily experienced by the runner, yet they are often shared as a sense of pride upon successful accomplishment. Runners wish to demonstrate their mental strength and physical endurance to their spectators as a way to engage spectators.

Runners often seek emotional support in moments of desperation and self-doubt that arise during periods of strenuous exertion, such as the depletion of energy or the urge to stop. They long to have their experiences understood and to be supported emotionally in these instances. Sometimes, it is not necessarily about sharing the pain

and hardship itself, but rather about sharing the moment wherein the runner experienced pain and requesting emotional support.

It is essential for technology to recognise moments of high exertion, low motivation, or extreme emotional states among runners. These runners may have difficulties communicating due to fatigue or the focus that is required while racing, making it difficult to articulate their emotional needs. Additionally, they may be in pain and not wish to interact, but instead remain focused on their bodies and remain in an associative state rather than engaging with external stimuli while suffering. Furthermore, runners may desire to be understood in terms of their physical conditions, including any pain or energy levels experienced, particularly during extreme circumstances. Without recognition of these situations, spectators' interactions with runners can become deleterious to runners.

Consequently, there is a need for technology to enable runners to express and share their emotional states, potentially through the design of an automatic recognition system (see 4.4.2). Such a system could improve communication between runners and spectators, providing support and motivation in moments of pain and low motivation. With this, spectators can be better informed of the physical and mental conditions of the runner, and can provide appropriate support while the runner is in need. Hence, exertion, sensations, emotions and motivation states are critical to understand the runners' experience and needs, not just performance measures.

4.4.2 Potential for building SARE-ARS for automatically capturing runners' experience

Advances in affective computing and sensing technology provide the basis for this new direction in designing SARE-ARS technology. For example, pain levels can be measured using physiological sensors [119, 207] and activity sensors, in both static and sport situations [132, 225]. Fatigue and muscle tiredness can be captured through EMG and pressure sensors (e.g., in shoes [254]). Skin temperature can be

tracked by temperature sensors or even thermal cameras which are becoming smaller and smaller (e.g., attached to headset [55]). Body movement sensors can also detect affective states that emerged in our studies, such as pride and excitement [123, 132], low confidence and concentration [198]. Physiological sensors can track high arousal, excitement [75], and frustration [76]. In addition, movement sensors might be able to discriminate between periods of associative strategies, possibly characterized by a more controlled body stance to save resources, from dissociative states where the body (and the head in particular) may be less controlled as engaged in exploring the external world [37, 198, 200].

New opportunities for SARE-ARS interventions are provided through increasingly used low cost portable EEGs [61] in HCI [193] and the growing body of work on EEG measurements during long distance sport events (e.g., review [53], cycling[73], running[68], running and mood[275]). This work shows the possibilities offered by EEG to discriminate between levels of fatigue and suggests that associative strategies to cope with increased physical demand may result in output in the pre-frontal cortex. Simultaneously, work on pleasurable tactile experience illustrates how EEG measurements of the sensory motor cortex (among other brain areas) can discriminate pleasurable tactile sensations (lack of cortical activity inhibition) from not-pleasurable/neutral tactile sensations (cortical activity inhibition) on the arm [241]. Such measurements could detect pleasurable thermal sensations on the skin caused by reactions of different skin mechanoreceptors.

Thus, there is potential for SARE-ARS experience sharing technology to be built on such automatic recognition capabilities. Yet, more foundational work is needed to better understand how running affects expression patterns (behavioral, physiological, neural), and to deal with possible interferences and possible confound (e.g., sweat as discussed in [73]). Separating exertion measures from emotional measures in physiological signals is also a challenge, but methods have been proposed to discriminate physiological signatures [281] and could be further studied in this

context. Multi-modal detection of such patterns is also shown to produce more reliable 4-C meanings interpretation [198, 282].

4.4.3 How can SARE-ARS be used?

The findings of the study also highlighted additional design implications for enhancing the experience of both runners and spectators. Although these implications are not directly related to RQ1, they do provide valuable insights into the requirements for building SARE-ARS and offer potential use cases for their implementation.

4.4.3.1 SARE-ARS supporting the design of annotated SARE map

The findings of this chapter propose the development of an annotated SARE map featuring multimodal representations of the runner's SARE, which could benefit both runners and spectators alike. Spectators would be able to gain a better understanding of the runner's journey and provide more effective support by experiencing the event from the runner's perspective, while also making informed decisions about where to go and when to cheer. Runners, on the other hand, would be able to recall and share their experiences after the event. The SARE-ARS system could also be used in conjunction with the annotated map to provide real-time visualization.

Our findings suggest that using SARE-ARS to support the creation of an annotated SARE map could be more interactive and engaging for spectators, rather than simply providing them with the map. This is because, for some runners, inferring the runner's state (particularly during non-critical moments) can be an enjoyable part of the supporter's role as they leverage their knowledge and observe the runner's behavior during the race. Such results suggests other ways to improve spectators engagement beyond those mentioned in the literature [64, 94, 112, 114, 267, 276].

4.4.3.2 SARE-ARS supporting physical/virtual micro-stage encounters

The findings of the study have important design implications for SARE-ARS to enhance the runner-spectator interaction during physical and virtual micro-stage encounters.

Our in-depth analysis shows (DyPECS:IL) that these micro-encounters go beyond the moment of sighting and involve tension building, efforts invested, and appraisal of success. Success is determined by the fulfillment of roles and experiences of both runners and spectators during these encounters. Hence, incorporating SARE-ARS supported by digital media and movie-making strategies [175] can create tension and surprise before encounters and facilitate appraisal during and after them. SARE-ARS can provide information about the runner's 4-C states to spectators, allowing them to deliver appropriate support and potentially rewarding supporters for fulfilling their role. By doing so, SARE-ARS can provide novel ways of enhancing interactive technology designed to enhance the sharing of experiences [62, 219, 276].

Additionally, our findings about micro physical encounters could inspire the design of virtual/digital encounters for runners and spectators (R&S). Using SARE-ARS, runners' experiential progress could be communicated to remote spectators, enabling a digital encounter that goes beyond simple cheering to include emotional aspects such as the possibility of missed opportunities and frustration. This would provide novel and valuable possibilities for engagement, particularly given the increasing number of remote spectators who follow runners internationally from their homes through apps.

The use of SARE-ARS can provide valuable insights into the appropriate level of interaction needed to ensure a positive experience for both runners and spectators. By leveraging the system's automation capabilities, runners and spectators can benefit from real-time feedback and personalized recommendations on how to optimize their interaction, creating a seamless and engaging experience for all participants.

4.4.3.3 Supporting Spectator's Runner-Identity Process (Not relevant to SARE-ARS)

Our findings have implications for supporting the spectator's runner-identity (SRI) process, although they are not directly related to SARE-ARS or RQ1. As such, we have included them in Appendix B-Chapter 4.

4.5 Summary

In response to **RQ1: What are the components of SARE that matter when sharing, and why**, this chapter employs a range of qualitative approaches, including secondary analysis of pre-collected qualitative data (supplemented with newly collected data). The findings provide a rich and detailed understanding of runners' subjective and affective experiences during long-distance running events (LDRE), shedding light on the complexity of sharing and how the different states of SARE runners direct the type of support they need from spectators and when silent support is more effective. Runners find value in sharing their discomfort and hardships with others during long-distance running events, as it is considered part of the beauty of the experience. They also seek emotional support during moments of desperation and self-doubt and want their experiences to be understood. The findings in this chapter also suggested that gathering self-report data about a runner's level of exertion, pain, emotion, and the moment when they perceive the desire to give up is crucial when building SARE-ARS to support runners and spectators during LDRE.

Moving forward, the subsequent chapters of this thesis will build on these findings to provide a comprehensive understanding of how such types of SARE can be collected via runners' self-reports.

Chapter 5 Developing R-ESM to collect data for building SARE-ARS

5.1 Introduction

The previous chapter provided an understanding of the affective and subjective running experience (**SARE**) that technology needs to sense and capture to support runners and their audiences or supporters. It suggested that AI-based wearable technology capable of automatically recognizing a runner's SARE. In this chapter, SARE will be referred to as more specific subjective states such as subjectively perceived exertion, perceived pain and discomfort sensations, desire to stop or continue running, and emotional state, could help with the management of both runners' own experience and spectators' supporting tasks. Therefore, this chapter investigates the feasibility of building runners' SARE Automatic Recognition Systems (**SARE-ARS**).

Building an ARS system requires gathering both people's self-report of their state of SARE (called ground truth in the context of machine learning) and the physiological and behavioural data that reflect those states. Hence, this chapter addresses the following question:

RQ2: How can a ubiquitous ESM based on voice and wearables be designed to gather in-situ self-reports and running data for the purpose of creating SARE-ARS?

As seen in Chapter 2, most methods employed in sport psychology and HCI for collecting self-reports from runners are largely controlled, such as inducing exertion in lab-based contexts[165]. Therefore, this chapter investigates the possibility of using an Experience Sampling Method (ESM) to enable the data collection of runners' SARE self-report during their own running activity outside the lab and focuses on capturing the continuous dynamic variation of runners' SARE rather than their states at specific phases of exertion. We also look at runners' acceptability of such frequent self-reports (e.g., every 1 min.) as, generally, in ESM studies, people are asked to self-report only a few times per day.

To address this question, a voice- and wearables-based ESM was implemented as a mobile app to wirelessly deliver speech-based ESM prompts and record runners' voice-based self-reports via earbuds, while simultaneously collecting physiological and behavioural data that relates to runners' SARE by using wearable sensors.

5.2 Prototyping a Voice & Wearable-based Runner ESM System (R-ESM) for Data Collection

To facilitate the process of gathering data from runners, a wearable and voice-based ESM was implemented as a digital and wearable system for runners' ESM (hereafter called R-ESM). As shown in Figure 5-1, this system includes a central Android mobile application that manages multi-modal wearable sensing devices to collect runners' behavioural and physiological data, such as a smart insole [236] with a pressure sensor for collecting stride pressure signals, an Empatica smart bracelet [1] for tracking physiological signals (e.g., heart rate, blood pulse, and galvanic skin response), as well as acceleration-based motion signals. The system also communicates with earbuds eSense [127], a smart headset to deliver speech-based ESM prompts and record runners' voice-based self-reports.

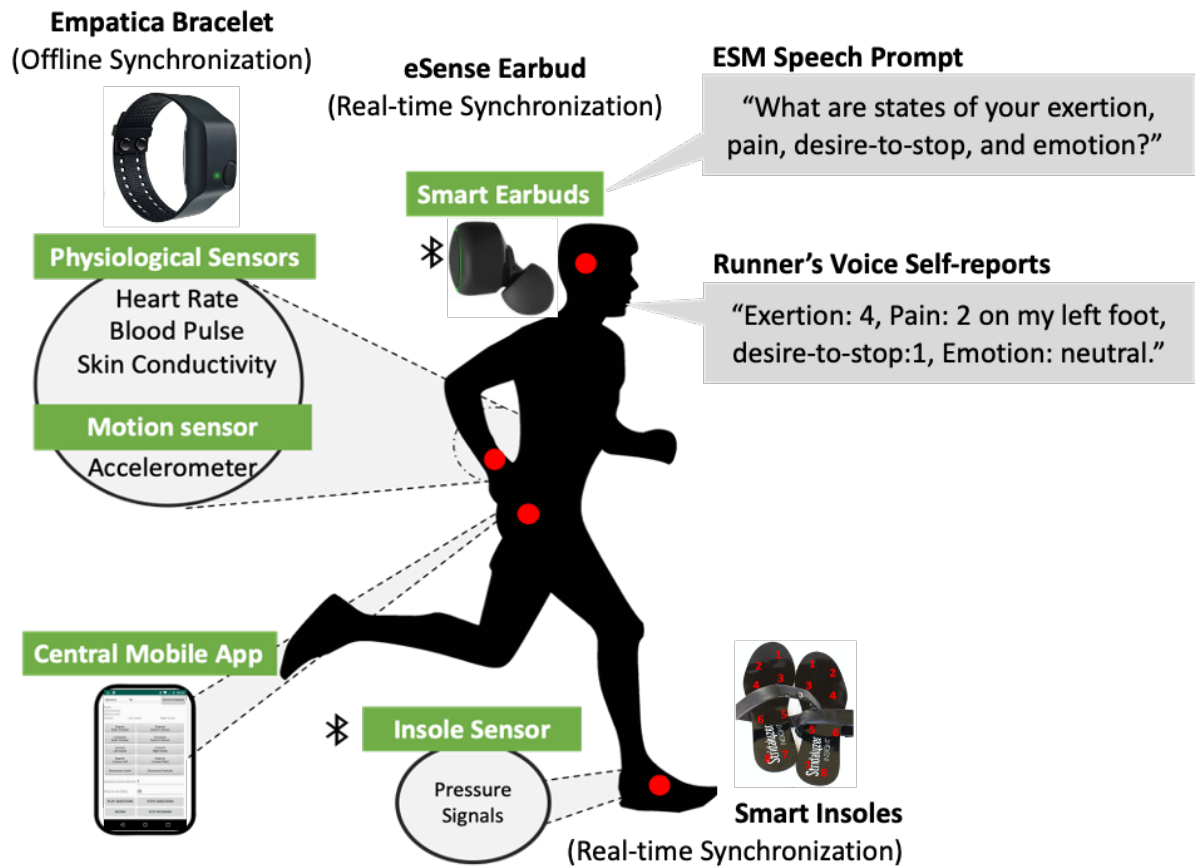


Figure 5-1: Wearable & Voice-based Runner-ESM (R-ESM) system

Next, we introduce each part of the wearable sensors included in the ESM and then introduce the ESM protocol design.

5.2.1 R-ESM using pressure Insoles for collecting feet pressure data

As outlined in Chapter 2, fatigue can have an effect on the pressure distribution within the feet. This motivates the use of smart insole devices as a non-invasive wearable to collect biomechanical running data. We first introduce the selection process for the pressure insole and then present an initial exploration of using insole to capture runners' emotional information.

5.2.1.1 Market research for available pressure insoles

This section first provides a review of some off-the-shelf smart insoles commercially available at the time of this study (November 2018), including the following four pressure insoles: Arion Pro Pack⁹, DigitSole¹⁰, Stridalyzer Insight¹¹, and IEE Smart Foot¹² Sensor.

The Arion Pro Pack, priced at \$259, is an 8-pressure-sensitive sensor system notable for its flexibility and ultrathin design at 2mm. It incorporates a multi-axis accelerometer, gyroscope, a Bluetooth-based chest band for heart rate monitoring, GPS, and on-board memory. Data generated from the sensor system includes footstrike, cadence, balance, stability/variability, contact time, flight time, pace, speed, and distance. However, raw sensor data could not be accessed directly, and had to be requested from the company separately.

The DigitSole Pressure Sensor, with a retail price of 129.90€, containing three Force Sensitive Resistors, offers a performance-based running analysis; however, it does not afford access to the raw pressure data.

IEE Smart Foot Sensor has the ability to Dynamic monitoring of plantar foot pressure and 6 DOF inertial sensing information and there is Data Logger + Insoles pair that costs 99.00 EUR but the full kit costs 400EUR for rent and 2800 EUR to buy.

Stridalyzer Insight is a pressure insole product (Figure 5-2) that includes 10 Force-sensitive Sensors (FSR) placed at eight important locations (see Figure 5-3 for a

⁹ <https://www.arion.run/product/arion-pro-pack/>

¹⁰ <https://www.digitsole.com/store/connected-insoles-running-run-profiler/>

¹¹ <https://www.retisense.com/product/stridalyzer-insight-sensor-insoles/>

¹² <https://www.iee.lu/en/products/sports-healthcare/smart-foot-sensor>

reference to foot anatomy): Hallux, other lesser toes, Front 1 (between 1st and 2nd metatarsophalangeal joints, two sensors), Front 2 (centre of 3rd-5th metatarsophalangeal joints, two sensors), centre of arch, centre of midfoot, centre of heel, and outer of heel. The insole kit provides raw pressure data as well as other gait analysis features such as foot orientation, foot vibration, stride length, ground contact time, stride rate (cadence), strike type, pronation type, overstriding/understriding, weight distribution (body balance), impact forces, push-off forces, and loading rate (dF/dt). It also provides a Developer Toolkit and Application Programming Interface (API) for researchers to integrate the insole in their own applications for real-time and customisable usage. Despite its numerous features, the product does come with a significant price tag of \$499 and is at an early stage in its development cycle.



Figure 5-2 Stridalzyer Insight Smart Insole Sensor (Left: picture of the real product; Right: picture of the concept product)

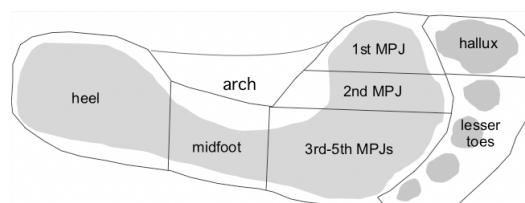


Figure 5-3: visualization of foot anatomy

The sensing frequency has been set to 8 Hz (default), but it is configurable between 1 Hz and 50 Hz. The insole is flat (0-drop from heel to toe) with a minimal bending radius of 50 mm. The circuit within the insole is 3mm and the circuit around the arch part is 8mm. The circuit battery life is estimated at 20 hours of continuous usage and may be recharged via a micro-USB charging mechanism.

After taking into account various factors such as price, features, and access to raw data as well as the opportunities for customization, the RetiSense Stridalyzer Insole was selected for use in this study. Two pairs of insoles (one 25cm insole to be used participants feet size between 24cm ~26cm and the other 28cm insole to be used for participants whose feet size between 27cm ~29cm) were purchased together with the developer kit.

5.2.1.2 Exploring insoles in Acted Emotional Runs

A pilot study was conducted to explore the potential of pressure insoles to capture discriminative and characteristic information for runners' subjective and affective running experiences. Three participants (see Table 5-1) were recruited from the researcher's lab for the initial exploration, and they were invited to simulate experiencing emotions while running with pressure insoles. This initial study did not aim to experimentally examine the correlation between affective states and running pressure during running.

Table 5-1 Participants Information for Initial Insole Exploration

ID	Gender	Weight(Kg)	Feet Size	Insole Size
1	Male	67	25cm	25cm
2	Male	69	27.5cm	28 cm
3	Female	71	24.5cm	25cm

5.2.1.2.1 Procedure

Participants were guided to engage in four Acted Emotional Runs, in which runners were asked to act as if they were feeling motivated, enjoyable, painful, and tired while running. These four align with the four states discussed in the findings of Chapter 4.

A 3-minute elicitation interview was employed to encourage participants to recall memories associated with the four emotional conditions while running, and then act as if they were feeling the same while running.

The Acted Emotional Run occurred in a 20-meter-long, straight, carpeted corridor located at the University College London Interaction Centre (UCLIC). For each condition, participants ran back and forth, traversing a total run distance of 40 meters.

Participants were not interrupted during the run and there were no requirements regarding completion time. Participants were instructed to jump once both before and after the run, to mark the start and the end of the data, respectively. They were given five minutes for rest after each run. The iOS application Stridalyzer from the product kit [5] was used to collect data from the pressure insoles with a sample rate of 25 Hz, with the app maintained on the phone's screen and battery saving mode disabled.

5.2.1.2.2 Analysis

The data for each acted emotional condition were segmented based on the jump mark. Due to the small sample size, statistical analysis and comparison among participants were not deemed essential. Instead, participant data were visualized to provide a tangible way to explore whether the insole data cap could capture discriminative information for the four acted emotional runs. To do this, time series-based line charts was used to plot the trends, and boxplots was used for plotting distribution using Python.

5.2.1.2.3 Results

The main objective of the analysis reported in this section is to explore the sensitivity to expected typical changes in pressure patterns that may occur during running in response to different emotional states, levels of exertion, and pain, as suggested by the literature. It should be noted that this analysis does not aim to demonstrate the ability to discriminate between different states but rather to simply assess the extent to which these low-cost sensors respond to stereotypical changes, serving as our starting point in this study. An analysis of changes in real patterns in relation to changes in states will be carried out in Section 6.4 using machine learning techniques as an initial proof of concept for using wearable sensors to capture behavioural cues related to such states.

To provide an overview of the insole pressure sensitivity in relation to stereotypical running, we present the plots of aggregated measurement, i.e., Ground Contact Time (GCT/ms). Ground Contact Time (GCT) is a measure of how long the foot remains in contact with the ground during each stride, and it is derived from the readings of the eight pressure sensors, namely Hallux, OtherToes, Front, Front2, Mid, Arch, Heel, and Heel2. The visualizations of each individual sensor's response data can be found in the Appendix 11.3.2 for reference. We use the aggregated sensor pressure measurements rather than measures of individual sensors given the difference in fitting of our insoles to participants and also differences in running styles. However, it is important to note that due to the limited sample size of three participants, the statistical significance of the findings may be limited and therefore not the focus of this section.

The boxplot shown in Figure 5-4 represents the distribution of P1's left foot GCT for the four acted emotional conditions: feeling tired, painful, enjoyable and motivated (labelled as tired_running, painful_running, enjoyable_running and motivational_running in the figure). The X-axis indicates the four different acted emotional runs, and the Y-axis indicates the GCT. The GCT with higher values

indicates longer ground contact time. In this boxplot, the median GCT for “tired running” is approximately 380ms, indicating that half of the strides’ GCT during this run were above 383ms and the other half were below. The interquartile range (IQR) for “tired_running” extends from approximately 380ms to 401ms. This means that the middle 50% of strides’ GCT for “tired running” fall within this range. The boxplot’s whiskers indicate the minimum (approximately 360ms) and maximum values (approximately 420ms) within 1.5 times of IQR. As expected, the distribution is positively skewed, meaning that more data points are concentrated towards higher GCT values. There is also at least one outlier data point (i.e., an exceptional high value of 485ms approximately) beyond this range.

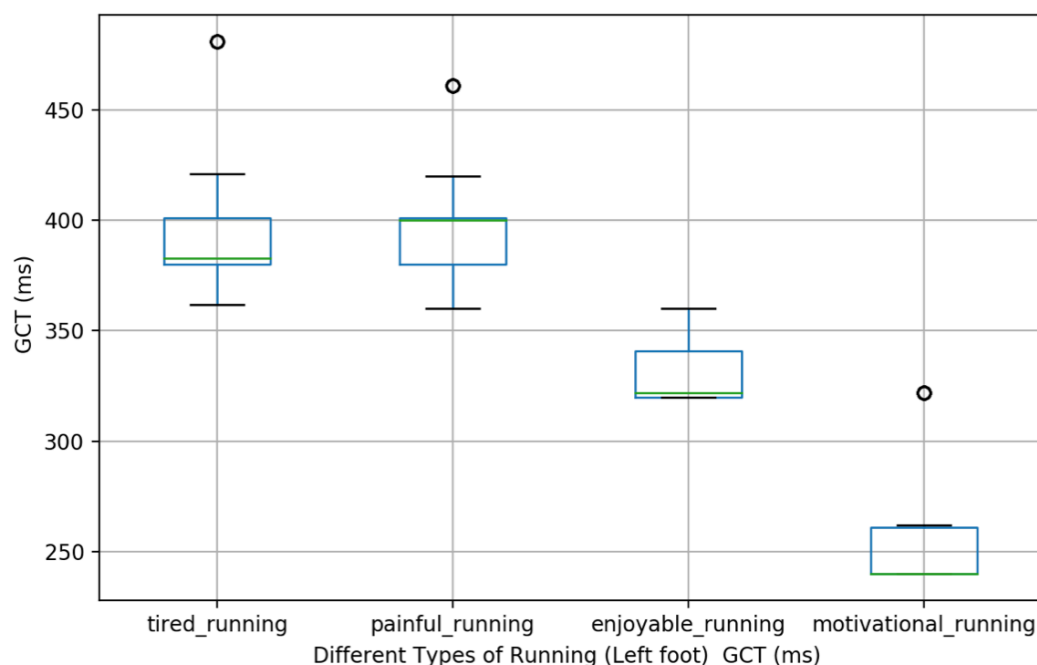


Figure 5-4: Box Chart of P1’s Ground Contact Time (GCT) on the Left Foot under Four Acted Emotional Conditions: Feeling Tired, Painful, Enjoyable and Motivated (Labelled as Tired_running, Painful_running, Enjoyable_running and Motivational_running in the Figure)

For P1's "painful_running", the median left foot GCT is around 400ms, indicating an increased level of ground contact time compared to "tired running". The IQR for "painful running", spans from approximately 380ms to 401ms, suggesting a similar spread of left foot GCT compared to "tired running" but with a negatively skewed distribution of the pressure values.

P1's "enjoyable_running" exhibits a median level of the GCT approximately 320ms, lower than both "tired running" and "enjoyable running". The distribution of the GCT level appears to be positively skewed, with the boxplot ranging from around 317ms to around 340ms. No outliers are present.

P1's "motivational_running" displays the lowest median level of the left foot's GCT among these four runs, which is close to 225ms. The distribution of the GCT appears to be positively skewed, with the box ranging from around 240ms to 261ms. Potential outliers are observed as circles above the higher whisker.

Overall, these boxplots indicate that, "motivational_running" and "enjoyable_running" tend to have lower GCT, whereas "tired_running" and "painful_running" tend to have higher GCT. Figure 5-5 displays that P2 tends to have lower GCT (possibly indicating higher cadence and shorter ground contact time) while feeling enjoyable and motivated compared to feeling tired and painful.

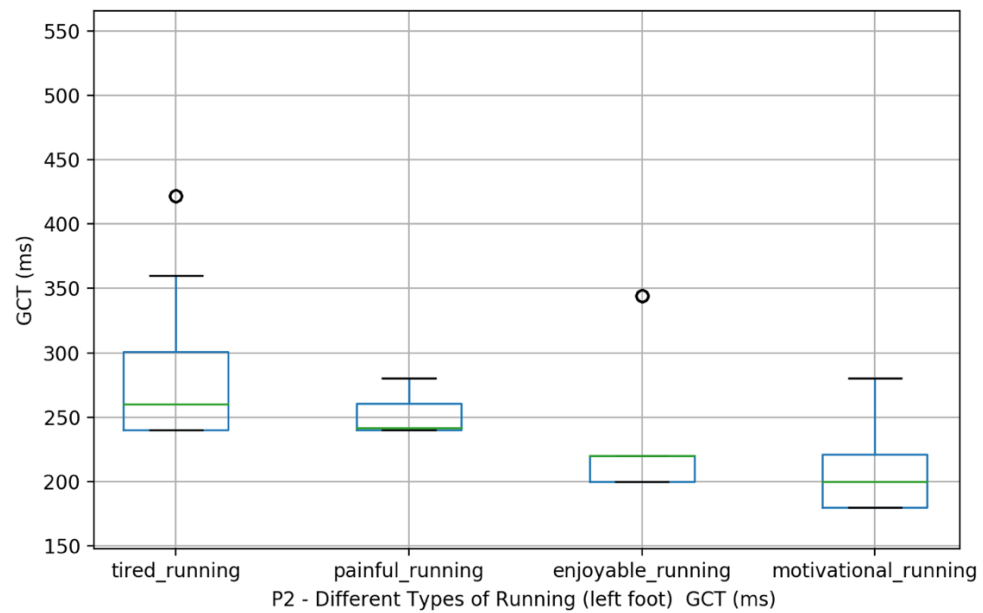


Figure 5-5: Box Chart of P2's Ground Contact Time (GCT) on the Left Foot under Four Acted Emotional Conditions: Feeling Tired, Painful, Enjoyable and Motivated (Labelled as Tired_running, Painful_running, Enjoyable_running and Motivational_running respectively)

However, P3's left foot GCT data does not exhibit a similar pattern. As shown in Figure 5-6, when feeling tired, the median level of P3's left foot GCT is approximately 400ms. The distribution of GCT levels appears to be negatively skewed, with the boxplot ranging from around 45ms to around 445ms. When feeling painful, the median level of P3's left foot GCT is approximately 320ms. The distribution of GCT levels appears to be highly negatively skewed, with the boxplot ranging from around 20ms to around 320ms. P3's GCT distribution is much wider compared to P1. From the figure, we can observe that P3 tends to have a higher average GCT when feeling enjoyable and motivated, in contrast to when feeling painful and tired. These changes in GCT confirm the expected variability in how individual runners may cope with physical and emotional states via their bodily features, as discussed in Chapter 2, Sections 2.2.

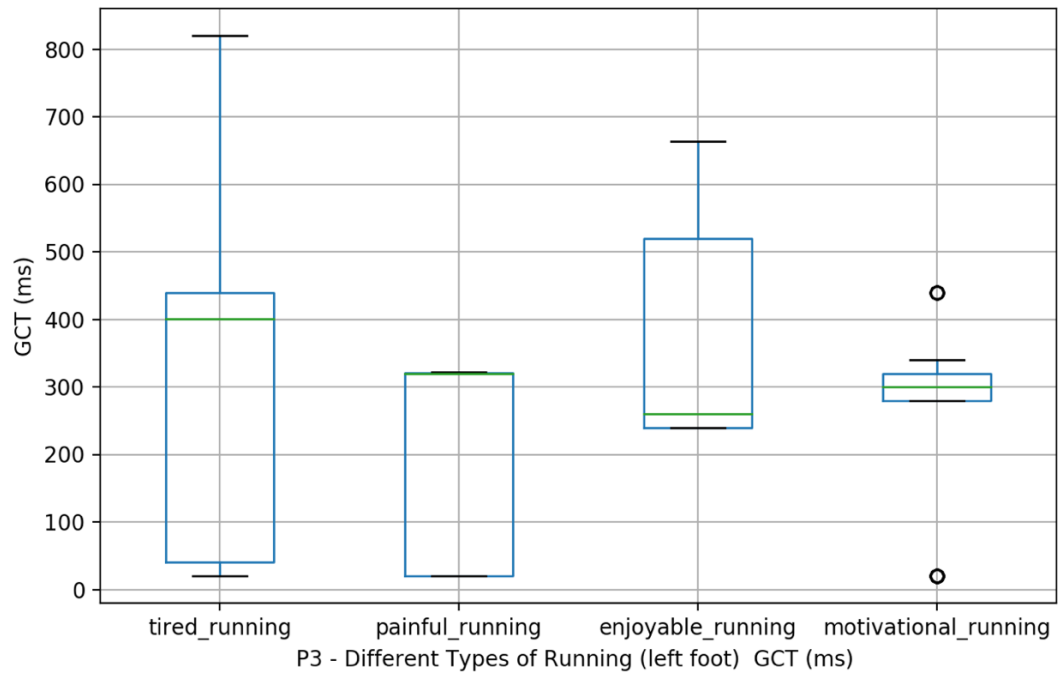


Figure 5-6: Box Chart of P3's Ground Contact Time (GCT) on the Left Foot under Four Acted Emotional Conditions: Feeling Tired, Painful, Enjoyable and Motivated (Labelled as Tired_running, Painful_running, Enjoyable_running and Motivational_running respectively)

The data presented in this analysis is derived from a limited sample size, consisting of data from only three participants. It is important to note that the findings and conclusions drawn from this study may not be generalizable to a larger population. However, despite the limited sample size, the results offer valuable insights indicating that the insole sensors exhibit sensitivity to the typical variations in runners' subjective and emotional states, as expected from existing literature in the field of sport psychology related to running.

5.2.1.3 Summary

Nevertheless, these findings were not intended to examine the exact pattern but to demonstrate that sensor data could discriminately differentiate between the four different acted emotional runs. Such results are indicative of the potential of the insole to be employed in the R-ESM system for collecting data to construct SARE ARS. Furthermore, it suggests that the left and right insole can be used as individual modalities, as people appear to have independent features on each foot when constructing the features for such ARS.

5.2.2 R-ESM using Empatica for collecting physiological and movement data

As discussed in the literature, physiological measurements such as heart rate and blood oxygen levels can be indicative of runners' perceived exertion. Furthermore, affective computing research has demonstrated that galvanic skin response is linked to emotional state. Consequently, the Empatica E4 bracelet (Figure 5-7) was utilized in conjunction with the R-ESM system. This bracelet is a reliable, non-invasive, and easily accessible wearable sensor which has been extensively employed and validated in affective computing research to collect data used to infer users' emotional state.

As introduced in [1], the Empatica E4 utilizes a three-axis accelerometer (ACC), which is sampled at 32 Hz, to collect motion signals. Additionally, a photoplethysmograph (PPG) sensor measures heart rate at 1 Hz while the blood volume pulse (BVP) at 64 Hz. Moreover, a skin conductance sensor (also called electrodermal activity (EDA) or galvanic skin response (GSR)) is used to measure electrical conductance of the skin at 4 Hz, providing information on emotional arousal and stress levels. Finally, a skin temperature (TEMP) sensor is sampled at 4 Hz to track changes in body temperature during exercise.



Figure 5-7: Empatica E4 Bracelet [1]

5.2.3 R-ESM using eSense headset for delivering speech prompts and recording voice reports

R-ESM also uses a wireless headset to deliver the speech prompts and recorder users' voice reports. eSense [127] was chosen as it is a very lightweight sensor earbud, weighing 20 g and measuring 18 x 20 x 20 mm, per earbud. As shown in Figure 5-8, it is equipped with a microphone and dual mode Bluetooth (Bluetooth Classic and Bluetooth Low Energy), powered by an ultra-thin 40 mAh LiPo battery. Additionally, it offers the capability of using accelerometer and gyroscope sensors to track head gesture and body trunk position (see Figure 7: Overview of eSense).

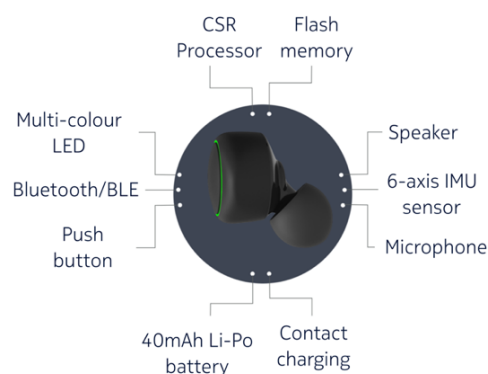


Figure 5-8:Overview of eSense [127]

5.2.4 R-ESM prompt design

The SARE self-report protocol was designed to include two parts, the ESM prompting parts to gather the ground truth and a spontaneous think to gather runners' free thoughts to understand better their perspective on their running experience.

This study focused on four components of SARE that were identified as important through sport psychology literature and the user study conducted in Chapter 4: runner's perceived exertion, pain level, desire to stop/continue running (hereafter referred to as desire-to-stop), and emotional state (valence). An iterative process was used to design the prompts structure for these four components. Initially, the ESM prompts questionnaire for gathering exertion levels and pain followed the descriptions and rating structure of the Borg Scale [4]. However, participants in the pilot study found it difficult to remember the meanings of the levels in such a scale without needing to read it. To facilitate verbal self-report while running, a simpler 5-point scale was adopted for the perceived exertion level, perceived pain level and desire-to-stop. Lastly, the prompt posed for the emotional state was based on a 3-point scale. This three-level categorical appraisal of one's emotional state was more easily interpretable during running, and it was suggested that such a scale would better account for individual variability in participants' emotional responses to their physical states [243]. The specific prompts and scale for each of the four SARE components with their respective meanings are shown in Table 5-2.

Table 5-2 Prompts for the four SARE components

Exertion Prompt: How difficult do you find the current run to be? (e.g. strain and fatigue in your muscles, or feelings of breathlessness)					
Numeric Rating	1	2	3	4	5
Descriptive Rating	Not at all	Very Light	Somewhat Hard	Very Hard	Extremely Hard
Explanation	Not feeling anything	It's just noticeable	It feels hard, but I can maintain the effort	I have to push myself	The strongest exertion level I have ever experienced
Pain Prompt: How much pain, if any, do you feel?					
Numeric Rating	1	2	3	4	5
Descriptive Rating	Not at all	Very Light	Somewhat strong	Very strong	Extremely strong
Explanation	Not feeling anything	It's just noticeable	I feel the pain is affecting me, but I can cope with it	It starts to be difficult to run with the pain	The strongest pain I have ever experienced
Desire-to-stop Prompt: To what extent, do you want to stop?					
Numeric Rating	1	2	3	4	5
Descriptive Rating	Not at all	Very Light	Somewhat strong	Very strong	Extremely strong
Explanation	Not feeling anything	This thought is just noticeable	I do think about stopping, but I want to keep running	I'm about to stop.	I have to stop now.
Emotion Prompt: What is your current emotional state? (positive, neutral, or negative)					
Descriptive Rating	Positive		Neutral		Negative

These four ESM prompt questions were converted into speech using text to speech conversation techniques [57], with each prompt lasting approximately three seconds, followed by seven seconds for participants to provide a rating. To facilitate this process, participants in the pilot testing were trained to use the scales for the prompt questionnaire before running. However, the average duration of this delivery and rating process was 40 seconds, which was considered too intrusive for the participant's run. To address this, the four prompts were merged into a single 3-second speech prompt: *"What is your level of exertion, pain, desire-to-stop, and your emotional valence?"*, followed by an average of seven seconds for participants to provide the rating in the format: *"Exertion: 4, Pain: 2, Desire to Stop: 1, Emotion: Neutral"*. This process only required an average of 10 seconds. Then, various frequencies ranging from 5 to 0.5 minutes per prompt were tested, with participants indicating that one minute was the minimum rate they could accept. Consequently, the ESM probing question was set at every 1 minute, resulting in participants being interrupted for 10 seconds during a 1-minute window of running (interruption ratio: 1/6). This allowed for an increased prompting rate, thereby facilitating the collection of more data for machine learning building.

5.2.4.1.1 A think-aloud structure

In addition to responding to the rating-based ESM prompts, this study was interested in capturing runners' free thoughts in order to gain a better understanding of their perspective on their running experience; for example, what matters to them and where their attention is directed. To this end, runners were encouraged to think aloud if they wanted between self-reports. The efficacy and feasibility of the think-aloud method, wherein runners verbalize their subjective experience, has been established [74, 197, 222] in prior research (e.g., [222]). However, it is important to acknowledge that thinking aloud may introduce potential additional effects on runners' subjective experience, such as the extra cognitive efforts as well as the perceived time discussed in the Section 2.3.1. Therefore, in our study, we will also consider how the think-aloud method might influence and potentially alter runners'

feelings during the interview. This will shed lights on how think alouds can be used as an appropriate approach in the context of continuous

Prior to the run, three think-aloud tasks (Math Problem, Anagram, Naming/Free Association Task) [74] had been prepared to help familiarize our participants with the think-aloud techniques, following the procedure outlined in [222], in order to assess the participants' performance on their think-aloud tasks. Tasks are described in Table 5-3: Think-aloud Task for Training .

Table 5-3: Think-aloud Task for Training [222]

Task Name	Task Description
Math Problem	<i>“Could you please think aloud as you multiply 18 times 6 in your head.”</i>
Anagram	<i>“Now I would like you to solve an anagram. I will show you a card with scrambled letters; it is your task to find an English word that consists of all the presented letters. For example, if the scrambled letters are KORO, you may see that these letters spell ROOK. Any questions? Please think-aloud while you solve the following anagram:<NPHEPA = HAPPEN></i>
Naming/free association task	<i>“Now I would like you think aloud as you name twenty different animals. Simply verbalize what passes through your mind as you name them.”</i>

5.2.5 R-ESM App development

In order to synchronize all wearable devices to be used by runners in real time, the R-ESM mobile application was developed, including four components (see Figure 5-9: R-ESM Mobile App Architecture): acquisition of pressure data from both insoles, customization of the ESM schedule based on the participant's planned running time, management of eSense to deliver speech-based ESM prompts and record voice-based self-reports, and acquisition of motion data from eSense.

The R-ESM developed individual Bluetooth Low Energy (BLE) management functions for each sensor, registering and monitoring its BLE connection. The R-ESM connects

to the eSense using its service UUID. However, the left and right insoles share the same name and service UUID; thus, the sensor's manufacturing data was then utilized to identify each insole. Specifically, the ninth byte of the advertised data indicates the ASCII code for either “Left” (76) or “Right” (82). Received data from the Bluetooth advertisement was retained in the form of a byte array of 20 elements. To make use of the eSense Sensor, the R-ESM imported “*io.esense.esenselib*” and “*StrideAnalyticsLib*” to parse the advertised byte-based sensor data into meaningful pressure data and motion data. The R-ESM app was developed and tested on the following Android devices using Android Studio 3.4.1: LG Nexus 5 (Android OS 6.0.1, API level 23) and Motorola G7 Power (Android OS 9.0.0, API level 28). Figure 5-10 Shows the final interface pf the Android based R-ESM app. The Empatica E4 bracelet data was synchronized with R-ESM app offline based on their shared timestamp.

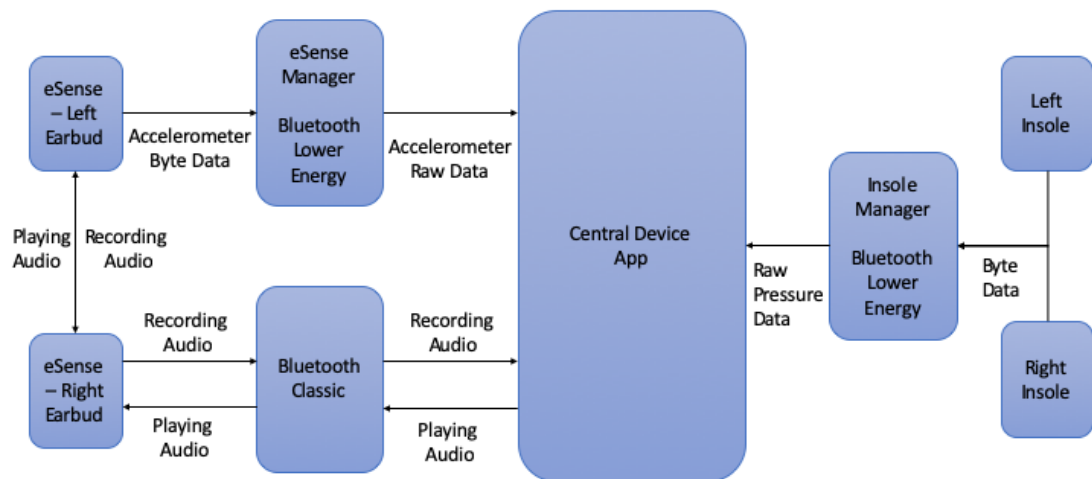


Figure 5-9: R-ESM Mobile App Architecture

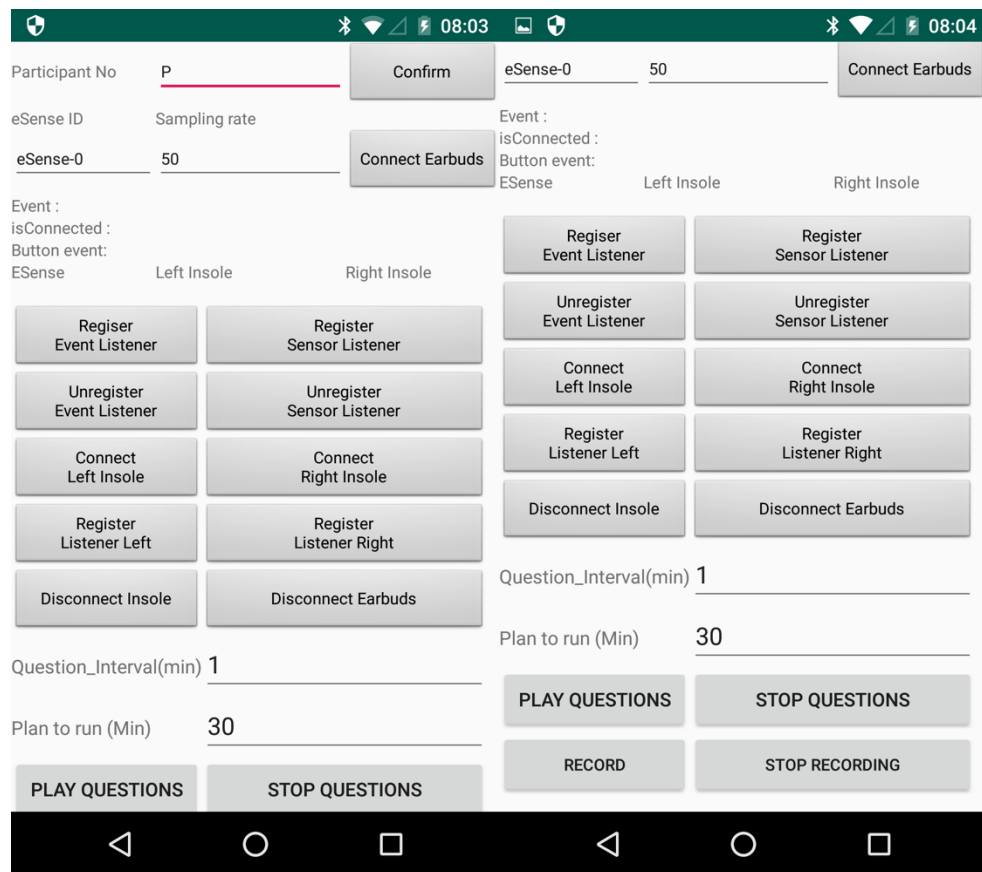


Figure 5-10: Screenshots of the Android based R-ESM App

The R-ESM app was designed to facilitate the offline annotation of data by providing alignments between participant's voice self-reports and the corresponding time recorded by the app. When an ESM prompt was presented, its time-stamp was connected to the participant's follow-up voice self-report. This was done for two reasons; it was simpler and more controlled than attempting to extract an exact time from the recordings, while also reflecting the fact that participants typically begin mentally preparing their self-report upon hearing the speech prompt, thus using the prompt's time as reporting time is reasonable.

The R-ESM app associates self-report data to the sensor data following a one-minute discrete interval of 30 seconds before the prompt time and 30 seconds after the

prompt time (visualised in Figure 6-1). Such labelling approach is similar to the quantised labelling (also called archetype labelling) as suggested in [60]. This was due to the difficulty of precisely determining when a feeling begins and ends, as feelings are not experienced within such a discrete temporal frame as self-reports. Utilising a 30 second window before and after the voice report time helps to maximise the coverage of meaningful information in the sensor data. However, further exploration is required, for instance, other labelling options such as 10 seconds before and 50 seconds after the prompt time, or 0 seconds before and 60 seconds after the prompt time, could also be tested.

In summary, the R-ESM system captures the following types of data as shown in Table 5-4.

Table 5-4 Data Type Captured in the R-ESM

Type			Specification
Rating-based reports	voice	self-	<i>“Exertion: 4, Pain: 2, Desire to Stop: 1, Emotion: Neutral”.</i>
Open-ended alouds	voice	think-	Participants’ verbal feedback on their running experience without any specific prompts.
Pressure signals			Pressure signals (in KG/cm ²) on 8 different areas of the foot, such as 1): Big Toe, 2): Smaller Toes, 3): Front 1: 1st and 2nd Metatarsal bones, 4): Front 2: 3rd, 4th and 5th Metatarsal bones, 5): Arch: Inside arch of the foot, 6): Mid: Outside of the midfoot, 7): Heel 1: Medial/inner heel, 8): Heel 2: Lateral/outer heel.
Physiological signals			Electrodermal activity (EDA) in (μS), heart rate (HR) in (bpm), temperature (°C), and blood volume pulse (BVP)
Movement signals			Acceleration in the x-axis (Acc_x, g), y-axis (Acc_y, g), and z-axis (Acc_z, g).

The next chapter will present a user study in which R-ESM was used by runners in their run.

Chapter 6 Understanding the efficacy of R-ESM: a user study

6.1 Introduction

Following Chapter 5, this Chapter presents a user study to expose runners to the R-ESM system during their run. Machine learning techniques were employed to validate the feasibility of mapping the ground truth to the physiological and behavioural data collected from the runners to build the SARE-ARS. In addition, post-running interviews were gathered and analysed (e.g., semi-structured interviews and thematic analysis) to gain insight into runners' self-reporting experiences and inform the design of R-ESM.

The chapter is organized as follows. First, it outlines the user study method. It then discusses the results of the ARS modelling and the interview analysis.

6.2 Method

6.2.1 Participants

The criteria for participation in this study were mid-distance or long-distance running (5KM or 30 minutes and above). No requirements for running speed or performance were necessary. Participants were recruited from friends, university students, gyms, and running clubs through social networking applications (e.g. Facebook, Twitter) and physical flyers. In total, 11 participants (5 male and 6 female) were recruited and asked questions regarding their basic personal information, running history, running routine as well as the planned running location during the study. Participants were permitted to run anywhere they chose, provided it was a safe environment with no traffic (e.g. an outdoor track (OT), indoor treadmill (IT), outdoor park (OP), etc.). Information about their foot strike pattern and foot/shoe size was gathered as reference in the analysis of insole sensor data. Participants were categorised into four types of runners, i.e., novice/beginner, intermediate, advanced and elite, based on their experience level using standard introduced in Section 3.2). Participants' information was summarized in Table 6-1.

Table 6-1: Participants' Profile (Indoor Treadmill (IT), Outdoor Track (OT), Outdoor Park (OP), M: Male, F: Female)

Participant	Session ID	Gender	Time-Location	Running Experience Level	Foot Strike Pattern	Shoe Size & Insole Size ¹³
P1	P1-1	M	25min-OT	Novice	Rearfoot	CN42.5/26CM
	P1-2		35min-OT			INSOLE 28CM
	P1-3		35min-IT			
	P1-4		30min-OT			
P2	P2-1	M	43min-IT	Advanced	Midfoot	UK9/27.3CM INSOLE 28CM
P3	P3-1	F	35min-IT	Advanced	Midfoot	UK6/25CM INSOLE 25CM
P4	P4-1	F	35min-IT	Novice	Midfoot	EU39/25CM INSOLE 25CM
P5	P5-1	F	30min-OT	Intermediate	Rearfoot	UK8/26.7CM INSOLE 25CM
P6	P6-1	F	40min-OP	Intermediate	Midfoot	UK6 INSOLE 25CM
P7	P7-1	F	35min-OP	Novice	Rearfoot	UK6 INSOLE 25CM
P8	P8-1	M	35min-OT	Novice	Rearfoot	EU44/27.3CM INSOLE 28CM
P9	P9-1	M	35min-IT	Intermediate	Rearfoot	CN42/25.6CM INSOLE 25CM
P10	P10-1	M	60min-OP	Novice	Rearfoot	UK6 INSOLE 25CM
P11	P11-1	F	25min-IT	Novice	Midfoot	EU39/25CM INSOLE 25CM

¹³ <http://www.shoesizes.co>

6.2.2 Study procedure

Prior to the study, participants were trained for 10 minutes to become familiar with the sensors, self-reporting and think-aloud techniques following the procedure in [222]. As shown in Figure 6-1 Study Protocol, participants had a slow, self-paced trial run to warm up and ensure they did not experience discomfort from the equipment. During the running session, the R-ESM system presented questions and collected data, as previously described. Overall, 11 runners (five males, six females) participated in 14 running sessions; nine participants completed one session, one runner completed three sessions, and one runner completed two sessions. The average duration of running sessions was 34.7 ± 15.1 minutes.

During the study, participants were informed that if they were interested, they had the option to participate in multiple sessions voluntarily for longitudinal data collection. The additional data collected from these participants served the purpose of machine learning training and allowed for the future development of person-based machine learning models. Out of all the participants, only P1 conducted multiple sessions.

All of the data collected from P1 during the sessions were used for machine learning training. However, we acknowledge the potential limitation that including additional data from P1 might have resulted in the model learning specific patterns that are unique to P1. This not only makes the model potentially more attuned to P1 but also to other participants who exhibit similar patterns. Since our objective in this thesis was to develop the machine learning model as a proof of concept for building SARE-ARS, we recommend that future research consider the imbalanced distribution of data among different participants and the variability in self-reported SARE states within each participant to enhance the robustness and generalizability of the findings.

After the running session, participants were invited to participate in a semi-structured interview about their experience of reporting to the system, beginning

with easy warming-up questions such as *“What is your most memorable feeling during this run?”* and *“How did your state of mind change during the run?”* They were then asked to elaborate on their experience of using and interacting with the R-ESM system during that moment and to provide feedback about the system and the study. The interview lasted approximately half an hour.

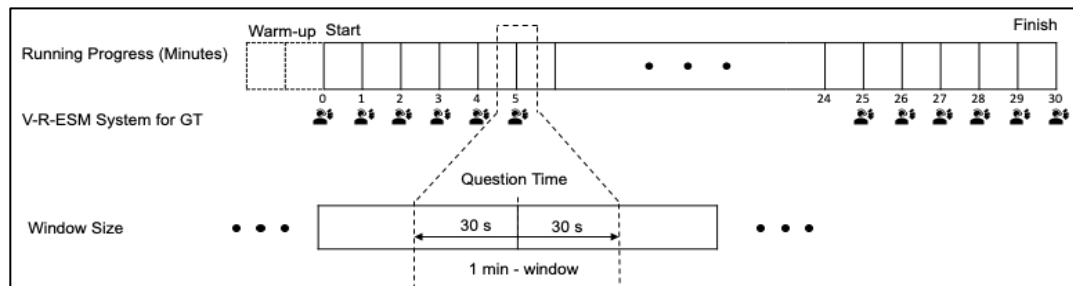


Figure 6-1 Study Protocol

6.2.3 Data Analysis Method

6.2.3.1 Self-report and qualitative data analysis method

Voice self-reports were transcribed and numerically coded using the questionnaire 5-point scale for exertion, pain, desire-to-stop, with positive, neutral, and negative emotion being assigned values of 1, 0, and -1, respectively. A total of 315 *4 self-reports were collected from all participants. The ratings were quantitatively analysed and visualised to present an overview of the data, including time-series based visualisations, the distribution of the reported ratings, and the correlations between each type of SARE. Additionally, think-aloud audio data was transcribed using Nvivo, and an inductive qualitative content analysis was conducted by reading through the data line by line, following the method used in [222], to evaluate it as a possible way to extend the question-based ESM with more free natural self-report to gather or refine the GT. Each key word or key meaning was extracted as a meaning unit and then sequentially grouped as sub-themes and final themes, with the number of

participants that reported on each theme also being extracted. Such results will be presented in in Section 6.3.

Thematic analysis [43, 44] was used to analyse the semi-structured interviews.

6.2.3.2 Sensor data pre-processing method

Sampling and Data Cleaning: Real-world in-the-wild data collection using sensors is not always as clean and organized as expected. In this study, we discarded data frames with the following issues: lacking self-report labels, continuous empty frames (e.g., due to Bluetooth failure, broken sensors and circuits caused by frequent use in running), and frames containing many abnormal outlier values (e.g., consistently maximum outliers).

Data Masking: To address the empty values present in the pressure insoles dataset, a linear interpolation method [205] was employed to estimate the missing value based on its known neighbouring values. This is due to the fact that the feet contact pressure is not constant, but rather dynamically increasing or decreasing during the landing phase[155]. For the sensor data such as heart rate, electrodermal activity, and accelerometer data, a forward fill approach was utilized to fill in the subsequent missing data frames.

Data Resampling & Normalisation: Despite the sensor being set to a fixed sampling rate, missing data points are often observed in the collected sensor dataset. This could be due to a loss of signal during the Bluetooth transmission phase (e.g., signal loss due to wireless advertisement and reception). For instance, during the data collection study, the pressure sensor was set to a sample rate of 10 Hz, meaning that 10 data frames were expected per second. However, in some cases, a small number of frames might have been missing or duplicated within a second, leading to an uneven sampling frequency. Thus, a resampling technique [206] was employed to reset the initially desired sample rate (e.g., 10 Hz for the insole sensor), supplemented with forward filling if needed. Subsequently, MinMaxScaler [8]

technique was employed to normalise each type of sensor data to the range of 0-1, accounting for such individual variability among participants [54].

6.2.4 Machine Learning Methods for SARE ARS

This section outlines the methods of constructing Automatic Recognition Systems (ARS) using data collected from self-reports and running, including data processing, and feature extraction, methods used for training and testing, as well as validation of the ARS.

Ground Truth Recategorization: The original SARE self-report labels (as shown in Figure 6-5) were further recategorized into three groups to achieve a more balanced dataset. To do this, K-Means clustering ($n\_clusters = 3$) [7] was employed in order to achieve a balanced self-reports dataset as ground truth, with re-clustered results of three new categories (No or Little:0, Moderate:1, High:2) [55].

Feature Extraction: For pressure insoles, in addition to the original 8 types of pressure sensor data (χ_i refers to the ID of each pressure sensor on a insole), 7 extra types data firstly were computed, including: overall foot pressure (the sum of all pressure sensor data, $\sum_{i=1}^8 \chi_i$), front foot pressure (the sum of the four sensors on the front part of the insole, $\sum_{i=1}^4 \chi_i$), mid foot pressure (the sum of two sensors on the middle part, $\sum_{i=5}^6 \chi_i$), back foot pressure (the sum of two sensors on the heel part, $\sum_{i=7}^8 \chi_i$), inner foot pressure (the sum of four sensors on the inner side, $\sum_{i=1,3,5,7} \chi_i$), outer foot pressure (the sum of four sensors on the outer side, $\sum_{i=2,4,6,8} \chi_i$), the inner-outer difference (the difference between the sum of two inner and outer sides). Then, Statistical features were extracted, including the minimal, maximal, mean and standard deviation (std) values of the 15 individual pressure data, 4 Physiological Sensors data, and 3 Movement sensor data. This led to a total of 148 features for each one-minute window.

Training, Testing & Validation: To build the classification models for three levels of exertion, pain, emotion valence and desire-to-stop, we used the machine learning

algorithms: Random Forest (RF), Decision Tree (DT), K-nearest Neighbours algorithm (k-NN), Naive Bayes classifier, Decision Tree classifier, Perceptron algorithm, Multi-layer Perceptron (MLP) classifier, and Support Vector Machines (SVMs). These were selected as they are commonly used in affective computing [200-202]. Leave-one-out cross validation approach was used to evaluate the trained models. Leave-One-Out Cross Validation (LOOCV) and k-Fold Cross Validation (KFCV) are two popular validation techniques. LOOCV is beneficial when the training set is small, as it avoids the need to divide the data into multiple folds. KFCV is generally preferred when the training set is large, as it reduces computational efforts. According to ISL, LOOCV produces test error estimates with lower bias but higher variance. This is due to the fact that each training set contains all but one of the training examples, which leads to higher correlation between the test error estimates. On the other hand, k-Fold Cross Validation results in test error estimates with higher bias but less variance, as the overlap between training sets is relatively small. All these were implemented with Scikit-learn [6].

6.3 Results from Self-reports and Think-alouds data analysis

This section presents the findings from the analysis of the runners' self-reports of their SARE and think-alouds data. This includes the demonstration of the fluctuating dynamics of a runner's SARE while running in a real-life context, the distribution of the collected SARE ratings, the correlations between each type of SARE, and the content of the think-alouds data.

6.3.1 Fluctuating dynamics of a runner's SARE while running in a real-life context

Next, we demonstrate what the potential fluctuating dynamics of a runner's SARE are while running in a real-life context. For instance, Figure 6-2 illustrates the example of parts of P4's self-reported ratings of their perceived exertion, perceived pain, desire to stop, and emotional appraisal of the current moment.. The x-axis represents a part of their running session, measured in minutes (from the 1st to the 15th minute), and

the y-axis depicts the ratings of various self-reported states. P4's perceived exertion is visualized in blue, perceived pain in red, desire to stop in grey, and emotional valence in yellow. The levels of these four ratings display a non-linear and fluctuating pattern, with numerous ups and downs observed throughout the selected timeframe. The data from participant P4 is selected solely for demonstration purposes, but fluctuations are observed among all participants, although the patterns differed to some extent. This confirms the sports psychology literature in Section 2, which states that running experience is dynamic and fluctuating and more importantly, further validates that the proposed approach is effective in capturing the changes and dynamics of SARE in real-life contexts.

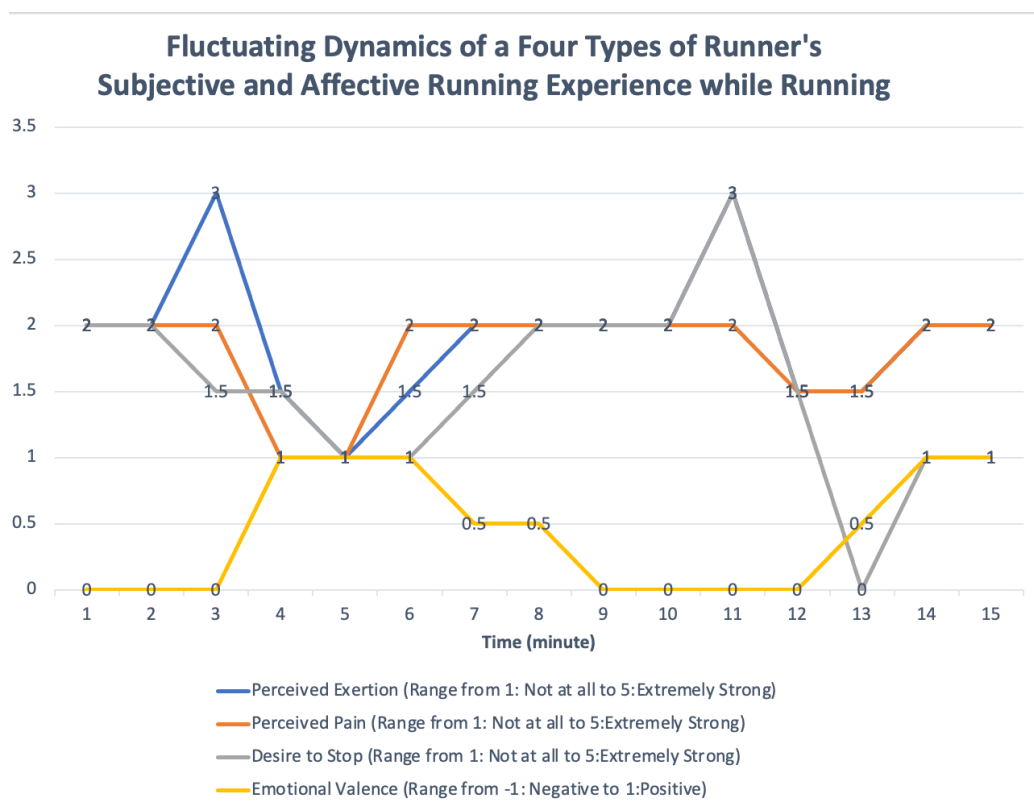


Figure 6-2 Fluctuating Dynamics of SARE (using P4's data as a demonstration).

6.3.2 Distribution of self-report SARE ratings

We created four set of boxplot graphs(see Figure 6-3) to visualize the distribution of 10 participants' self-report values for the different variables. It should be noted that P11's data was removed due to sensor malfunctioning issues. Each figure represents a different variable, namely exertion, pain, desire-to-stop, and emotion. In each figure, the x-axis represents the participant ID, while the y-axis indicates the corresponding self-report values for each variable. It should be reminded that for all variables but Emotion, the range of possible self-report values was between 0 (not at all: exerted, feeling pain, feeling a desire-to-stop) and 6 (very much: exerted, in pain, feeling a desire-to-stop). For emotion, the self-report range went from - 1 (negative) to +1 (positive) valence with 0 representing a neutral emotional state.

These plots show a data imbalance across the 4 variables. Participant P1's exertion self-report values have a median of 3, with a distribution ranging from 1 to 3.5. The boxplot indicates a negatively skewed distribution, suggesting a higher frequency of lower exertion levels. The upper whisker line represents the maximum value of 6, within 1.5 times the interquartile range (IQR). Participant P2 has a median exertion value of 2, and the box plot shows a relatively balanced and symmetric distribution ranging from 1 to 3, with an upper whisker of 4. The box for P2 is narrower compared to P1, indicating a more centralized distribution of exertion levels. For Participant P5, there is no box present in the plot. Instead, a single line and three dots are displayed. This suggests an extremely skewed distribution with the exertion self-reports centred around a value of 2, with outlier values of 1, 3, or 4. Similarly, for Participant P7, the pain self-reports, as well as the "desire-to-stop" self-reports from P2 and P8, also show an absence of a box, indicating extremely skewed distributions. Regarding the emotion self-reports, most participants (P2, P4, P5, P7, P8, P9, P10) exhibit box plots with an extremely skewed distribution. This is possibly due to the fact that the emotion scale we used was very narrow and people did never experience really negative experiences.

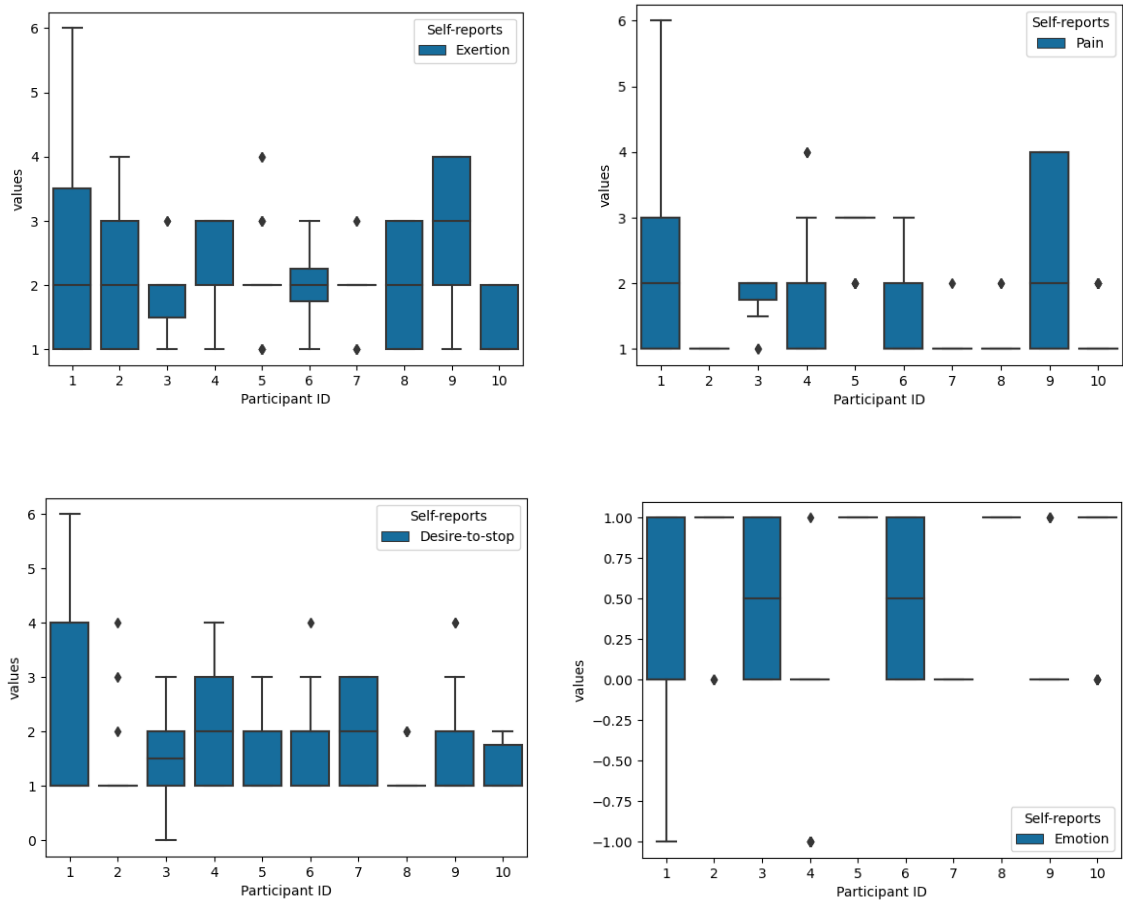


Figure 6-3 Self-reports distribution of each participant:for the four variables exertion, pain, desire-to-stop, and emotion (as indicated in the respective legend) for 10 participants.In each graph, the X-axis represents participant ID, while the Y-axis indicates corresponding self-report values. Box plots displaying median levels with boxes representing the first quartile, median, and third quartile. Outlier dots represent minor outliers. The upper whisker extends the top of the box to the maximum value in the dataset while the lower whisker indicates minimum values in the dataset within 1.5 times the interquartile range (IQR)

These boxplots effectively showcase the variations and individual differences in the central tendencies of each variable across participants, highlighting the unique individuality reflected in these subjective self-reports.

The variation observed in the boxplot distributions of each participant indicates an imbalance in the levels of self-reported exertion, pain, desire-to-stop, and emotion. Generally, there are fewer data points representing critical moments when participants experience high levels of exertion or pain. This could be due to participants encountering these intense moments less frequently compared to milder instances. According to our ethics application, we could not expose people to very high demand or competition. Regardless of the specific reasons, the data imbalance can pose challenges when utilizing them for computational purposes or machine learning applications.

In order to understand the amount of instances collected for each level of each variable to be modelled, Figure 6-4 shows the distribution of self-report values for exertion, pain, desire-to-stop, and emotion by levels. In each histogram, the Y-axis represents the count of instances for a specific level of self-report and the X-axis represents the range of self-report values. In the Histograms for exertion, pain, and desire-to-stop, the X-axis ranges from 0 to 6, whereas in the plot for emotion, the X-axis ranges from -1 to 1.

When plotting the self-reports data using histograms, as seen in Figure 6-4, there is a noticeable skew in the distribution of data points for levels of exertion, pain, and desire-to-stop. There are more data points concentrated towards lower levels, ranging from 1 to 3, in comparison to higher levels above 4. Similarly, for emotions, more data points are skewed towards neutral and positive states compared to negative states. Consequently, the modelling may not adequately represent the full range of levels and states that are not well-represented in the dataset.

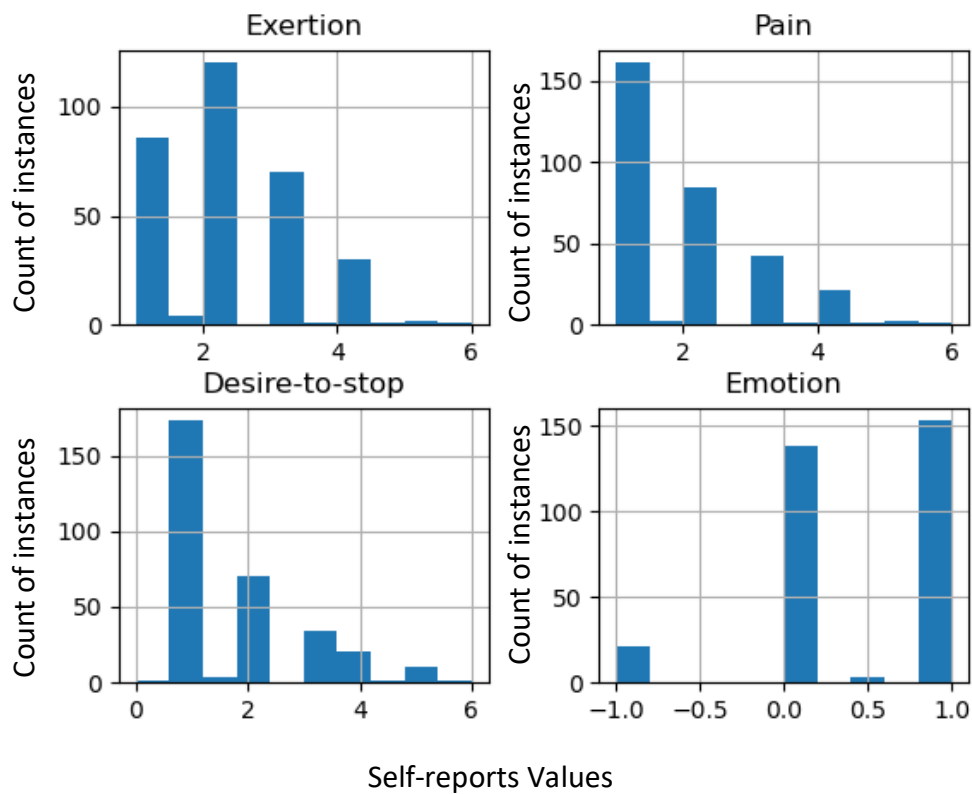


Figure 6-4 Overall Self-reports distribution: Histogram plots depicting the distribution of self-report values for exertion, pain, desire-to-stop, and emotion; Y-axis represents count of instances for a specific self-report level, while X-axis represents the self-report levels. Exertion, pain, and desire-to-stop subplots have X-axis range of 0 to 6, while emotion subplot ranges from -1 to 1.

Furthermore, as it can be seen in these plots, a few participants did not strictly adhere to the original integer ESM scale of 0 to 5, instead they used decimal ratings such as 2.5 and 4.5 and values exceeding the scales (e.g., 6). In order to account for individual variability observed in the boxplots in Figure 6-3, the self-report ratings for each variable separately were normalised to the range of 0 to 1 using `sklearn.preprocessing.MinMaxScaler`[8]. The results of which are shown in Figure 6-5. This normalization is a typical step taken before building affect recognition model given the subjective interpretation of the meaning of such scales and also the

different ability to cope with physical and psychological demand [54]. However, it should be noted that a longitudinal data collection would provide more reliable normalization allowing people to better reflect the link between the scale range and their level of experiences.

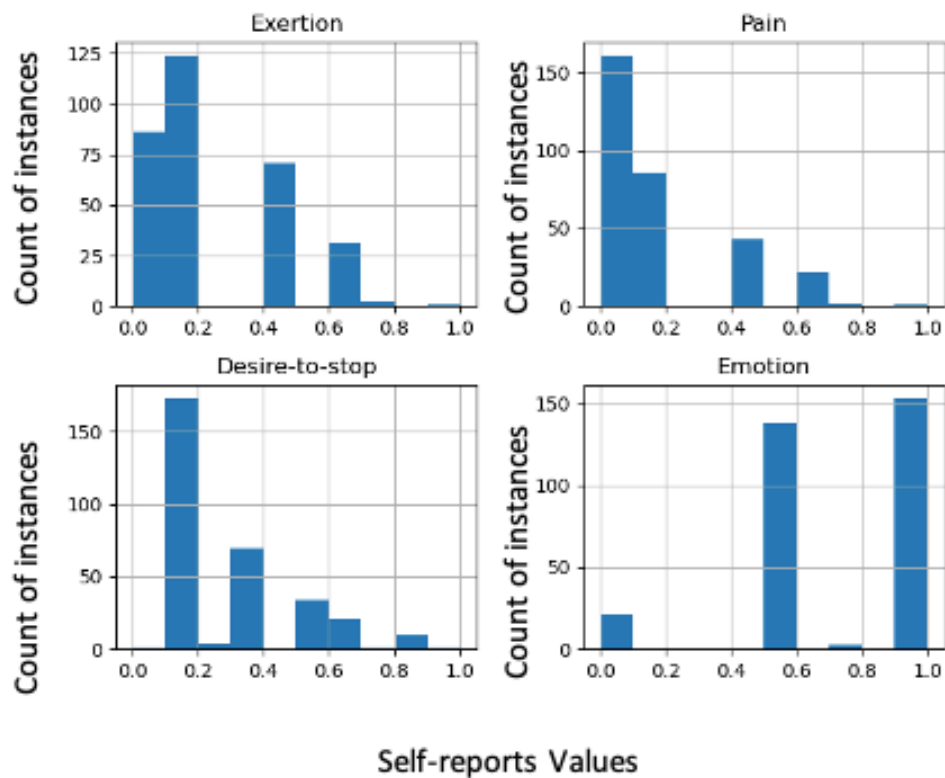


Figure 6-5 Normalized self-report values for exertion, pain, desire-to-stop, and emotion, ranging from 0 to 1 (as represented on the X-axis), to account for individual variability. The Y-axes represents the counts of instances and the x-axes the normalized ranges of self-report.

6.3.3 Demonstration of interrelations between each SARE ratings

Our objective in this analysis was to explore the correlation between each self-reported SARE state, aiming to preliminarily assess the reliability of the ratings and investigate whether they exhibit specific patterns related to participants' actual

feelings. By examining if pain and emotion, for example, consistently show significant correlations, we can gain insights into the consistency of participants' responses and potentially uncover underlying relationships between different SARE states. This initial exploration provides a valuable opportunity to evaluate the truthfulness of the self-reported SARE ratings in accurately capturing participants' genuine experiences.

Given the limited points and the fact that this are ranking data, we use the Spearman correlation coefficient. The results are shown in Table 6-2, and visualied in Figure 6-6. As expected, Exertion and Desire-to-stop demonstrate a moderate positive correlation ($r = 0.612$), suggesting that increased levels of Exertion are associated with a stronger desire to stop. Similarly, Exertion and Pain exhibit a slightly positive correlation ($r = 0.455$), while Desire-to-stop and Pain display also a slightly positive correlation ($r = 0.445$), suggesting that higher levels of Desire-to-stop may be associated with increased Exertion or Pain. This implies that the perception of pain can influence the desire to stop running. Instead, self-reports of emotion states show weak negative correlations with Exertion ($r = -0.191$), Pain ($r = -0.323$), and Desire-to-stop ($r = -0.131$). These findings show the expected interactions between these states. This lower correlation coefficients with the emotional valence of the state may be due to the highly skewed data collected for this variable and/or possibly a more complex relationship.

It's worth noting that due to the small sample size, the analysis only provides a preliminary understanding of the complex interrelationships among the SARE types. Therefore, we refrain from drawing strong conclusions regarding the extent and direction of the correlations. This analysis mainly meant to verify that the self-reports were valid in that they reflect expectation from the literature and hence usable to progress to the next step in our study, and verify as a proof of concept a possible relationship with the sensor data. Nevertheless, these findings highlight and reflect the literature showing that the SARE types are not independent but rather interrelated in complex ways that needs more in-depth investigation when gathering them as self-reports data.

Table 6-2 Correlation between each type of SARE (using Spearman correlation coefficient).

	Exertion	Pain	Desire-to-stop	Emotion
Exertion	1.000000	0.455416	0.612193	-0.190523
Pain	0.455416	1.000000	0.444786	-0.323124
Desire-to-stop	0.612193	0.444786	1.000000	-0.130581
Emotion	-0.190523	-0.323124	-0.130581	1.000000

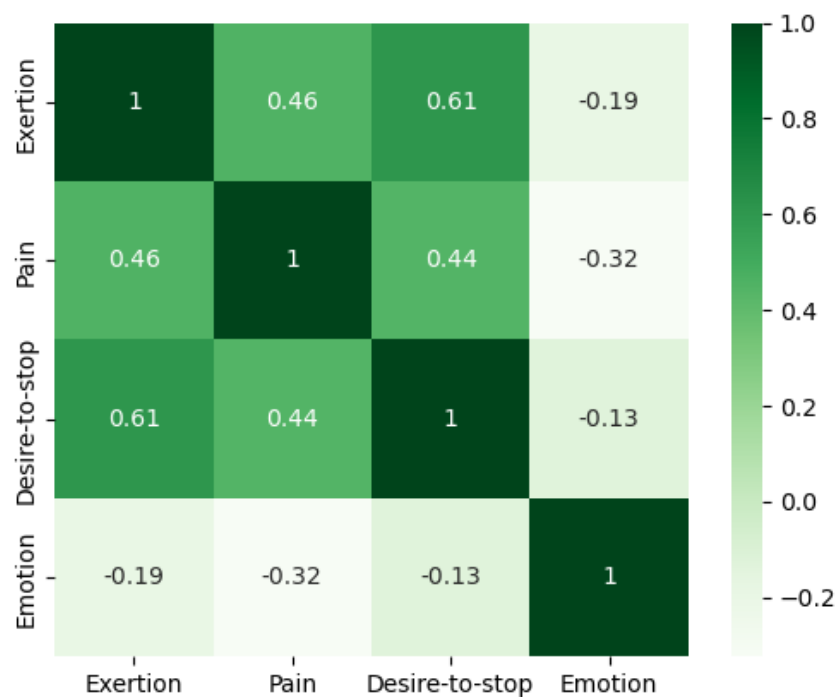


Figure 6-6: Correlation Heatmap: Exertion, Pain, Desire-to-stop, and Emotion (Spearman Method); The correlation coefficient value ranges from -1 to +1, where -1 represents a perfect negative correlation, +1 represents a perfect positive correlation, and 0 represents no correlation. Colour gradients in the heatmap indicate the strength and direction of the monotonic relationship between two variables, with darker colours indicating positive correlations and lighter colours indicating negative correlations.

6.3.4 Analysis results of Think-alouds

In addition to self-reporting, participants were invited to think aloud to see if they would engage in the think aloud process while running as in [222]. This aims to evaluate it as a possible way to extend the question-based ESM with more freedom of self-report to gather or refine the GT. This section reports the analysis of the think-alouds data but discuss it together with the interviews data. As shown in Figure 6-7, all participants but one (P4) engaged in the think aloud process at different degrees. As reported later in the interviews, the variability in engagement with the think-loud process was partly due to the participants' environment (e.g., crowded gym vs more isolated places) but also possibly due to individual personalities. Four main themes emerged: 1) clarification and characterization of self-reported GT labels, 2) reflections on current running, 3) external environment, 4) general life events. Table 6-3 Think-aloud statistics shows a breakdown of think-alouds.

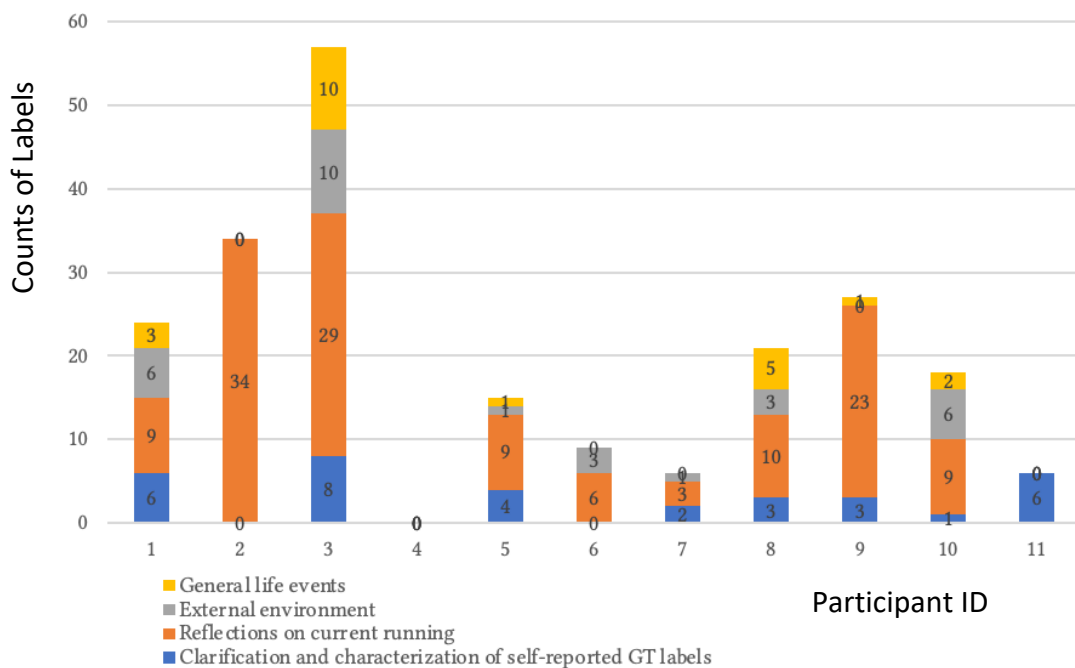


Figure 6-7 Frequency of think-alouds for each participant
(X-axis = Participant ID; Y-axis = Count of Labels)

Participants sometimes elaborated on their self-reported ratings of the four ESM questions. For example, P1 said: *“Exertion – 3: I feel that my lung is a bit under pressure. Desire-to-stop – 4: maybe because it's a bit late. Emotion – neutral: but I was feeling a bit worried because it looks like it'll be raining soon”*. These results confirmed the fact that runners are ready to engage in thinking aloud during running providing further insights on their state. In addition to clarifying their rating, most of the time participants elaborated on their running postures and progress as well as about other types of feeling related to running. For example, in relation to the style of running, P5 said *“I'm thinking about my running posture, like swing my arms, trying to [have my stride] impact near the front of my feet rather than on my heel, cos I know what's bad for my knees”*. When considering their feelings, P8 said *“Running alone is boring”*. In addition, beyond running, participants reported their attention on observations and thoughts about their surrounding such as *“I noticed the guy next to me [...] we're keeping a similar pace. Oh, he's a bit slower than me. Actually, quite slower.”* (P3). Possibly due to the open-ended nature of the think-alouds we purposely used in our study, our participants also show that their attention shifted away from running to focus on other affairs in their life. P8 for example started to think about *“Now, I'm thinking of my evening: what would I do in the beginning, probably a shower afterwards. Then [...] study, then [...] programming.”*

Table 6-3 Think-aloud statistics

Major Themes and Sub-themes (# of meaning units)	Mentioned by # of participants
Characterization of GT (N = 33)	N = 7 (P1,5,7,8,9,10,11)
Reflections on current running (N = 153)	N = 11
Body sensations (n = 53)	n = 11
Running postures/style (n = 45)	n = 7 (P1,2,3,5,8,9,10)
Running status & progress (n = 44)	n = 8 (P1,2,3,5,8,9,10)
Subjective feelings (n = 11)	n = 7 (P1,3,5,7,8,9,10)
External environment (N = 30)	N = 7 (P1,3,5,6,7,8,10)
General life events (N = 22)	N = 6 (P1,3,5,8,9,10)

6.4 Results of constructing SARE-ARS

As a result, after excluding non-usable data, the gathered data for machine learning training includes 308 instances of one-minute windows of pressure data from both left and right insoles, and 247 instances of one-minute windows from the Empatica bracelet sensor data (physiological and acceleration signals). 247 instances contained all insoles data, Empatica data and self-reports.

Simple classification models were developed and evaluated to investigate the feasibility of using the low cost physiological and behavioural sensors together with the ESM-based ground truth gathered in the wild to build an affective recognition system. Figure 6-8 illustrates the distribution of labels across three levels of each of the states to be modelled (this refers to using all features from all sensor together). As depicted in the figure, there are more counts of self-reports indicating lower or moderate levels of exertion, pain, and desire-to-stop compared to higher levels. Additionally, positive or neutral emotion self-reports outnumber negative emotion self-reports. This observation suggests an imbalance in the dataset, potentially introducing bias in the machine learning process. While addressing the imbalance is not the main focus of our thesis, it underscores the significance of designing better systems to address this issue and the importance of future machine learning approaches that can effectively handle dataset imbalances.

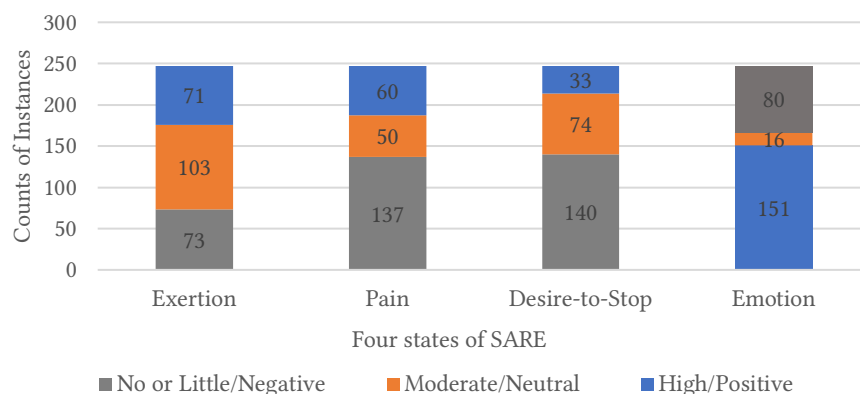


Figure 6-8 Ground Truth Labels Distribution

Since the goal of this study was not to compare machine learning algorithms and optimize performance, only the results obtained using the Random Forest were reported as it performed best in terms of discriminating between levels of exertion, of pain, of emotional valence and of desire-to-stop. The model was built with the following settings: `information criterion='gini', n_estimators =100, max_samples=0.25`.

Table 6-4 reports the results for each state with respect to the different sensed modalities: left and right insole pressure signals, physiological signals and arm movement signal. For each affective and physiological state, the performances are reported as follow: individual F1-score achieved for each of the levels (No or little, Moderate, High) of exertion, pain and want to stop and three categories of emotional valence (Negative, Neutral, Positive). In addition, we also report the overall weighted F1-score and classification accuracy with the best results obtained always for the combinations of all three modalities highlighted in bold: Exertion F1 = 0.79, Pain F1 = 0.84, Desire-to-stop F1 = 0.78, and Emotional valence F1 = 0.85. It can also be noticed that all the F1 scores for each level of each of the states are well above chance level. However, in case of Want-to-stop, the F1 score for the high-level label is much lower than the overall average F1 score (while still being above chance level). This is possibly due to the low number of instances available for this label as shown in table 1 (Desire-to-stop).

Table 6-4 Classification results for the 4 SARE states (Number in brackets after the F1 scores indicate the number of instances for that level)

EXERTION	No/Li ttle (F1)	Mod erate (F1)	High (F1)	Number of Instances	Weighted Avg. F1	Accur acy %
insole L + R	0.76 (96)	0.72 (127)	0.7 (85)	308	0.73	0.73
HR+EDA+BVP+TEMP+ ACC	0.73 (73)	0.74 (103)	0.71 (71)	247	0.73	0.73
Insole L&R + HR+EDA+BVP+TEMP+ ACC	0.83 (73)	0.78 (103)	0.75 (71)	247	0.79	0.79
PAIN	No/Li ttle (F1)	Mod erate (F1)	High (F1)	Number of Instances	Weighted Avg. F1	Accur acy %
insole L + R	0.86 (163)	0.66 (68)	0.78 (77)	308	0.78	0.79
HR+EDA+BVP+TEMP+ ACC	0.86 (137)	0.61 (50)	0.76 (60)	247	0.78	0.79
Insole L&R + HR+EDA+BVP+TEMP+ ACC	0.89 (137)	0.73 (50)	0.81 (60)	247	0.84	0.84
DESIRE-TO-STOP	No/Li ttle (F1)	Mod erate (F1)	High (F1)	Number of Instances	Weighted Avg. F1	Accur acy %
insole L + R	0.84 (171)	0.72 (98)	0.5 (39)	308	0.76	0.77
HR+EDA+BVP+TEMP+ ACC	0.87 (140)	0.73 (74)	0.43 (33)	247	0.77	0.79
Insole L&R + HR+EDA+BVP+TEMP+ ACC	0.88 (140)	0.76 (74)	0.44 (33)	247	0.78	0.8
EMOTIONAL VALENCE	Nega tive (F1)	Neutr al (F1)	Posit ive (F1)	Number of Instances	Weighted Avg. F1	Accur acy %
insole L + R	0.86 (172)	0.75 (30)	0.76 (106)	308	0.82	0.82
HR+EDA+BVP+TEMP+ ACC	0.88 (151)	0.71 (16)	0.75 (80)	247	0.83	0.83
Insole L&R + HR+EDA+BVP+TEMP+ ACC	0.89 (151)	0.91 (16)	0.77 (80)	247	0.85	0.85

6.5 Results of Qualitative Interview

This section reports the thematic analysis resulted in four themes discussed below:

1) acceptance of R-ESM, 2) companionship, and emotional experience of using the system, 3) creating a mindful running experience, and 4) adapting to the dynamism and complexity of the running experience.

6.5.1 Perceived relevance and acceptance of R-ESM

The designed ESM prompts were perceived as easily relatable and empathetic to participants' running experience, which kept them engaged in the self-reporting process. As one participant (P3) remarked, *"I can tell that it makes sense for the system to ask this question, as it is the same ones that I ask myself, like postural checks, techniques, feelings. When running every day, one spends a lot of time considering these elements."* P2 also commented, *"The question about exertion or effort, good or bad, and how strong it is, is not too difficult to answer. You should more or less know where they are in the range normally."* Despite the repetitiveness of such ESM prompts, answering these questions soon became harmonious and habitual. As P3 noted, *"To me, it was straightforward; I was not hearing the entire question as I was familiar with it. When hearing it twice, one knows what is being asked and can think of the answers."* P7 corroborated this sentiment, *"I [...] got into a rhythm, it was very clear."* P8 recognised the benefits of using it in more rigorous training and expressed their enthusiasm for its incorporation, stating *"I would make use of it more when I am engaged in more professional and serious training, because it's about keeping data on how I run."*

When participants were asked to what extent they perceived the speech ESM prompt to be interruptive or non-interruptive, they responded positively, e.g., *"It didn't interrupt me"* (P8) and *"I didn't feel intrusive"* (P7). P7 reflected on this hands-free experience, seeing talking back to the ESM prompt as an effortless way of

documenting their feelings. They explained, *“I don't know how else you would track it. Maybe by tapping? But that's more intrusive, because you don't want to take out something, and then you have to do a bit of interaction”* (P7).

6.5.2 Companionship, and emotional experience of using the system

The R-ESM system provided runners with a sense of companionship through its voice-reporting feature. Runners felt as if they were having a conversation with someone, which made them happier and more engaged in their running experience. P10 referred to their interactions with the system as *“a friend,”* and commented that, *“it's good, it's talking to me like a friend. It keeps me entertained and occupied.”* Similarly, P3 stated, *“It was interesting having these inputs, like this question every now and then. It felt like I had company.”* This prompted talking aloud, which further contributed positive and happy emotional experiences, as P5 remarked, *“It prompted me to talk. [...] I found that talking to the device made me feel happier (compared to normal running). Whereas if I had just stayed silent, I wouldn't have been so happy, and I don't know why that was the case. It feels like I have someone to talk to even though you're just talking out loud.”* Such positive associations could have encouraged the development of mental habits of interacting with the ESM prompt. For example, P5 anticipated the prompts before they arrived, stating, *“I was anticipating them before they would come in a little bit.”* They also expressed doubt, wondering, *“Why isn't it coming?”* and expressed relief when the prompts did arrive, saying, *“then the question would come, and I'd be like, 'Oh, that was fine then.'”*

Following perceived social and emotional companionship, participants expressed a desire for personalizing the system and making it more natural or human-like. They wanted the reporting experience with R-ESM to be more akin to natural conversations, even to be used in their daily runs: *“It could also be a good conversational assistant. Right now, it's so much like 'Please rate your blablabla from one to five', instead if it asked 'How do you feel? Are you tired now?' [...] maybe you could find certain phrases or words that are more common between people”* (P8).

Some participants expressed a preference for more human-like traits in the system, such as a human voice rather than a computer-generated, robotic one. They suggested the voice of a person familiar to the user (e.g., a friend, family member, or fellow runner) could be utilized. P3 commented, *“I think if it is like a person that you know, it would be an interesting experience. It will feel more familiar instead of being a synthetic voice [...] So, it's like they are asking you how you're doing during the run [...] especially for someone that you typically share your running experiences with.”*

In contrast to the positive emotional responses, some participants expressed feeling social pressure when conversing with the system in a social context, expressing concern about the potential for judgement from others. Participant 7 (who ran in the gym) remarked, *“It felt a little bit weird talking like that, and people were looking at me.”* Participant 4 additionally commented, *“As I was talking aloud, I was also considering whether the girl next to me would hear me talking to myself.”*

6.5.3 A mindful & introspective running experience

Although the R-ESM tool was developed for the purpose of collecting behavioural running data and self-reported ground-truth labels, our qualitative findings highlight that all participants engaged in a mindful running experience due to listening to and answering these questions while running. The system was not simply perceived as a companion to help time pass by, but to help focus and reflect on their running. Participants reported that the system led them to focus on the present (run) whereas without such a system, their mind would wander. P5 explained their normal state of mind, *“quite often my state of mind would have ever so slightly changed every minute, even though that seems really fast. Yeah, my state of mind can change so much when you're running.”* Others reported that they did not often actively reflect on their experience. For example, P3 mentioned usually listening to music passively during a run rather than actively focusing on their run (also P7).

Participants attributed this mindful attentional focus with the R-ESM tool on the current moment by listening to and answering the questions asked by it. P8 reflected on this, *"It brings me back to thinking about myself, in terms of exertion and so on. [...] thinking about, 'Oh, how do I feel now', and that's not what you usually do when you run. You know, (what you usually do is) you just run and then if you feel something you internalize it in a way that you adapt."* (Being present *"in-the-moment"* and focusing on their run triggered runners to check in with their body during the run. P3 explained, *"...maybe you are thinking something completely different ...Then, at some point, I had that input (the question from R-ESM) that I start to think like, 'okay, let's check my posture. I feel like my shoulders are a bit tight.' So, I want to release and relax the shoulders and the arm.' I was saying that aloud like 'now let's stop for a moment that I'm willing to breathe better and I will look at my posture.' So, I will [...] make sure that I'm running a very economical way."*

While the R-ESM asked for numeric answers, it encouraged runners to further think about, explore and elaborate on their states of mind. P3 said, *"You probably wouldn't want to just limit yourself to a number, you also want to say, I reply like that [either a higher or lower exertion/pain level] because I have this issue [...]"* Thinking about the numbers led runners to deliberate rationally as to what the numbers meant to them over a run or to their running experience. P11 said, *"It forces me, for me it's a positive thing, to rationally say this exertion that I'm feeling. Having the question and me answering made me reflect like, 'Okay, before I was 1 [on the scale of exertion], so I was not tired. And then I started getting a bit tired, is now two? I was rationally thinking: is this how tired I can get? can I get more tired? Can I get less tired? It made me compare throughout the session. So, it's not only a number. [...] It has more nuance, more richness [...] So, this question help[s] me say, 'you're tired, but it's not the worst that you can be'"* P2, a professional runner, also reported engaging in this reflective process, *"[By talking about my exertion level], I knew I was reaching only 3. So, I was deciding, okay, [...], Now I'm going for 4."* P4, even if not think aloud (See Figure 4), reflected further about her experience beyond giving a number to the

system. *"I tried to be consistent with how I was feeling [...] I had to spend a bit of time as well or attention in thinking, what do I mean by my myself?"* P8 agreed, *"because of the scores, [I started to think], how do I feel?"*

The reflective thinking process helped runners to achieve a more objective understanding of the perceived subjective running experience and evaluate their efforts during the run. P5 elaborated, *"[...] when you have the numbers in front of you [difficulty level showed on the machine in the gym, you can actually see what your output is. Whereas running [outside], apart from today, I never have any gauge of how hard I'm going. So, I always devalue my exertion because I'm not seeing any numbers. I've always pushed myself harder in the past because I thought I haven't achieved very much."* Attaching a number to their subjective feelings at different temporal stages created a richer mental visualization of the running experience. *"But I was also thinking, 'how do I feel? I still feel positive and I'd have no pain. Oh, suddenly a little bit pain.' Because of the scores, you have this graph, it changes in your head whenever it asks."* (P8) P7 felt it helped to visualize their run, *"It was quite interesting for me as a self-reflection. [...] I felt like I could understand how I'm changing over time, even though it wasn't a massive change. I feel like there was a general pattern of, being okay at the beginning, then feeling tired and then stopping a bit and walking, then being okay again and then being tired. So, in that sense, it was a bit predictable, but also good to [...] reflect on myself."* (P7)

The R-ESM tool could also potentially help with reflection and sense making of running data after running. P11 reflected, *"if I only see that my heart rate went up, I don't know how I felt at that point, right? [...] I end up [thinking], "...today I had the heart rate of 160. I should have felt like crazy. But I didn't [...]. So, having this [data] [...] especially if I can discuss it with somebody afterwards, I have a way to (support how I) think about it. It's interesting as it puts them in a frame of reference that I am giving, not [just] the heart rate"* (P11). The self-report audio itself could also be reassuring to others and contextualize the sensor data. P11 elaborated: *"[...] This person [friend] is at the maximum of their heart rate, but they are reporting an*

exertion level of three or two [i.e. not high exertion] If somebody like their mother saw this [heart rate data] all the time, she would be worried [...] Like, 'is this too high? [...]' it's interesting to see the mismatch [between the runner and mom's interpretation only based on heart rate]. But, if she could have some reassurance [from the audio recorded self-report] like, 'No, I'm fine, It's fun.'"

6.5.4 Incapability of simple discrete questionnaire for unpacking dynamism and complexity of the running experience

As shown earlier in section 6.3.1, runners SARE in a real-life context were fluctuating and dynamic.

The interviews findings further confirmed the dynamism and variation of the running experience comprising continuous changes such as a pattern of cyclic ups and downs in physical exertion levels and emotional valence. Participants had higher physical and cognitive energy, emotional bounce and excitement in the beginning of the run (e.g., P1, P4, P5, P7, P8, P9) and felt more exertion as run distance increased (e.g., P2, P3). Participants found talking and answering difficult at high exertion levels. P11 said, *"towards the end, I started saying like Exertion level three. I feel I felt like my voice was cutting because I was tired or more tired than before. So [...] There was a bit more 'heehaa' (difficult breathing sounds) in between."* P4 agreed, *"...when I was more tired, would like maybe to say just the numbers, not the whole thing because that took longer and then my breath was a bit short"* (P4). Participants desired that the system could adapt to their physical state when posing questions during running. P3 proposed, *"the system (could) detect that my heart rate is the way too high, my speed is changing frequently, or my breathing is... you know these kinds of measures you might take."*

In the interview, participants explained how they self-reports differently at the different stage. P3 felt pain and exertion during run, but only when towards the finish, his emotional state changed from neutral to positive because of the perceived

achievement. *"I feel the pain in the foot. I feel very tired. But at the end of the session, you have achieved something good, and your emotion changes from neutral to positive."* P3 agreed, *"if you're doing like a particular fast training, if you are not exhausted at the end it means that you didn't put enough effort and you're not improving, whereas if you are exhausted at the end of recovery run, that means that it was not good because you were putting too much effort."*

The interweaving of complex experiences and sensations made it difficult for runners to always be able to separate them for reporting as independent of each other, such as when asked about the sensed exertion, pain, motivation and emotion by the system. For instance, P5 found exertion and pain difficult to distinguish, *"[...]as your pain increases [...], it's harder to distinguish the exertion, like just physically in your lungs and heart from the pain [...]. I think that's partly why my exertion goes up is because like mentally I'm dealing with the pain and just pushing through it. The pain could also potentially lead to negative emotions, "like this feeling heavy or a bit of this pain, it was a bit annoying."* (P4) P11 said that the discomfort or pain might increase her perception of exertion, which also led to negative emotions and reduced motivation. *"since I felt heavy and I didn't feel comfortable and stuff, maybe I also perceived my exertion to be higher. And then if you see my heart rate, it's the same. You know? [...] I remember that I had the pain, and I think it affected negatively everything, so I really wanted to stop, and I was bored within half."*

Participants' attentional focus was dynamic during the study and they could be focusing on their internal state (e.g., body sensations, mental activities) or external things (e.g., environment, other people). The attentional state could either be narrow (e.g., focusing on one thing) or broad (e.g., randomly wondering different things). For instance, P2 was internally and narrowly focusing on self-controlling of the running speed, postures and techniques. *"I was planning actually, planning the session. it is like, like a metronome. it is something like you have to do this this this, check yourself, check the posture. Then I control the water like the time [a couple of minutes] and then push a few a bit more."* In contrast, other runners (e.g., P4, P7) mentioned that

their focus were external or broad, with focus on observing how other runners run to learn their techniques. *“I was noticing other runners and what they're doing, so I feel like I was trying to learn from them.”*

6.6 Discussion

This chapter presented a study to address ***RQ2: How can a ubiquitous ESM based on voice and wearables be designed to gather in-situ self-reports and running data for the purpose of creating SARE-ARS?***

A mixed quantitative and qualitative approach was employed to study the ability of people to report frequently on their states of mind while running, as well as to understand how such R-ESM systems should be designed to enhance self-reporting experience.

Visualizations and analyses of the collected self-reports data demonstrate that the R-ESM was able to capture the fluctuating dynamics in runners' SARE, as well as a balanced distribution of each type of feeling. The think-alouds data not only confirms the importance of the four types of SARE as ground truth labels, but also greatly expands the details with more granular categorizations. This further informs the need to further unpack such rich and complicated SARE when using them as ground truth for machine learning purposes.

Utilizing features extracted from physiological signals (GSR, BVP, HR, Temp) and behavioural signals (arm acceleration, insole pressure data) and machine learning techniques, the study demonstrated that the gathered question-based ground truth led to recognition performances well above chance level in naturalistic conditions. In particular, the best model based on the combination of all signals achieved a weighted average F1 of 0.79 for three exertion levels, 0.84 for three pain levels, 0.78 for two desire-to-stop levels, and 0.85 for three emotion valence levels.

The results from the qualitative data further supported the feasibility of using such systems to collect data during people's personal (not lab) running sessions. Interview results showed that runners did not find the one-minute spaced questions intrusive and were generally able to answer them. This is an important finding in the context of running, since previous studies have either used a lab-controlled protocol (e.g., leading people to a physiological predefined state of exertion [46]), which does not reflect running in the wild, or ESM questionnaires only before and after a running session [98], which is not sufficiently granular for building machine learning models, or required runners to stop at regular intervals to fill in a questionnaire, thus breaking the running experience [203] as well as the validity of the physiological and behavioural data due to the imposed rest. Participants found the system useful, suggesting that such a method could engage runners beyond data collection study sessions by providing them with new opportunities to enhance and facilitate their running. They also highlighted how R-ESM systems should be designed in order to enhance this outcome.

Three design implications for a R-ESM system that emerged from the qualitative studies, as well as how data collection could be improved to build more robust and wider-in-scope running-related affect recognition systems. These include facilitating mindfulness and self-reflection, providing emotional and social companionship for disclosure, as well as providing context-adaptive prompts for more balanced self-reports.

6.6.1 Facilitating mindfulness and self-reflection

The findings of this study (in 6.5.3) demonstrated a ESM-induced mindful introspection process that was perceived by participants as useful, leading to an improved quality and accuracy in subjective self-reports.

As presented, by responding to the ESM questions, runners felt more reflective and aware than in their normal run routine, which further triggered their desire to

investigate running more deeply, ultimately allowing them to better understand their running styles and associated physiological and psychological needs. This further introspection and reflective mental activities were also observed in the think-aloud data, wherein participants expanded upon their self-report responses to the ESM questions with additional information and descriptions that disentangled the different dimensions of their self-reports (pain, exertion, emotion, and motivation). Participants perceived this introspective process as beneficial for enhancing their running experience beyond mere data collection.

Similar effects of ESM-induced introspection have been observed in other studies, for example ESM has been found to help people become more aware of the underlying meanings or reasons for their feelings, beliefs, and behaviours, by encouraging them to reflect on unnoticed events and to more accurately respond to ESM [242]. Such facilitated self-reflection while self-reporting can also be seen in this review paper [86].

Differed from current literature that discovered ESM-triggered self-reflections over longer periods of time (e.g., days or weeks) with fewer ESM questions (e.g., a few per day) in more static contexts (e.g., at home) [242], this study provides evidence of such ESM-triggered self-reflection over a shorter period of time (~30 minutes) with a higher frequency (i.e., one minute) during physically demanding activities in outdoor situations.

Furthermore, this chapter's findings suggest that this mindful introspection process led to improvements in the quality and accuracy of subjective self-reports. For instance, participants reported developing the ability to more reliably respond (e.g. adjust their scales to more accurately quantify their feelings) to the ESM questions, which may indicate an improvement in the accuracy of the data. This further adds to the existing evidence that ESM can improve self-reports data, from the participants' retention rate (i.e., when the ESM activity is perceived as relevant and of value to participants, their retention is maximized [93, 239]) to the accuracy of the data. It

was also observed in this study (section 6.5.1) that the perceived relevance and usefulness would increase participants' acceptance and retention of such self-reporting process during running.

Overall, this introspective and mindful effect can be used as a design implication to guide the design of the R-ESM system, thereby both usefully supporting the runner's own experience and maximizing the quality of data collected for constructing ARS.

6.6.2 Providing emotional & social companionship for disclosure

Interview data from participants showed that converting the R-ESM from a simple prompting-based mechanism into a more conversational and interactive system, such as a chatbot, could result in a more engaging self-reporting experience during running and a richer collection of ground truth. As shown in findings 6.5.2, this interaction with the R-ESM system while running not only evoked a sense of social companionship and positive emotional responses towards the system, as validated by previous studies (e.g., [244, 266]), but also generated further demands for more natural conversations with the system and for it to possess more human-like traits, such as the voices of familiar others. In fact, it has been shown in various studies [137, 138, 152] on the use of chatbots for data collection that participants perceived the data collection task as a social interaction and that they engaged more with questionnaire delivered by a text-based chatbot conversation [129] than with normal questions.

Additionally, a chatbot could facilitate the extraction of more useful information, providing a critical supplement to traditional rating-based self-reports when evaluating complex SARE. As shown in results Section 6.5.4, participants had the opportunity during post-run interviews to discuss the complexity of their SARE, summarizing the change pattern and providing explanations for such changes. For instance, runners discussed how challenging it was to distinguish between high levels of pain and exertion, and how they emotionally appraised pain differently, ranging

from neutral to positive, due to their perceived achievement at the end of a run. Rating-based self-reports in R-ESM were not able to encourage such a level of disclosure.

This further suggests the potential for a conversational chatbot that can prompt users to disclose additional clarification for their rating-based self-reports. In mental health literature, chatbots designed with disclosure behaviours have been demonstrated to encourage more active self-disclosure in participants reporting [148]. In fact, runners in their think-alouds (6.3.4) disclosed a range of discrete emotions in addition to the three classes of affective response (i.e., positive, neutral, and negative) required by the R-ESM in this thesis. Capturing these various facets and discrete emotions could be of immense value, as different emotions (despite sharing the same valence) have different functions and thus influence running differently. Consequently, chatbot technology could be used to facilitate runners' descriptions of their SARE, enabling more specified emotional categories to be applied and thus refining the runners' self-report. This could contribute to a more accurate and nuanced dataset for building recognition systems.

The chatbot could also utilize its natural language understanding capability to extract information about runners' attentional and cognitive state using think-alouds data. As shown in 6.3.4, think-alouds data exposed information regarding runners' attentional and cognitive states, such as associative states (e.g., focusing on improving posture or breathing to reduce injury) and dissociative states (e.g., distracting oneself from running to deal with boredom). This was in line with the literature review discussed in Chapter 2. While this study did not focus on automatically discriminating between moments of associative and dissociative states, findings presented in Chapter 4 indicated the importance of differentiating between these moments to help spectators decide when attracting the runner's attention would be beneficial or detrimental. Such self-reports about associative and dissociative states can be used as ground truth for training an Automatic Recognition System (ARS).

Additionally, designing R-ESM as a chatbot should take into account its use in a publicly shared context. Findings have revealed that runners experienced social embarrassment and refrained from expressing themselves loudly aloud when responding to the R-ESM in a public social context, which is in line with previous research that has shown that talking/interacting with a chatbot in public can be perceived as strange and awkward [102, 153, 258]. Thus, a chatbot has the potential to be socially aware and adaptively disappear or remain silent, yet appear at the appropriate time to prompt users to recall the missing moment in order to maximise data collection coverage.

Building an intelligent and affect-aware chatbot [10] is promising, as Natural Language Processing (NLP) and Sentiment Analysis (SA) software [47, 187] are becoming increasingly powerful, particularly the fast-growing Large Language Models (LLMs) such as GPT-3 [80]. The findings of this study can inform the factors used to fine-tune a chatbot for conversations and their SARE, such as posing questions to characterize the dynamic changing and reasons of SARE, using the topics mentioned in their think-alouds (e.g., running environments and, potentially, personal life events) as topics for small talk.

6.6.3 Providing context-adaptive prompts for more balanced self-report collection

The findings in this study also suggested that the ESM prompts can be designed as context-adaptive to support more tailored and balanced self-report collection.

In this study, the R-ESM system provided prompts at fixed 1-minute intervals throughout the entire running session. However, the fixed sampling rate could potentially fail to capture moments in which runners experienced important feelings. As evidenced by the think-alouds data, runners spontaneously reported additional important self-reports (e.g., knee pain) at times when there was no ESM prompt. These moments were less frequent but had more significant impact on the runners' experience than those in which runners just felt a medium level of exertion without

pain. As a result, not capturing such moments could lead to an imbalanced self-data distribution when training the machine learning model. As demonstrated in Section 6.3.4, the number of collected self-reports regarding high-level pain or negative emotional states is much smaller than that of other self-reports. This lack of high-level pain data can hinder the effectiveness of ML models' learning and prediction of high-level pain moments. As discussed in literature, imbalanced data distribution is a common problem in real-life datasets [87, 154, 213]. By leveraging the chatbot's interactivity, the system can be designed to increase the frequency for collecting self-reports in such critical moments with insufficient numbers, while reducing the frequency of data collection for self-reports with a sufficiently large sample size, thus reducing the burden on users.

The frequency and style of questioning can be tailored to support data collection during special situations. For example, when the exertion level is high, the system can switch to a non-conversational mode and ask questions one at a time, thereby enabling participants to provide very short self-reports (as suggested by one participant).

Furthermore, such context-adaptive prompting may promote increased engagement and reflection, allowing runners to become more mindful of the moments they may have otherwise overlooked. This provides an opportunity to make the system more useful and to provide reflective experiences for runners [56, 151], as well as to produce more reliable self-reports [93] and to facilitate decision-making and behavioral changes [124].

Another important point of discussion is the validity and granularity of the timing and duration of self-reported feelings (referred to as the Window in machine learning terminology [176]). The selection of the optimal Window is a primary focus of research in the field of machine learning. This chapter's findings suggested a solution from a HCI perspective, that the optimal Window could be interactively fine-tuned through a chatbot's conversations with users during the data collection stage.

6.6.4 Discussion on using talking aloud or thinking aloud for gathering runners' self-reports

Our results demonstrate that think aloud has an impact on both the primary task and experience of runners. Participants reported a sense of time passing faster and experienced distractions from physical sensations while engaging in self-reporting via thinking aloud to the R-ESM, potentially leading to a more positive overall experience. The findings indicate that engaging in talking or thinking aloud during the use of the R-ESM has resulted in a socially friendly, mindful, and retrospective experience for runners.

This positive impact of think aloud on runners' experience aligns with existing literature in which runners engaged in talking aloud to express their feelings and monitor their running progress [192, 222]. Such induced positive experience could also partially explain the tendency of rating-based SARE self-report to skew towards middle ranges (i.e., lower physical and pain demands, and a greater desire to continue running) and positive/neutral emotions in those boxplot charts demonstrated in Section 6.3. However, it is important to note that the self-paced nature of runs in our study, which are typically easier compared to racing or competitive running, could also contribute to these positive outcomes.

However, while thinking or talking aloud has demonstrated a positive impact on the overall running experience, it is essential to recognize the potential risks associated with its use as a method for collecting self-reports. One potential concern is the tendency for think-aloud reports to emphasize one type of experience while potentially overlooking other critical sensations that require attention and care. For instance, runners might neglect their discomfort or pain sensations in their lower body while predominantly discussing the excitement of their achievements. As shown in the findings of Section 6.5.3, thinking aloud can divert attention in some of our runner participants. Thus, when employing thinking aloud as a data collection technique, it could be beneficial to design prompts that guide participants' attention

to capture a more comprehensive and accurate representation of runners' experiences, encompassing both positive aspects and critical feedback. Moreover, gathering data using think aloud may require a diverse range of running sessions involving different levels of effort demand, as our study was limited to mostly a single session with a specific level of running.

In addition, it has become evident that incorporating a sense of agency or collaboration in the ESM could be beneficial for facilitating their thinking-aloud process. Participants have expressed a desire for a more interactive and collaborative environment to enhance their engagement and participation in thinking aloud. Some participants in our study seemed to be already perceiving the R-ESM system as a listener or a friend, expressing their thoughts aloud to it during their SARE. They have expressed a desire for an enhanced version of this listener, in the form of a conversational agent that can offer increased interactivity and emotional and social support. These findings align with the concept of addressee perception discussed in the thinking-aloud literature in Section 2.3.1 (e.g., [223]). Hence, incorporating a chatbot as an additional form of listener might serve as a beneficial means to facilitate the thinking-aloud process for an effective but also more meaningful (to participants) data collection. Future research (outside the scope of this thesis) could explore these aspects further by design different types of listeners within the chatbot, taking into account individual preferences (e.g., gender, type of voice, personality).

6.7 Summary

Running as a popular exercise for maintaining welling is received increasing attention in the research field and industry with the aim of building SARE-ARS. This chapter explored the feasibility and viability of using R-ESM for collecting both self-reported ground truth data and behavioural and physiological data for recognizing levels of four experiential states (perceived exertion, pain, desire-to-stop, and emotional valence) during running sessions in the wild. Runners' SARE during running is fluctuating and dynamic. The machine learning recognition performances well above

chance level. In particular, the best model based on the combination of all sensor signals obtained a weighted average F1 of 0.79 for discriminating between three exertion levels, 0.84 for three pain levels, 0.78 for two desire-to-stop levels, and 0.85 for three emotion valence levels.

In addition, analysis of think-alouds and interview data show runners' acceptance of such system during running owe to its non-intrusion, ease and relevance to their activity. More importantly, runners developed an emotional response and a sense of social companionship towards the R-ESM. The R-ESM created a mindful running experience for runners by facilitating introspection and self-reflection. Such findings suggested design implications for R-ESM systems to maximize runners' engagement in using such systems and optimizing the data collection for building wearable runners' experience recognition systems.

This chapter discusses the interview findings and suggests three design implications for the R-ESM. These implications include facilitating mindfulness and self-reflection, providing emotional and social companionship for disclosure, as well as providing context-adaptive prompts for more balanced self-reports. These design implications have the potential to not only result in richer data collection for the development of automatic recognition systems, but also to offer a more personalized, engaging, and useful experience for the runner during the data collection process, ultimately leading to a more natural experience.

Chapter 7 Developing a voice-based Chatbot for SARE self-reporting

7.1 Introduction

Findings in last chapter describes the use of the R-ESM to prompt runners for self-reports every minute during their run. Qualitative findings suggest that the R-ESM should be designed as a chatbot to provide runners with a more conversational, mindful, and engaging self-reporting experience.

In this Chapter 7, we provide a detailed account of the iterative design and implementation of the chatbot prototype. The iteration process began with an initial proposal for the structure and scripts of the chatbot's dialogue. This resulted in an initial dialogue structure that was used as material for expert review study for further improvement. The chapter follows with the technical implementation of the chatbot prototype. The chatbot will then be used as a probe in a user study (in Chapter 8) to determine how the conversation structure should be designed to facilitate self-reporting about SARE.

7.2 Iteration 1: Proposing Three Initial Design of Chatbot's Dialogue

Iteration 1 involves the proposal of three initial dialogues that expand on the R-ESM's probing question: *"What is your level of exertion, pain, desire-to-stop, and your emotional valence?"* The first dialogue structure involves the chatbot asking questions related to exertion, desire-to-stop, and emotion. The second involves the chatbot asking questions related to pain, desire-to-stop, and emotion. Desire to stop and emotion are asked together with both exertion and pain because the findings in Chapter 4 show that they are often interrelated. The third dialogue, led by the chatbot, aims to offer a mindful experience based on participants' requests in Chapter 6. Each initial conversation will be introduced in detail below.

7.2.1 Chatbot's Dialogue for Perceived Exertion, Desire-to-Stop and Emotion

The proposed initial script for the Chatbot to inquire about users' perceived exertion, desire-to-stop, and emotion is presented in Table 5-1. The Chatbot proactively invites users to engage in conversation and initiates the discussion by requesting them to rate their perceived exertion level. Subsequently, the Chatbot seeks to ascertain the factors that contribute to the users' perceived exertion level. This design follows the literature that highlights the subjective nature of exertion, which is influenced by various factors, including muscle sensations (e.g., soreness, tension, and tightness), breathing patterns, cardiovascular responses, subjective heaviness or force, arousal, exercise intensity, and subjective drowsiness [40, 41, 180]. The Chatbot provides prompts for users to consider the potential factors while reporting.

Furthermore, the literature indicates that the rating of perceived exertion is also impacted by an individual's motivational and emotional state. Accordingly, the Chatbot inquiries about the user's affective state together with the perceived exertion, treating them as a cohesive conversation instead of an individual topic. These questions include inquiries about their emotional valence and specific emotion types, as well as their motivational orientation to continue pushing, maintain, or decrease their current effort. By probing these variables simultaneously with perceived exertion, the Chatbot aims to collect self-reports that can contribute to a more comprehensive understanding of the user's psychological and physiological states during running.

Table 7-1: Chatbot’s Dialogue for Perceived Exertion, Desire-to-Stop and Emotion

Role	Example Dialogue Scripts
Bot	Hi there! Is now a good time to talk?
User	Yes
Bot	Can you use a number from 1 to 5 to describe, how difficult is running? 1 means not at all, whereas 5 means almost maximum.
User	4
Bot	What makes you feel the exertion and difficulty? Is it about your muscles, heart, or breathings? or something else?
User	(User gives an open answer)
Bot	What type of emotions can you see in your mind?
User	(User gives an open answer)
Bot	Have you been feeling positive, negative, or neutral?
User	Positive
Bot	Now, I’d like to know about your motivational state. Do you want to push harder? remain the same? or take it easier?
User	Remain the same
Bot	Thank you for sharing.

7.2.2 Chatbot’s Dialogue for Perceived Pain & Discomfort, Desire-to-Stop and Emotion

Table 7-2 outlines the proposed initial script for the Chatbot to inquire about users' perceived pain and discomfort, desire-to-stop, and emotion. Similarly, to the one presented above about exertion, the Chatbot actively engages users in conversation and inquiries about any pain they may be experiencing. If the user responds affirmatively, the Chatbot proceeds with a follow-up discussion. As demonstrated in the literature and findings in Chapter 4, pain is highly subjective and can have physical or psychological origins. It may be caused by injury or discomfort located in certain parts of the body, or it may be a result of increased stimulation or effort, such as that experienced by runners. Therefore, the Chatbot poses questions to help runners characterise their pain. It first inquiries about the location of the pain, followed by whether or not the runner is familiar with the sensation. The Chatbot then asks the

runner to provide a description of the pain, allowing for further characterization before assigning a quantified rating of its intensity.

Additionally, as evidenced by the findings in Chapter 4, some runners may perceive pain as a source of pride or positive experience, viewing it as an achievement or a sign of success in overcoming challenges. As a result, they may be highly motivated despite the pain, while others may not share this perspective. Therefore, the Chatbot further inquiries about the runner's emotional valence, specific emotion types, and motivational orientation to continue pushing, maintain their current pace, or ease up (formed as desire-to-stop earlier).

Table 7-2 Chatbot's Dialogue for Perceived Pain, Desire to Stop and Emotion

Role	Example Dialogue Scripts
Bot	Hi there! Is now a good time to talk?
User	Yes
Bot	Great! I just want to check with you if you feel any pain?
User	Yes
Bot	Oh, I'm sorry to hear that. Can you tell me a bit more? Where is it?
User	(User gives an open answer)
Bot	Do you usually feel such pain during running? or is it unusual?
User	(User gives an open answer)
Bot	Please can you use any word to describe this pain?
User	(User gives an open answer)
Bot	Can you pick a number from 1 to 5 to describe the pain? 1 is very tiny bit, 5 is maximum level of pain
User	3
Bot	What type of emotions can you see in your mind?
User	(User gives an open answer)
Bot	Have you been feeling positive, negative, or neutral?
User	Positive
Bot	Now, I'd like to know about your motivational state. Do you want to push harder? remain the same? or take it easier?
User	Take it easier
Bot	Thank you for sharing.

7.2.3 Chatbot's Dialogue for Mindfulness and Awareness of SARE

Table 7-3 presents the initial script of the chatbot aimed at guiding users towards mindfulness and awareness of their subjective and affective running experience (SARE).

To develop precise chatbot scripts for promoting mindful and reflective running, a lightweight secondary analysis of articles on mindful running was conducted. Thirty articles from the online magazine: Runners' World ¹⁴ were selected using the keywords "*mindful running*." The articles were analysed using a top-down approach, where a coding guideline was proposed based on the definition of mindfulness in literature (e.g., [39, 45, 117, 164]). The coding guideline utilized in this study included the following components: paying attention to (observing), elaboration and description, or investigation of whatever appears [233], acting with awareness, nonjudging of and nonreactivity to inner experience [28], and experiencing the present moment [39].

Applying this coding guideline during the analysis of articles involves identifying what should be attended to, elaborated and investigated during running, as well as how runners practice non-judgmental and in-the-moment awareness. The results of the analysis indicated attentional focus on the body divided into multiple sub-areas (i.e., head, neck, and shoulders; core; arms; and feet), attentional focus on the mind and thoughts, attentional focus on posture and breathing techniques, and awareness of one's surroundings and external cues. Furthermore, being in the moment could be achieved through practicing paying attention to the above, while being non-judgmental involves refraining from suggesting what is wrong and how to regulate or improve it.

¹⁴ <https://www.runnersworld.com>

The aim of our research is not to provide instructions for improving running behaviours, but rather to enhance runners' ability to self-reflect and be aware of their experience. Therefore, the chatbot's scripts were designed to remind runners to pay attention to their mind, body, movement, and environment using phrases such as *“in-the-moment”* and *“non-judgmental.”*

Table 7-3 Initial Chatbot’s Dialogue for Mindfulness and Awareness of SARE

Topics	Chatbot’s Dialogue Scripts
Focus on mind	Now try to observe the thoughts in your mind. Be aware of your thoughts and feelings at the present moment.
Focus on body	Now, we are asking you to scan and pay attention to each part of your body.
Focus on running movement	Now, focus on your posture and breathing. You can make whatever adjustments to feel comfortable.
Focus on external environment	Now shift your attention to what is surrounding you. Pay attention to things such as objects, landmarks, the ground, the weather.
Non-judgement	Please don't judge or associate any things with your observations. Just be aware of their existence

7.3 Iteration 2: Refining Chatbot’s Dialogues

A co-design study with 6 HCI expert reviewers was conducted to thoroughly analyse and improve the existing chatbot scripts to make the chatbot more user-friendly and provide runners with an engaging and useful experience. It aims to gathering contributions to the chatbot's dialogue scripts, including its linguistic choice, grammar correction, speaking style (e.g., word choice, language, personality, avoiding robotic or stilted speech), the chatbot's initial conversation structure and flow, and other potential design spaces.

7.3.1 Method

7.3.1.1 Co-design Study Material

The proposed three Initial dialog scripts were used as probe in the co-design study. The co-design study was conducted on Miro Board¹⁵ remotely.

7.3.1.2 Participants

This included one-to-one review sessions with six English-speaking HCI expert reviewers (three HCI researchers and three HCI designers) who also had interests in running. Participants were recruited from the author's work community.

7.3.1.3 Procedure

In the initial phase, the participants were introduced to the overarching research objective: to create a chatbot that would collect data on runners' subjective and affective experiences during running. Specifically, the chatbot would gather information on perceived exertion, pain, and the runners' motivational and emotional state.

After that, the participants were briefed on the expert review study's purpose, which was to receive feedback on improving the low-fidelity chatbot's script and conversation flow. The participants were informed that the chatbot needed to be designed to facilitate speech-based interaction while running, assuming it had natural language understanding capabilities. To create a more human-like conversation and provide empathetic responses and emotional companionship, the chatbot needed to promote mindful self-reflection and increased awareness of affective experiences.

¹⁵ <http://miro.com>

During the co-design session, a role-playing game was used where an expert reviewer played the role of a chatbot, and the author of the thesis played the role of a runner. The expert reviewer read the chatbot's script aloud to ensure that it sounded natural and did not slip into a formal writing style. The expert reviewer then acted as the chatbot and provided responses after the runner said something.

Three tasks were designed for the review session. Task 1 involved discussing and elaborating the chatbot scripts and creating alternative expressions sentences for the topic of perceived exertion and pain or discomfort. The expected output included a fully developed chatbot script for exertion and pain dialogues, alternative expressions, and design considerations. Task 2 involved the expert reviewers improving dialogues about mindful running with the assistance of a set of mindfulness cards. The expected output included a simple version of the chatbot to guide the mindfulness process. Task 3 required reviewers to iteratively design the overall conversation flow when all merging dialogues, with a focus on discussing the frequency of each question and reducing duplicated and repeated dialogue. The expected output included a final conversation script and flow diagram.

Each session lasted approximately 1.5 hours.

7.3.1.4 Analysis Methods

Following each co-design session, the expert reviewer's feedback and suggestions were incorporated into the initial conversation flow and dialog scripts, which were then used in the subsequent sessions in an iterative process.

7.3.2 Final dialogue scripts and conversation flow

After the co-design with 6 experts, the final conversation flow and seven detailed dialogue scripts were generated. Figure 7-1 presents an overview of the conversation between a runner and a Chatbot, highlighting (in red) 7 types of conversations that were carefully designed with detailed dialogue scripts (see a summary in Table 7-4).

Table 7-4 Overview of Seven Types of Conversation

Table 7-5	Chatbot welcome users and asks users' running plans and goals
Table 7-6	Chatbot proactively invites users to chat
Table 7-7	Chatbot asks user's perceived exertion
Table 7-8	Chatbot asks user's affective state
Table 7-9 & Table 7-11	Chatbot provides personalized and empathic responses based on Users' responses
Table 7-10	Chatbot asks the user's perceived pain or discomfort
Table 7-12	Chatbot's mindful guidance

At the beginning of a running session (highlighted in blue), the Chatbot introduces the purpose and structure of the session, followed by a conversation with the runner concerning their running purpose and plan (refer to dialogue scripts in Table 7-5).

After two minutes, the Chatbot invites the runner to engage in conversation by asking, *"Hi there, do you have a minute to chat?"* (Table 7-6 and Table 7-5 presents all alternative scripts). If the runner declines to chat, they are given the option to provide a reason, although this is not mandatory. If the runner responds affirmatively, the Chatbot initiates a discussion around predetermined topics, which include perceived exertion (refer to Table 7-7), pain or discomfort (refer to Table 7-10), and mindful body scanning exertion (refer to Table 7-12). This is then followed by a discussion about the runner's affective and motivational state (refer to Table 7-8). The Chatbot provides personalized and empathetic replies based on the runner's response (refer to Table 7-9 and Table 7-11). For the topic of mindful body scanning (refer to Table 7-12), the runner is expected to listen to the Chatbot without responding orally.

After a complete conversation, the Chatbot waits for two minutes before initiating a new conversation. It also checks whether a runner has previously reported their pain or exertion during the conversation (highlighted in green). If so, the Chatbot refers back to what the runner said previously and inquires about their current condition.

The conversation is designed to be continuous until the running session ends (highlighted in blue).

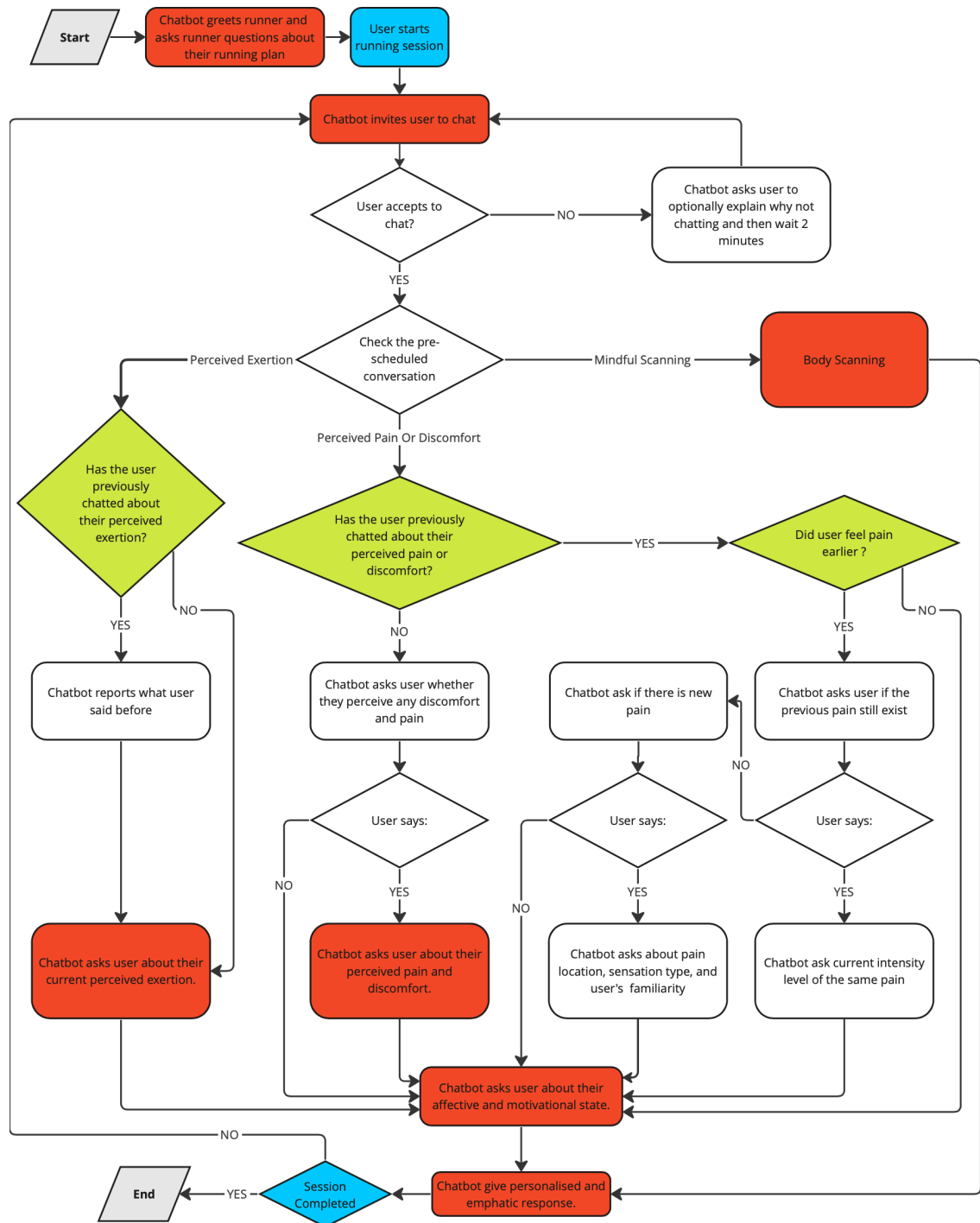


Figure 7-1: Overview of Final Conversation Flow.

The following tables outlines the detailed dialogue scripts. Multiple versions (denoted by V#) were designed to avoid repetitive responses and to provide a more dynamic conversational experience. Table 7-5 outlines the dialogue scripts for a chatbot that welcomes users and asks them about their running plans and goals including their purpose for running, the duration of their run, and the distance they plan to run.

Table 7-5 Chatbot welcome users and asks users’ running plans and goals

Sub-topics	Version	Dialogue Scripts
Welcome & Introduction	V1	<i>“Hi! I am your Run Bot. I am going to ask you a couple of questions while you run to learn more about your experience.”</i>
Purposes	V1	Everyone has their own reason for running. Maybe they are running for performance training, or physical or mental health. Maybe you are running for a charity event, for social reasons, or even to connect with nature. What’s the reason you are running today?
	V2	What’s the reason you are running today?
Duration	V1	For how much time do you plan on running today?
	V2	How much time do you want to spend on today's run?
Distance	V1	How far do you plan to run?
	V2	What's the planned distance?
	V3	How many miles or kilometres do you plan to run?

Table 7-6 presents the dialogue scripts in which the Chatbot proactively invites users to chat while running. If the user responds with “No”, the Chatbot responds with “Can you please explain why? Alternatively, you may simply say no.” If the user again

responds with “No”, the Chatbot further responds with “*Thank you for your response. I hope to speak with you again soon.*” If the user initially responds with “Yes”, the Chatbot initiates the conversation on a predetermined theme.

Table 7-6 Chatbot proactively invites users to chat

Version	Dialogue Scripts
V1	Hi! Is now a good time to talk?
V2	Hi, do you have a minute to answer some quick questions?
V3	Hi, do you have a minute to talk about your run?
V4	Hi, would you like to have a quick chat now?
V5	Hi there, are you free to answer some quick questions?

Table 7-7 presents dialogue scripts for a chatbot to ask a user to rate their perceived exertion on a scale of 1-5. The chatbot also provides an explanation of the scale, with 1 being extremely easy and 5 being maximum effort. The chatbot also asks the user to explain which factors influenced their rating. Factors such as muscle fatigue, stiffness or tightness, breathing difficulty, boredom or stress, uncomfortable sensations, lack of energy, or just simply not being in the right mood are all mentioned as possible influences.

Table 7-7: Chatbot asks user’s perceived exertion

Sub-topics	Version	Dialogue Scripts
Intensity	V1	Every runner experiences effortful moments regardless of their ability. Runners often refer to these subjective efforts as perceived exertion. Can you pick a number on a scale of 1 - 5 to rate how difficult you find the current run? 1 means extremely easy: it's like sitting or walking. 2 means Easy: You feel very comfortable;

Reasons or Factors	V2	3 means Moderate: it's within your comfort zone but you can tell you are working;
		4 means hard: it's outside your comfort zone;
		5 means Max effort: It is at your physical limit or past it. Between 1 to 5, how difficult are you finding the run right now?
	V1	1 being extremely easy and 5 being maximum effort. There are many factors that influence whether a run is difficult or easy. This could be due to muscle fatigue, stiffness or tightness, breathing difficulty, boredom or stress, uncomfortable sensations, lack of energy, or just simply not being in the right mood. Based on which factors did you rate your exertion level as XX (a number)?
	V2	When you rate the exertion level as XX (a number) what factors are you referring to?

Table 7-8 provides dialogue scripts for a chatbot to ask a user about their affective state. The topics covered include emotional valance, emotion description, and motivational status. For emotional valance, the chatbot could ask if the user's emotions are good, bad, or neutral. For emotion description, the chatbot could ask the user to describe their current emotional state. For motivational status, the chatbot could ask the user how motivated they are to complete their running goal.

Table 7-8 Chatbot asks user's affective state

Sub-topics	Version	Dialogue Scripts
Emotional valance	V1	Would you say your emotions are good, bad or neutral right now?
	V2	How would you classify your emotional state? Good, bad, or neutral?
Emotion Description	V1	What words would you use to describe your current emotional state?
	V2	Can you describe what types emotions you currently have?
Motivational state	V1	Think about your running goal for today. How motivated are you to complete it? Are you highly motivated? Or just so so, or lack of motivation?

V2	How motivated are you to still achieve your original running goal? Is your motivation high, middle, or low?

Table 7-9 shows how the chatbot responds differently to users based on their emotional state (positive, neutral, or negative), their exertion level (high, middle, or low), and their motivation level (high, middle, or low). The chatbot provides personalized and empathic responses that are tailored to the user's situation. For positive affective states, the chatbot will provide encouraging dialogue scripts to motivate the user to keep going. For neutral affective states, the chatbot will provide dialogue scripts that are supportive and encouraging. For negative affective states, the chatbot will provide dialogue scripts that are empathetic and supportive.

Table 7-9 Chatbot provides personalized and empathic responses based on users' responses regarding exertion, emotion and motivation

Users' Reports			Chatbot's Feedback Responses
Emotion	Exertion Intensity	Motivation Level	
Positive	High	High	You are doing great! You are certainly enjoying the run! Keep it going!
		Middle	You have been doing well so far! Keep it up!
		Low	<i>"It's normal to feel demotivated, but you have been doing really well so far! Keep it going and you will get the running flow soon!"</i>
	Middle	High	You are doing great!
		Middle	You have been doing great so far!
		Low	You might have pushed yourself very hard. Just take it easier and enjoy the run! "
	Low	High	You are highly motivated for today's run! Keep it going!
		Middle	You are doing great!
		Low	We all have good running days and bad running days. Just take it easy!
	Neutral	High	Great run! I'm proud of your efforts!
		Middle	This has been a good run so far. Good efforts. Keep it up!
		Low	Great run! I'm proud of your efforts

	Middle	High	You have a good running spirit! Keep running!
		Middle	This has been a good run so far. Good efforts. Keep it up!
		Low	It's normal to feel demotivated. Every runner feels this. Keep it going and you will get the running flow soon!
	Low	High	I like your motivation! Enjoy the run!
		Middle	This has been a good run so far. Good efforts. Keep it up!
		Low	It's normal to feel demotivated. Every runner feels this. Keep it going and you will get the running flow soon!
Negative	High	High	Although you are feeling down, I'm glad that you are still full of motivation to finish your goal! Keep the good spirit!
		Middle	This has been a good run so far. Good job. Keep it up!
		Low	It's normal to feel demotivated, since you have reported an exertion level of 4. I know you have been working so hard though! Keep it going and you will get the running flow soon!
	Middle	High	Great run! I'm proud of your efforts!
		Middle	This has been a good run so far. Good efforts. Keep it up!
		Low	This has been a good run so far. Good efforts. Keep it up!
	Low	High	Although you are feeling negative, i'm glad that you are still full of motivation to finish your goal! Good spirit!
		Middle	This has been a good run so far. Good efforts. Keep it up!
		Low	We all have bad days in which we are just feeling demotivated !

Table 7-10 presents dialogue scripts for a chatbot to ask the user about their perceived pain or discomfort. The topics covered include Existence, Bodily Location, Sensation, Duration, Changes, Familiarity, and Intensity. For each topic, the chatbot provides a variety of questions to ask the user. For example, under Existence, the chatbot may ask *“Do you feel any type of pain or discomfort?”* or *“Are you feeling any*

pain or discomfort?” Under Bodily Location, the chatbot may ask “Where is this uncomfortable feeling?” or “Where is this pain or discomfort located?” Under Intensity, the chatbot may ask “Can you pick a number from 1 to 5 to describe the intensity of the sensation?” or “With 1 being barely noticeable and 5 being an intense pain. How would you describe its intensity?” Sensation refers to the user reported sensation.

Table 7-10 Chatbot asks the user’s perceived pain or discomfort

Sub-topics	Version	Dialogue Scripts
Existence	V1	Do you feel any type of pain or discomfort?
	V2	Are you feeling any pain or discomfort?
	V3	Have you felt any type of discomfort or pain?
Bodily Location	V1	Can you tell me a bit more about it? Where is this uncomfortable feeling?
	V2	I'm sorry. Where is this pain or discomfort located?
	V3	Oh, sorry to hear that. Can you tell me a bit more? Where is it?
	V4	Oh. Where do you feel it?
Sensation	V1	I’m sorry to hear that. Can you describe the pain? What type of sensations do you have?
	V2	How would you describe the sensations?
	V3	What words can you use to describe the sensations?
Duration	V1	How many minutes has it been since you started to feel this XX sensation?
	V2	When did you notice it? Can you estimate a time in minutes?
	V3	How long have you been feeling this?
	V2	When did you notice it? Can you estimate a time in minutes?
Changes	V1	How has it changed? Has it gotten better, worse or stayed the same?
	V2	Think about an earlier moment, has it gotten better, worse or stayed the same?
Familiarity	V1	Is this XX sensation on your XX (body part) new, or have you had it before?
	V2	Is this a recurring feeling or is it new?

Intensity	V1	Can you pick a number from 1 to 5 to describe the intensity of the XX sensation? with 1 being just tiny and noticeable, 2 being low, 3 being middle, 4 being high, and 5 being maximum
	V2	With 1 being barely noticeable and 5 being an intense pain. How would you describe its intensity?
	V3	From 1 to 5, how would you describe the intensity level, with 1 being just noticeable and 5 being maximum.

Based on the user's responses to the questions, the chatbot provides personalized and empathic responses. Table 7-11 illustrates how the chatbot responds differently based on the user's emotional state (positive, neutral, or negative), pain or discomfort level (high, middle, or low), and motivation level (high or low). Depending on the intensity of the pain and discomfort, the chatbot's dialogue scripts range from encouraging users to stay positive and observe how the pain changes, to praising them for staying motivated and positive despite the uncomfortable sensations. Furthermore, the chatbot thanks the user, saying *“Thank you for your patience in sharing all these details. It is important to be aware of how you emotionally react to these experiences.”* If the user states that they are not experiencing pain, the chatbot responds, *“That is great. I am happy for you. I am glad to hear that you are not experiencing any pain or discomfort.”*

Table 7-11 Chatbot provides personalized and empathic responses based on users’ responses regarding pain, emotion and motivation

Users’ Reports			Chatbot’s Dialogue Scripts for Responses
Pain Intensity	Emotion	Motivation Level	
High	Positive	High	Great to hear that you are staying motivated and positive despite the uncomfortable sensations! Keep the spirit!
		Middle	You are doing great!

		Low	Stay positive! Be aware of how the pain changes! Please do not push too hard if it is getting worse!
Neutral	High		You are doing great! Great to hear that you are staying motivated and positive despite the uncomfortable sensations! Keep the spirit!
	Middle		You are doing great!
	Low		Having pain or discomfort during running can be annoying and demotivating to every runner. Please observe how these sensations change, how you are affected by them, and how you respond to them. Hope to hear from you again!
Negative	High		This negative feeling doesn't affect your motivation at all. I'm so proud of you!. Remember this spirit!
	Middle		It can be annoying, but you are amazing!
	Low		Having pain or discomfort during running can be annoying and demotivating to every runner. Please observe how these sensations change, how you are affected by them, and how you respond to them. Hope to hear from you again!
Middle	Positive	High	'Glad you stay motivated and positive despite the uncomfortable sensations!
		Middle	This has been a good run so far. Good efforts. Keep it up!
		Low	Stay positive! Be aware how the pain or discomfort changes! Please do not push too hard if it is getting worse!
Neutral	High		You have a good running spirit! Keep running!
	Middle		This has been a good run so far. Good efforts. Keep it up!
	Low		It's normal to feel demotivated. Every runner feels this. Keep it going and you will get the running flow soon!

	Negative	High	Glad you are not affected by the negative feeling. I know you are pushing hard, but please do take care!
		Middle	This has been a good run so far. Good efforts. Keep it up!
		Low	Having pain or discomfort during running can be annoying and demotivating to every runner. Please observe how these sensations change, how they affect you, and how you respond to them!
Low	Positive	High	Glad you stay motivated and positive despite the uncomfortable sensations!
		Middle	This has been a good run so far. Good job. Keep it up!
		Low	Stay positive! Be aware how the pain or discomfort changes! Please do not push too hard if it is getting worse.
	Neutral	High	Great run! I'm proud of you!
		Middle	This has been a good run so far. Good efforts. Keep it up!
		Low	This has been a good run so far. Good efforts. Keep it up!
	Negative	High	Glad you are not affected by the negative feeling. I'm so proud of you!
		Middle	This has been a good run so far. Good efforts. Keep it up!
		Low	Having pain or discomfort during running can be annoying and demotivating to every runner. Please observe how these sensations change, how they affect you, and how you respond to them

Table 7-12 outlines a mindful scanning exercise guided by the Chatbot to help runners become more aware of their bodily sensations during running. The exercise begins by prompting the user to pay attention to any discomfort they may be feeling, observe their thoughts, and identify any negative emotions they may have brought

185

from their daily life. The user is then asked to consider the sensations around their eyes, face, and breath, as well as the tension in their neck and shoulders. They are encouraged to investigate how the discomfort sensations change over time, and to reflect on the sensations in their arms, hands, and fingers. Attention is then directed to their posture, arms swinging, spine, lower back, abdominals, and hips before finally considering how their feet are landing and their ease of movement. Upon completing the exercise, the user is encouraged to remain aware of their bodily running experience. The script wording of this mindful scanning exercise was designed to be relevant to the conversation scripts in Table 7-7 and Table 7-10.

This would not only make the Chatbot's sound more coherent but also help the user to reflect on the topics of perceived exertion and pain or discomfort before conversing with the Chatbot. For instance, the mindful scanning exercise asks the user to *"Identify whether you have brought any negative thought or emotion from your daily life"*, which is relevant to the questions about the user's emotional state. Another example related to the theme of perceived exertion and its factors is, *"Breath is often the source of the perception of tiredness. Be aware of your breath, and how you inhale and exhale"*.

Table 7-12 Chatbot's mindful scanning guidance

Sub-topics	Chatbot's Dialogue Scripts
Introduction	Hi, i hope you are doing well. I'd like to invite you to start a body scanning exercise. This exercise mainly aims to guide you be aware of your bodily sensations during running. You don't have to talk, just listen to me.
Body & Sensation	Now, start by paying attention to the sensations on your head.
discomfort	Notice whether any discomfort exists.
Mind	Observe what thoughts are in your mind right now.
Emotion	Identify whether you have brought any negative thought, emotion from your daily life.

Non-judgement	If yes, just simply observe them.
Body & Sensation	Try to think about the sensations around your eyes, face.
Perceived Exertion & Factors	Breath is often the source of the perception of tiredness. Be aware of your breath, and how you inhale and exhale
Body & Posture	Now, check whether you are holding any tension on your neck and shoulders.
Sensation Changes	Investigate how the discomfort sensations change over time.
Body & Sensation	Think about what the sensations on your arms, hands, and fingers.
Sensation & Exertion	Ask yourself whether any of these sensations make you feel uncomfortable or difficult.
Posture	Pay attention to how your arms are swinging.
Body & Posture	Think about your spine and posture. Noticing whether you are hunching forward or slightly arching your back.
Body	Now, do a scanning on your “lower back, abdominals, and hips.
Body & Posture	Feeling your hips, check if it's rotating excessively or mostly driving forward.
Body & Posture	Think about how your abs (abdominal muscle) are engaged as you stride.
Body & Sensation	Now, scan your legs and feet and notice the sensation around them.
Body & Exertion	Think about the way your feet, how are they landing? Check whether you use the heel, midfoot, or forefoot, and think about the easiness of moving.
End	Well done! You have successfully completed the body scanning exercise. Hope you become more aware of your bodily running experience. Enjoy the run, and I'll speak to you soon.

7.4 Iteration 3: Implementing the Chatbot

This section includes the technical details of the APIs/platforms used for implementing the chatbot prototype. Voice recognition and natural language understanding capabilities are provided by the Google Cloud Speech-to-Text¹⁶ and Dialogflow¹⁷ APIs, respectively. The chatbot is implemented as a mobile application using the React Native framework. The application is connected to the internet via a mobile phone, allowing for use in outdoor settings. Finally, the chatbot is designed to provide empathetic responses and emotional companionship, with pre-designed dialogues scripts rather natural language generation to ensure interaction reliability.

7.4.1 A Hybrid Approach: Macro and Micro Conversation Pathways

Figure 7-2 outlines five components in the Chatbot prototype's architecture: the Runner component, the Chatbot Voice Interface component, the Macro Conversation Pathway (**Macro-CP**) component with a Proactive Interaction Mechanism (**PIM**), the Micro Conversation Pathway (Micro-CP) component, and the Cloud Fulfilment component. **The Runner component** serves as the entry point for users engaging in conversations and self-reporting. **Chatbot Voice Interface component** is responsible for facilitating speech-based interactions between the end user and the Mobile Phone App. This is achieved through the App's speech recognition of the end-user's voice, which is captured by the microphone in the end-user's earphone. The App then conveys the chatbot's response, conversed using text-to-speech, to the end-user via the earphone. (See Appendix 11.4 or a demonstration how the conversation was implemented following the architecture).

¹⁶ <https://cloud.google.com/speech-to-text/>

¹⁷ <https://cloud.google.com/dialogflow/>

The Macro-CP component is responsible for a controlled and proactive approach to trigger and execute the desired conversation topics (see Table 7-4) between the end user and the chatbot, ensuring reliable sampling and gathering of end user's self-reported experience. The Macro-CP component uses a program to control the timing, sequence, and frequency of each conversation topics listed in Table 7-4, and employs a proactive interaction mechanism (PIM) to proactively invite the end user to chat with the right conversation context being properly prepared for each topic. The Macro-CP component is implemented on the backend of the Mobile Phone App. Once the end user accepts to have a conversation on one topic, the Micro-CP component takes action.

The Micro-CP component is responsible for natural language understanding (NLP), performing interactive conversation turns regarding each predetermined sub-topic (see from Table 7-5 to Table 7-12), and extracting information from an end-user's expression. It is implemented with NLP-based platform Google DialogFlow that provides powerful natural language understanding capabilities to comprehend users' different queries and responses. Lastly, the **Fulfilment component** is responsible for performing backend processing, such as data storage and computations, as well as providing dynamic and empathic responses.

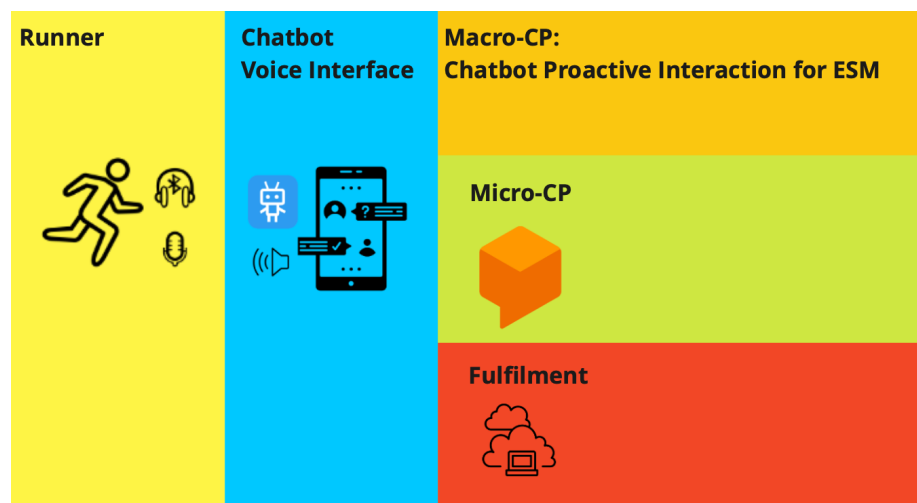


Figure 7-2 Five components in the Chatbot prototype's architecture

7.4.2 The Macro-conversation pathway (Macro-CP) with a controlled and proactive interaction mechanism (PIM)

The macro conversation pathway primarily involves a proactive interaction mechanism, wherein the chatbot invites the user to engage in conversations on pre-scheduled topics. However, this desired proactive interaction mechanism is not directly supported by the Google DialogFlow platform, as it is reactive by default and designed with a “*request-response*” mechanism that requires the end user to initiate the conversation.

It was challenging to implement the proactive interaction mechanism using DialogFlow. Although the DialogFlow provides some potential solutions, such as using the provided “*Event*” framework and the *agent.setFollowupEvent()* function to trigger another intent without the end-user’s expression, this does not seem to allow for the control of the timing of proactive triggering. Furthermore, adding this function may also override the existing responses to users. In addition, DialogFlow provides a “*Welcome Intent*”¹⁸ that can proactively interact with an end user, but this can only be used once when an end user begins the conversation, not recurrently throughout the session. Another approach is to utilize “*Push Notifications*”¹⁹ from a mobile phone’s Operation System level to initiate interaction with the end user. Nevertheless, this also has constraints, as the end user has to respond to the push notification, which is challenging in our research context, in which hands-free interaction is required, as a runner would not likely wish to physically operate a phone while running.

Designing proactive interaction mechanisms for a chatbot also has gathered interest in the HCI community. For example, [271] implemented a proactive speaker based

¹⁸ <https://cloud.google.com/dialogflow/es/docs/intents-default>

¹⁹ <https://miningbusinessdata.com/api-ai-push-notifications/>

on Google Home, using a Raspberry Pi to deliver pre-recorded voices to trigger the “*Be Proactive*” intent. Google Home then responded with proactive voice interactions, such as “*Is now a good time to talk?*” The Raspberry Pi was physically close to Google Home, but was kept invisible to participants in the experimental setup. The pre-recorded voices were also designed to be at a very low volume so that the participant could not hear them, yet they were still effective in activating the Google Home device. While this hardware-based setup is inspiring, it is not feasible in our study context, as runners cannot carry a smart speaker and the added gadget - Raspberry Pi - during their run. Therefore, we sought digital solutions on mobile phones, which are more accessible during running. As a result, we developed our own voice-based chatbot on a mobile phone, rather than using existing voice-based smart speakers like Google Home.

Therefore, we propose a proactive interaction mechanism in the macro conversation pathway. This mechanism involves a program written in the local Mobile Phone App that proactively invokes the context of an upcoming conversation using key phrases that are identifiable to each scheduled conversation topic. DialogFlow will then respond with messages inviting the user to engage in conversation about the scheduled topic, which will be delivered to the end user via speech by the Mobile Phone App. This approach allows the end user to be invited to a conversation with the context of the conversation already configured, thus improving the accuracy of the chatbot's natural language understanding capabilities.

In order to provide a more natural and hands-free voice interaction experience, the mobile Chatbot App was designed to automatically switch on the microphone, eliminating the need for the user to activate it. This allows users to simply speak after being invited to do so. However, sometimes people may not be able to speak immediately, as they need time to think of what to say. This prolonged speaking time can cause errors in voice interaction. When a user does not talk within two seconds,

the current Google Speech Recognition API: *SpeechRecognizer*²⁰ will treat it as if the user is not talking and return a *"No_match error"*, thus ending the current conversational turn and interrupting the conversational flow. As the *"SpeechRecognizer"* API does not provide any function like *"wait-until-I-speak"*, our solution is to program the App to keep switching on the microphone until a user begins to speak.

7.4.3 The Micro-Conversation Pathway (Micro-CP) using DialogFlow for more powerful and flexible NLP

Google DialogFlow²¹ was employed to implement the conversations in the Micro-CP. This platform provided a range of components to facilitate the construction of the Chatbot's conversation flow, including Intents, Training Phrases, Parameters, Entities, Slot Filling, Prompts, Contexts, and Fulfilment, which were all utilized to construct the chatbot for this study. Details about their implementation are documented in Appendix-11.4.2. For example, the Intent: *"chatPain"* describes the end user's experience regarding pain. Once the end-user input resembles one of the Training Phrases, the associated intent is matched and triggered (referred to as Intent Matching). To improve the chatbot's natural language processing ability, we trained the chatbot with some feelings and sensations related vocabularies. A selection of them can be seen in Appendix E-Chapter 7.

Parameters were then used to extract values from the end-users' inputs. Appendix-11.4.2. shows all parameters and the example values in this chatbot. For instance, when the end-user says *"My lower back feel sore"*, the Parameters *"\$pain_location"* will extract *"lower back"* as the value. The extracted parameters' value needed to be categorised with an information type using the Entity component. For instance,

²⁰ <https://cloud.google.com/speech-to-text>

²¹ <https://cloud.google.com/dialogflow>

the Entity called “@human_body_parts” can categorize information about different human body parts. In addition, we used pre-defined entity types, e.g., Entity: “@Number”, to categorise the value (e.g., 3 out of 5) extracted from the Parameter @pain_intensity. If the end user did not give the value, we used the “Slot Filling and Prompt” component to prompt the user again until they gave the desired answer. To make DialogFlow more accurately match end-user’s expressions to an intent, we also used the Context component to configure Input Context and Output Context for an intent so as not only improve the matching rate but also helps with information transition between intents.

The responses of the chatbot were implemented by leveraging both the simple static responses provided within the Intent component and more dynamic, advanced, and personalized responses enabled in the Fulfilment component. Additionally, Speech Synthesis Markup Language (SSML)²² was utilized to generate more natural and easy-to-follow audio responses for end users. For instance, the chatbot's responses were set to a low pitch using the Prosody element of SSML, as research has found that a low pitch provides a more calming and relaxing experience than higher pitches of a robotic voice [189]. Furthermore, during the mindful scanning exercise, SSML was used to set the chatbot's speaking rate to “slow” and with “break time” (e.g., 3 or 5 seconds) before and after keywords and sentences, in order to allow users more time for self-reflection of their feelings and body sensations while running the exercise.

7.4.4 Chatbot dealing with exceptional conversations

The chatbot was programmed to handle exceptional conversations where users' inputs were not recognized. In such cases, the chatbot triggers the Fallback intent²³,

²² <https://www.w3.org/TR/speech-synthesis11/#S3.2.4>

²³ <https://cloud.google.com/dialogflow/es/docs/intents-default#fallback>

which is a solution for handling exceptional situations in DialogFlow. The chatbot responds automatically with a pre-defined message to prompt users to repeat their input, such as *"Sorry, I didn't get it. Could you please say it again?"* or *"Sorry, I missed your answer. Would you mind saying it a bit louder or slower?"* or *"Sorry, could you please rephrase it and say it in a full sentence?"* Although fallback messages can increase the coverage of data collection, too many fallbacks can affect the quality and natural flow of the conversation. Therefore, the fallback prompts were restricted to a maximum of three times in a row.

Additionally, the chatbot was configured with a master termination approach for the user to stop the conversation in any case when the conversation becomes problematic or if the user does not want to continue the conversation for any reason. In such case, a user just has to simply *"Bye"* or *"Bye-bye"*, the chatbot will stop the current conversation and wait for 2 minutes to invite the user chat again.

7.4.5 Testing and debugging Prototype

The chatbot was tested with two friends during outdoor running sessions. However, the chatbot application was unable to remain active in the background of the phone's operating system and became dormant. Additionally, when the phone's screen was locked, the app also became inactive. After conducting research on potential solutions, we realized that programming a solution would be too complex and time-consuming. Therefore, we decided to keep the chatbot running in the foreground of the mobile phone with the touch screen activated. However, this presented new challenges as participants frequently touched the phone screen accidentally, causing interruptions to the chatbot conversation by activating the drop-down menu or navigation button on an Android phone (Figure 7-3). To address these issues quickly, a simple solution of using cardboard to prevent accidental screen touch was used, as illustrated in Figure 7-4. This was a relatively straightforward fix that enabled us to proceed with the study and utilize the chatbot prototype in the experiment presented in the next chapter.

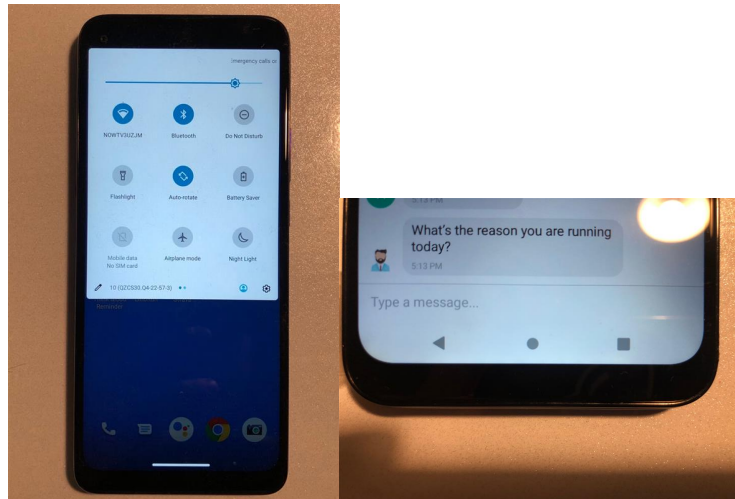


Figure 7-3: Drop-down menu (left) and navigation button (right)

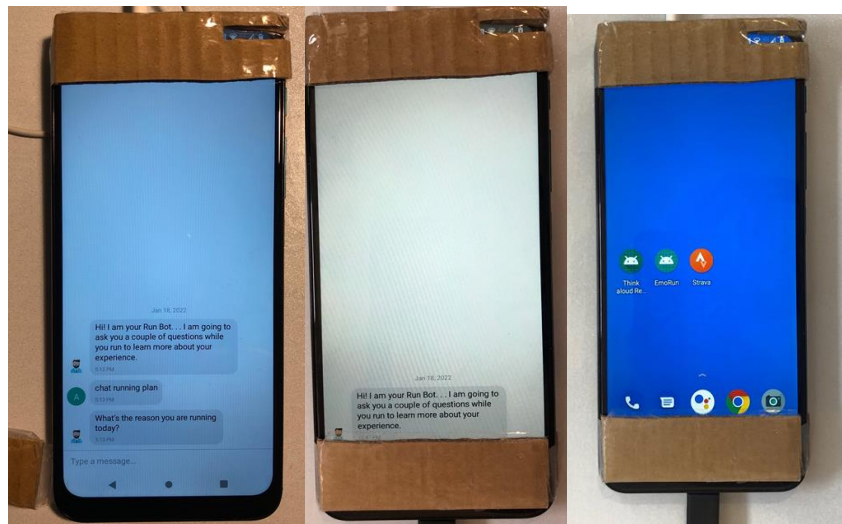


Figure 7-4: Using cardboard to disable accidental touch.

Chapter 8 Understanding runners' experience of using Chatbot for SARE self-reflecting and reporting

8.1 Introduction

Chapter 7 presented a voice-based chatbot prototype to be used as a research probe. In contrast to the R-ESM system in Chapter 6, which only delivers the same speech-prompts following a fixed schedule, the prototyped chatbot engages in initial small talk, as well as proactively invite runners to discuss topics of interest. Furthermore, the chatbot was designed to report back to the runners what they had said before, and guide them in engaging in mindful scanning while running. The objective of this Chapter 8 is to understand how such more conversational, even if still simple structures, support SARE self-report during long-distance running.

RQ 3: How can chatbot conversation flow be designed to facilitate voice-based SARE self-reports while running?

This chapter is structured as follows: Section 8.2 details the methods used in this study, followed by Section 8.3 which presents the quantitative findings. Section 8.4 presents the qualitative findings, while Section 8.5 provides a discussion of the results and their design implications.

8.2 Methods

An in-the-wild study was conducted, in which runners were invited to use the designed chatbot prototype for SARE self-reporting during their own running and to compare it with the R-ESM approach used in Chapter 6. The R-ESM prompt was slightly modified from a rating and number-based question to an open-ended question that prompted runners to 'think aloud' their thoughts and feelings more strongly as they reflected on their subjective experience. The reason for exposing the participants to the two forms of self-reporting (R-ESM vs chatbot) was to increase participants' awareness of how a chatbot-based conversation could facilitate or hinder self-reflection and lead to more comprehensive interviews after the running

sessions. The semi-structured interviews were analysed using thematic analysis, and conclusions were drawn concerning the design of a suitable chatbot with accompanying design recommendations.

8.2.1 Participants Recruitment

In this study, nine participants (five females, four males, refer to Table 8-1 Participants Profile) were recruited through online advertising on social media and local running clubs. To be eligible, participants had to be 18 years or older, capable of running for at least 30 minutes at a comfortable pace, and have established plans for two runs. It is noteworthy that the participants were not recruited to run for the study, but rather to trial the chatbot prototype within their own running routines. Each participant received a £25 Amazon voucher as compensation for completing the study. None of the participants had taken part in any of our previous studies. In total, we recruited nine participants, comprising five females and four males.

Table 8-1 Participants Profile (Indoor Treadmill (IT), Outdoor Track (OT), Outdoor Park (OP), M: Male, F: Female)

Participant	Session 1 (Min)	Session 2 (Min)	Gender	Age	Location	Running Experience
P1	22	32	M	30	OP	Beginner
P2	53	30	F	55	OP	Intermediate
P3	51	N/A	F	42	OP	Advanced
P4	42	36	M	60	OP	Advanced
P5	30	43	F	34	OP	Intermediate
P6	30	59	M	50	OP	Advanced
P7	35	31	F	39	OP	Advanced
P8	29	28	M	32	OP	Beginner
P9	28	18	F	25	OP	Beginner

8.2.2 Study Procedure

In adherence with UCL and UK Government guidelines, a COVID safety procedure was implemented to minimize the potential risk of COVID infection during the study. This included strict disinfection of devices and adherence to social distancing protocols when meeting participants outdoors. Participants were provided with experiment devices and prototypes, which included a smartphone pre-installed with both the R-ESM App and the Chatbot App, a waist belt phone holder, and headsets. Participants were given clear instructions on how to use the prototypes and were left to independently use them without any intervention or support from the researcher. It is important to note that the pressure insoles were not included in this study as they were no longer the focus of this thesis. Additionally, their use required support from the researcher and the insole prototype was found to be uncomfortable during running, which could have affected the participants' self-reporting experience.

8.2.2.1 Session 1: Running with the R-ESM with open-ended prompt every 2 minutes

In the first running session, participants self-reported to the R-ESM app that provided speech-based prompt during running. This prompting was designed to slightly differ from the rating and number-based question in Chapter 5. It was an open-ended audio prompt that asked runners to 'think aloud' their thoughts and feelings more strongly as they reflected on their subjective experience, saying: *"Hello, could you please think aloud about your current feelings, thoughts, and experiences right now? You can talk about anything you want -if you're not sure what to say, physical exertion, discomfort, pain, emotion, and motivation are all good topics to start with?"*. The prompts were scheduled at the interval of every 2 minutes to be close to the conversation frequency in the second running session. Participants were informed that they had the flexibility to decide what to verbalize (which could be one or more of the suggested topics, or anything else related to their feelings) and when to talk (which could be any other time even without prompts). This process was audio-recorded by the App throughout the run.

Prior to the first running session, participants were provided with a brief training which included two exercises to ensure they were familiar with the think aloud procedure. The first exercise was the “*Naming Task*”, as described in [222], wherein participants were instructed to verbalize the thoughts that passed through their minds as they named twenty different animals. The second exercise was designed specifically for this study, wherein participants were instructed to verbalize their thought processes as they reflected on their current feelings regarding one of themes of discomfort or pain, exertion, emotion, and motivation.

8.2.2.2 Session 2: Running with the Chatbot

During the second running session, participants used the Chatbot prototype to report on their experiences with regards to discomfort or pain, exertion, emotion, and motivation. Prior to the run, participants received instructions on how to interact with the Chatbot using voice commands. They were trained to handle unexpected issues that might arise due to technical limitations in chatbot speech recognition, such as noise or low volume. They were shown how the Chatbot would respond to such issues by asking them to repeat or rephrase their statements or to speak slower or louder. Participants were also informed that they could end a conversation at any time by saying “*Bye-bye*” in extreme cases and start a new conversation two minutes later.

The two running sessions were arranged on different days within the same week. Participants were interviewed after each session in semi-structured interviews. The first interview was shorter (around 20 min) and focused on their experience of thinking aloud SARE while running. The second interview was longer (between 55 and 70 minutes) and focused on the experience of conversing with the Chatbot compared to the R-ESM. The interviews began with general questions about their running experiences to warm up the participants.

The average duration of session 1 was 35.56 minutes (Std = 10.78), while session 2 had an average duration of 34.62 minutes (Std = 12.12).

The participants' conversations with the Chatbot were logged locally on the mobile system, including details of the Chatbot triggered intents. The think-alouds and self-reports with the R-ESM were also logged.

8.2.3 Data Analysis

All interviews were audio-recorded and transcribed analysed using thematic analysis. Findings were present in Section 8.4

The logged conversation history with the chatbot was quantitatively analysed to evaluate how effective the designed chatbot was functioning during the study process. The measurements extracted include the number of ***Interactions***, ***Unsuccessful interactions***, and ***Fallback***. These terms are defined in Google DialogFlow²⁴: ***Interactions*** refer to a conversation turn between the user and the chatbot (i.e. a chatbot poses a question or suggest a task, and the participant responds); ***Unsuccessful interaction*** refers to a situation in which the user's phrases were not matched with any intent and, thus, not recognised by the chatbot. In such a case, the chatbot triggers the ***Fallback*** intent to prompt users to repeat again for maximum three times in a row (see details of the prompts in 7.4.4). These three basic measurements can be used to infer the quality of conversation data, users' conversational experience and behaviours, and the success of fallback interactions. In this thesis, we further propose the following metrics: ***Fallback Rate, Fallback Success, Conversation Smoothness, and Conversation Usefulness*** (with respect to the self-reporting task). We hence computed: 1) the number of fallback interactions triggered, 2) their length and temporality, and 3) whether the fallback interaction

²⁴ <https://cloud.google.com/dialogflow/cx/docs/basics#interaction>

successfully maintained the conversation and obtained useful information. The rationales and formulas for computing these metrics will be explained in the following Section 8.3.

8.3 Quantitative Results: Chatbot and Conversation Quality

Before presenting the main findings from the interviews, the effectiveness of the designed chatbot in supporting the collection of runners' self-report data was measured using the following metrics. It should be noted that as of now, we do not have a baseline for comparison and will present these metrics as an individual assessment.

Fallback Rate, referred to as how often the Fallback Interactions occurred among all Conversation interactions.

$$Fallback\ Rate = \frac{Count(All\ Fallback\ Interactions)}{Count(All\ Conversation\ Interactions)}$$

The Fallback Rate is used to infer the conversation smoothness and quality. A higher Fallback Rate indicates more problematic conversations with lower speech recognition or Natural Language Understanding performance.

Therefore, **Conversation Smoothness** can be computed as

$$Conversation\ Smoothness = 1 - Fallback\ Rate$$

Furthermore, we categorized ***Fallback*** into the following subtypes: ***FallbackSuccess***, which is further classified as ***1FallbackSuccess***, ***2FallbackSuccess***, and ***3FallbackSuccess***, and ***FallbackFailure***, based on the observations of fallback interactions in our study. More specifically, **FallbackSuccess** identifies instances when the chatbot's fallback response successfully prompted the user to continue the conversation. Upon the first fallback response being triggered, if a user's expression was successfully matched with an intent, this was named as **1FallbackSuccess**. If not,

a second fallback was induced. If a user's expression was successfully matched with an intent, this was termed **2FallbackSuccess**. If this too was not successful, a third fallback was initiated. If the user's expression was successfully matched with an intent in this instance, this was referred to as **3FallbackSuccess**. If not, this was referred to as **FallbackFailure**, followed with the chatbot ending the current conversation to avoid an endless dead loop. **1FallbackSuccess** triggers one conversation interaction. **2FallbackSuccess** triggers two conversation interactions. Both **3FallbackSuccess** and **FallbackFailure** trigger three conversation interactions. Therefore, the total counts of Fallback Interactions can be counted as below:

$$\begin{aligned}
 &Count(All\ Fallback\ Interactions) \\
 &= Count(1FallbackSuccess) \times 1 \\
 &+ Count(2FallbackSuccess) \times 2 \\
 &+ Count(3FallbackSuccess) \times 3 + Count(FallbackFailure) \times 3
 \end{aligned}$$

In addition, we have further specified **FallbackSuccess** into two specific types: **"FallbackSuccessPertinent"** or **"FallbackSuccessNotPertinent"**. According to [145], conversation breakdowns can include 1) when the chatbot fails to recognize the user's intent and asks them to repeat their request, or 2) when the chatbot responds with "false positive" responses that may not be relevant to the user's query. Building on this, we examine the pertinence of users' responses to the chatbot's fallback questions. Specifically, we differentiate between positive and false positive responses. A positive response occurs when the user provides an answer that is relevant to the question, and the chatbot can proceed without further fallbacks. In contrast, a false positive response happens when the user provides an answer that is not directly related to the chatbot's question but still satisfies the conversation, preventing further fallbacks.

"FallbackSuccessPertinent" refers to a positive situation when the chatbot's Fallback message successfully triggers the user to respond with relevant information that are

202

pertinent to its question (highlighted in Green in Table 8-2). For example, as shown in participant 1's conversation data when the chatbot asked, "How motivated are you to still achieve your original running goal? Is your motivation high, middle, or low?", P1 responded, but possibly due to noise, their answer was recognised by the chatbot as "hello". The chatbot then triggered the Fallback message to ask again, "Sorry I did not understand it. Would you mind saying it again?". P7 followed with "low", which is an answer with relevant information to this conversation context. Such a response is pertinent to the chatbot's question; hence they are more likely to be used as useful self-report data.

"FallbackSuccessNotPertinent" refers to a false positive situation when the participant's response is not related to the chatbot's question, despite the conversation can continue (highlighted in Blue in Table 8-2). For instance, the chatbot asked P7 a question, *"Great to hear that! How are you feeling emotionally? Good, bad, or neutral?"* and received an unrecognised answer *"Curtis"*. Then, it triggered the Fallback message to ask P7 to say it again. In this case, P7 replied with *"You're so clever"*, which triggered the system's predefined intent (e.g., for small talk) and smoothly and successfully the chatbot managed the exceptional situation. P6 also used *"Love you!"* that helped to overcome the fallback situation. Alternatively, participants (P2, P5, P6) also used phrases such as *"Bye Bye"*, *"Stop"* to terminate the current conversation proactively, which was seen as successful fallback because users proactively and smoothly handled the exceptional conversation. However, such response was not pertinent to the question and cannot be used as SARE self-reports for the research purpose. Therefore, the relevance of the conversation can be measured as **Conversation Pertinent**:

Conversation Pertinent =

$$\frac{\text{Count (All Conversation Interactions)} - \text{Count (FallbackSuccessPertinent)} - \text{Count (Fallback Failure)}}{\text{Count (All Conversation Interactions)}}$$

The type of *"Fallback Success"* can indicate whether a participant is actively engaged in the self-reporting experience. Among participants who have the same *"Fallback*

Success” rate, the one with a higher ***“FallbackSuccessPertinent”*** rate may be more actively engaged in the self-reporting task than the one with a higher ***“FallbackSuccessNotPertinent”*** rate, who might be less patient when interacting with the chatbot. Therefore, Participants’ Engagement in self-reporting task is measured as:

Task Engagement

$$\begin{aligned}
 &= (1 \times \text{Count}(1\text{Fallback SuccessPertinent}) \\
 &+ 2 \times \text{Count}(2\text{FallbackSuccessPertinent}) \\
 &+ 3 \times \text{Count}(3\text{FallbackSuccessPertinent})) \\
 &/(\text{Count}(\text{All Fallback Interactions}))
 \end{aligned}$$

Table 8-2 reports the findings about the exceptional conversation interactions occurred in each participant's conversation data. The first and second column shows participant ID and Fallback Interaction ID respectively. The third column, *“Chatbot”*, shows the conversation context in which the chatbot asked questions. The fourth column, *“Unrecognised User Phrases”*, shows the unrecognised user phrases that triggered the Chatbot's fallback message. The fifth column, *“User Phrases after 1st Fallback”*, shows the user phrases after the first fallback message. The sixth column, *“User Phrases after 2nd Fallback”*, shows the user phrases after the second fallback message, only if the previous user phrases were unrecognised. The seventh column, *“User Phrases after 3rd Fallback”*, shows the user phrases after the third fallback message, only if the previous user phrases were unrecognised. The eighth column, *“Chatbot End”*, shows the chatbot ending the conversation if the previous user phrases were unrecognised.

A colour scheme of red, green, blue and yellow was also used to highlight various aspects of the table. Red highlights the ***“Fallback Interactions”*** that occurred, referring to those situations in which user phrases were still not successfully recognised after the fallback message, thus triggering further fallback responses. Green highlights the ***“FallbackSuccessPertinent”*** situation, in which user phrases

were successfully recognised after the fallback message, allowing the conversation to continue, and were also related to asked question. Blue highlights the ***“FallbackSuccessNotPertinent”*** situations, in which user phrases were successfully recognised after the fallback message, allowing the conversation to continue, but were not related to the asked question. Finally, Yellow highlights the ***“FallbackFailure”*** situation, in which user phrases were not successfully recognised after receiving fallback messages for three times, resulting in the chatbot ending the conversation.

Table 8-3 presents the analysis results of Conversation’s Fallback Rate, Conversation Smoothness, Conversation Usefulness, and participants’ self-reporting Task Engagement. Figure 8-1 presents an overview of the overall conversation quality.

Table 8-2: Summary of Fallback Conversation Interactions occurred. Green = “FallbackSuccessPertinent”; Blue = “FallbackSuccessNotPertinent” i.e., participant’s response correctly understood but not pertinent to the chatbot’s question; Red = “Fallback Interactions” i.e. response not recognized triggers further fallback response. Yellow = “FallbackFailure” i.e., three subsequent fallback response attempts lead chatbot ending the conversation.

P_ID	F_ID	Chatbot	Unrecognised User Phrases	User Phrases after 1 st Fallback	User Phrases after 2 nd Fallback	User Phrases after 3 rd Fallback	Chatbot End
1	1_1	“Can you tell me a bit more about it. Where is the pain?”	lag	bye-bye			
1	1_2	Do you feel any type of pain or discomfort?	little bit	a little bit	a little bit	Bye	
1	1_3	How motivated are you to still achieve your original running goal? Is your motivation high, middle, or low?	hello	low			
2	2_1	Sorry to hear that. Can you tell me a bit more about it. Where is the pain?	niece	my niece	knee		
2	2_2	None	I'm feeling ok my police is a bit wobbly but that's all	I'm feeling ok my police is a bit wobbly but that's all	Kelly's is customers repeat everything aching but otherwise I'm feeling fine	customers repeat everything	YES
2	2_3	With 1 being barely noticeable and 5 being a maximum intense pain. How would you describe the intensity of the pizza feeling?	it's me	it's a 3			
2	2_4	I hope you are having a good time. Just to check, have you felt any type of discomfort or pain during the run?	Kelly's	my Achilles is aching as we're going up	at the back of my left		
4	4_1	Hi, do you have a minute to answer some quick questions?	little 2 to answer	what do you want me to say	warrant	I was speaking to a runner who went past me	YES
4	4_2	Hi, do you have a minute to answer some quick questions?	I can answer different way	i	was	find thesaurus	YES
4	4_3	Hey there! From 1 to 5, how difficult do you feel right now?	probably about 2	I am feeling difficulty of about 2			

4	4_4	What words would you use to describe your current emotional state?	it would be neutral	my emotional state is	still neutral, nearly finished		
5	5_1	What words can you use to describe it?	type shut up	Pipe Shop	Sharp	painful	
5	5_2	How motivated are you to still achieve your original running goal? Is your motivation high, middle, or low?	hi	hi	hi	middle	
5	5_3	Can you tell me a bit more about it. Where is the pain?	in the way	magnets	interacting with it	I will come	YES
6	6_1	Glad to hear that you don't experience any type of discomfort or pain. How would you classify your emotional state? Good, bad, or neutral?	bullet	i	bye-bye		
6	6_2	Are you emotionally feeling good, bad, or neutral?	up	feel good			
6	6_3	Where is the pain?	no pain	no pain. You misunderstood			
6	6_4	Hey there! You said your exertion level was 3 because of level of fatigue. What's your exertion level now?	level 84	level four	4	stop	
6	6_5	Good, bad, or neutral?	tell us the Co-op wages so I stopped I'd look at it	love you!			
7	7_1	Are you feeling any type of discomfort or pain?	i	no			
7	7_2	Great to hear that! How are you emotionally feeling? Good, bad, or neutral?	Curtis	bye-bye			
7	7_3	Are you emotionally feeling good, bad, or neutral?	i	bye-bye			
7	7_4	Can you explain or describe your positive feeling?	positive	positive	feeling good running	just feel very happy	
8	8_1	Great to hear that! How are you emotionally feeling? Good, bad, or neutral?	New	Neutral			
8	8_2	How motivated are you to still achieve your original running goal? Is your motivation high, middle, or low?	price	ebay	I'm very motivated now		
8	8_3	Hi there, are you free to answer some quick questions?	death	yes			
9	9_1	"Can you tell me a bit more about it. Where is the pain?"	lag	bye-bye			
9	9_2	Sorry to hear that. Can you tell me a bit more about it. Where is the pain?	niece	my niece	knee		

Table 8-3: Analysis of Conversation's Fallback Rate, Conversation Smoothness, Conversation Pertinent, and participants' self-reporting Task Engagement.

P_ID	All Conversation Interactions	1 Fallback Success	2 Fallback Success	3 Fallback Success	Fallback Failure	Total Fallback Interactions	Fallback Success	Fallback Success Pertinent	Fallback Success Not Pertinent	Fallback Rate	Conversation Smoothness	Conversation Pertinent	Task Engagement
1	45	2	0	1	0	5	3	1	2	0.11	0.89	0.96	1.00
2	49	1	2	0	1	8	3	3	0	0.16	0.84	0.98	0.63
4	56	1	2	0	2	11	2	2	0	0.20	0.80	0.96	0.45
5	58	0	0	2	1	9	2	2	0	0.16	0.84	0.98	0.67
6	65	3	1	1	0	8	5	2	3	0.12	0.88	0.95	1.00
7	54	3	0	1	0	6	4	1	3	0.11	0.89	0.94	1.00
8	35	2	1	0	0	4	3	3	0	0.11	0.89	1.00	1.00
9	43	1	1	0	0	3	2	1	1	0.07	0.93	0.98	1.00
Avg	51.71	1.71	0.86	0.71	0.57	7.29	3.14	2.00	1.14	0.14	0.86	0.97	0.82
Std	9.76	1.11	0.90	0.76	0.79	2.43	1.07	0.82	1.46	0.03	0.03	0.02	0.23

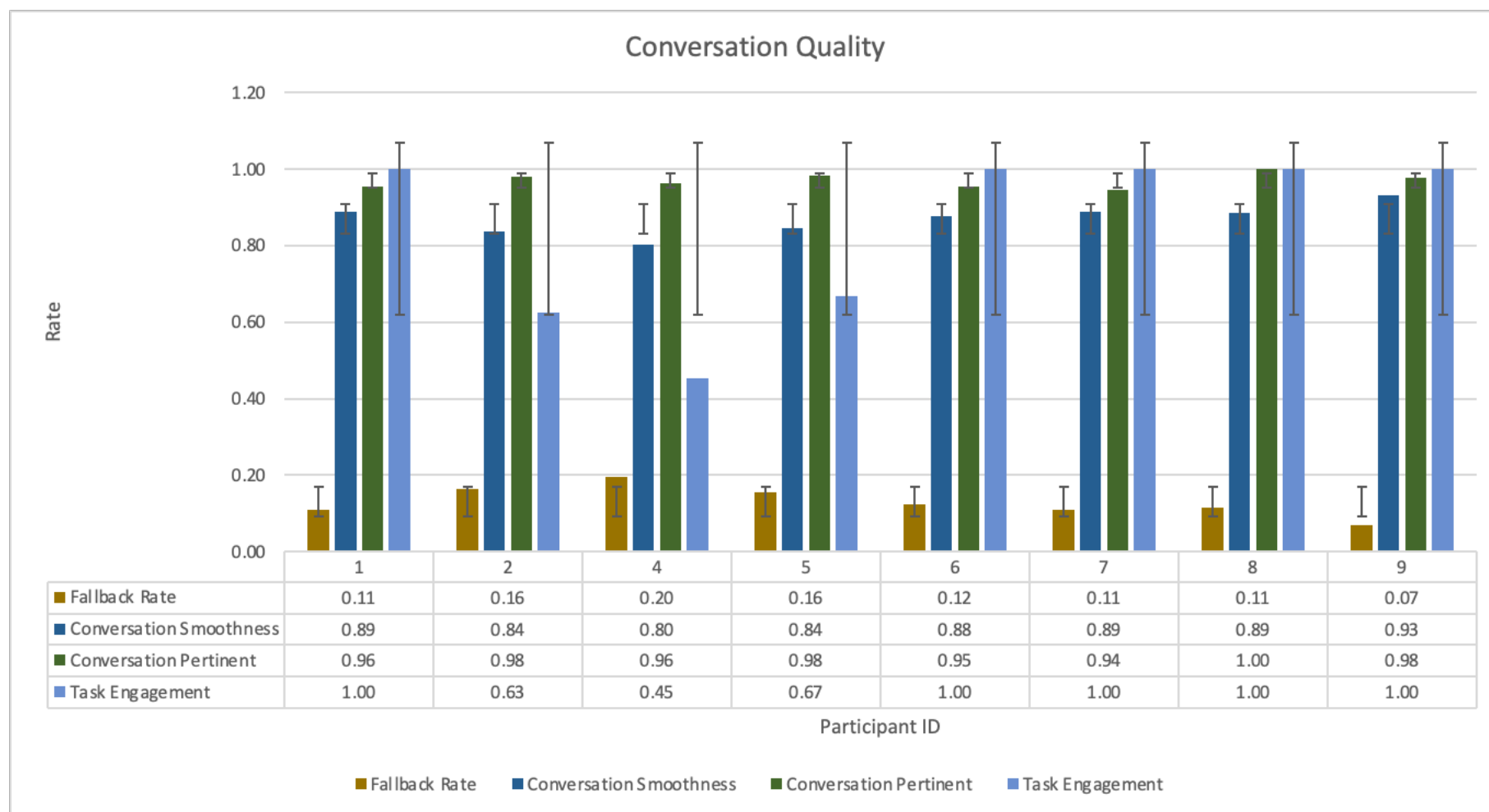


Figure 8-1: Overview of Conversation Quality

8.4 Qualitative Results

The thematic analysis of the interview data lead to the identification of five themes: verbalizing facilitates mindful reflections; Four types of structured prompts facilitate self-reflection; Self-reporting to a chatbot as a socially collaborative and empathetic experience; and the role of unstructured vs structured conversation to self-report. Finally, potential barriers that could impede the self-reporting experience were also identified

8.4.1 Verbalising facilitates mindful reflections

Participants' feedback in this study revealed more in-depth insights on how using voice to verbalize SARE can facilitate a more mindful and reflective process, which are further presented as eight subthemes as outlined in Table 8-4. Specifically, speaking aloud enabled further processing of the multitude of thoughts they experienced, which was perceived as an enjoyable experience. This led to an increased awareness and understanding of the internal self and the present moment. Furthermore, it was indicated that this generative thought process was beneficial in developing running strategies for regulating behaviours, as well as psychological strategies to motivate oneself when encountering obstacles and barriers. Experienced runners found that vocalizing their self-monitoring was similar to their existing habit of asking questions internally to reflect on while running.

Table 8-4 Verbalising facilitates mindful reflection: sub-themes overview

Overview of sub-themes
1. Vocalization for thoughts awareness and further processing
2. Vocalization for thoughts formulation to enhance clarity and completeness
3. Vocalization as a way of attention regulation and feeling externalization
4. Vocalization improves cognitive activity, including comparing feelings, pattern recognition, and sense-making
5. Vocalization as a means of self-determination and intrinsic motivation
6. Vocalization as a meditative and therapeutic intervention
7. Vocalization as an externalized version of experienced runners' self-monitoring strategy

8.4.1.1 Vocalization for Thoughts Awareness and Further Processing

Runners found that verbalizing their feelings while running facilitated not only the awareness of one's thoughts but also a further processing of the multitude of thoughts a runner experience, which was perceived as an enjoyable experience. As reflected by P7 in the first running session, *"When you are running, you obviously have loads of thoughts going through your mind, but I don't think you really process them when you're not thinking out loud. I suppose you do still think about how you're feeling, but I don't know whether you think that deeply. [...] Talking about it as you were doing it made you think about it more than how you were feeling and what the run felt like, [...] and then I processed about how I was thinking [...] It's hard to explain, but I quite enjoyed it. I could probably talk and run the whole way, expressing what I felt."*

8.4.1.2 Vocalization for Thoughts Formulation (enhanced clarity and completeness)

P1 argues that speaking out loud can help clarify thoughts, as it requires one to elaborate the thought in order to convert it into words, *"It makes you formulate your thoughts more clearly than if you just think them. When you think, you don't have to think about it much. But if you say them out loud, you have to finish the thought to convert it into words"* (P1).

8.4.1.3 Vocalization as a way of attentional regulation and feeling externalising

Participants became more aware, observant, and conscious when they thought aloud. This was evidenced by their reports of increased awareness of their physical and mental responses to their activity, as well as the need to be more present and appreciate the moment. For example, P2 noted they were more aware of how they were feeling, stating, *"When doing long distance and being tired, it's a lot of your body tells that you can't go any further any longer, so your mind shut up while you're moving. So, you don't particularly listen to it, you learn to block that. But the chatbot was quite interesting to sort of hop into actually thinking how I was feeling. I was more aware of how I was feeling, which I think sometimes you aren't. You just yeah, it made me it's like, I was aware that I had a bit of a pain in my Achilles, but until I verbalised it I wasn't taking very much notice of it."*

P8 said they had to be more observant and put words to what was happening, *“I had to consciously think about what was going on when I was thinking aloud, because I had to mention it, but I think that was a nice part of it, I had to be a bit more observant and yeah just trying to put words to what was happening. [...] I think I was present, most of the time during this run because I was consciously talking about everything that was going on. I was more conscious of what was going on than I normally am because otherwise I might be daydreaming or I would be thinking about something else. It's nice to be aware and be conscious and appreciate the moment.”*

Similarly, P3 reported they became more aware and conscious, stating, *“I started to think about. At a niggling my hip, so hip problem so that kind of coincided the same time with the first prompt came along, then I'm probably more aware of it, whereas usually if I'm talking to someone I tend to ignore it more but because I'm focused in that I might hit this play me up. [...] perhaps you just become a bit more aware of [...] how you're physically responding to your activity, and how you're mentally responding to your activity.”*

This process of introspection contributed to an increased awareness and understanding of the internal self and the present moment.

8.4.1.4 Thoughts Vocalisation improved cognitive activity: comparing feelings, pattern recognition and sense making

Participants felt encouraged to make sense of the causes for a perceived exertion and seek for a pattern in the changes in their feelings and experiences. For example, P3 commented, *“I started to think, 'Did that effort correspond to how I felt? Was I exerted? Or is it just in my mind that the hill was a lot harder? Did I get over the hill quicker than I thought?' It would be an interesting way to figure out how I actually performed physically against how I felt mentally. [...] Because I had said it out loud, I was able to compare the effort I was exerting now to what I had said previously. Obviously, the gradient had an impact on the effort required, whether it was a hill, flat or downhill”* (P3).

Other participants also indicated interests in extending such induced cognitive activity for making sense of their run. P2 commented, *“It would have been very interesting to find out the*

difference in feeling from the beginning of the runs to the end” (P2). Similarly, P8 also developed further interest in exploring “whether what I was feeling was also what was happening, such as going slower at the end. This would be interesting to see” (P8).

Such sense making would encourage participant's reflections and regulation of running behaviours. For instance, P1 reported feeling soreness in their lungs due to cold air and began to consider how to address this issue. They stated, *“The cold air was making my lungs a bit sore at times. I mentioned that a few times, and I probably wouldn't have thought about it as much if I wasn't forced to talk about it. Then I ended up changing my breathing to my nose instead of my mouth. Maybe that was because I was thinking about it more.”* When P1 started to comment on this body feelings, he also noticed [his] body trying to slow down, but then [him] made sure to stay the same speed. He further reflected, *“Normally, I don't remember thinking about it. You just let your body control what you do, and you might slow down. But when you actually think, 'Do I need to slow down now? Or am I just being lazy?', I realised that I didn't need to slow down and could continue running, which changed how I was running”* (P1).

Another example of such behaviour change can be observed in P2's case. P2 shared that ignoring pain is their normal response, but that talking about it helped to recognize and address it. *“When I go running on my own and think aloud, I become more aware of how I am feeling. This helps me to recognize the importance of attending to any pain or discomfort in my body, and to understand my emotions and feelings. Normally, if I am in pain, I will ignore the fact that I am hurting and carry on running, as this is what we often do.”* (P2)

8.4.1.5 Vocalising as self-determination and intrinsic motivation

This generative thought process can also be used to develop psychological strategies to motivate oneself when encountering obstacles and barriers while running. For instance, P5 reported that they think aloud about the wind, rain, and hills they encounter and become more aware of their fitness level, which helps to motivate them to push through. P5 stated, *“I've been commenting on the points that I find hard during the run, as a way of showing myself that I need to get a bit more fit. I did a lot of reflecting on when I ran up a hill and how that made me feel compared to when I'm running on a flat surface. I was saying 'I'm feeling*

out of breath', which made me realise how unfit I am. This gave me a bit more motivation internally. So I was like, 'Come on, it's only a tiny hill, you need to work a bit harder.' When I came up to the final stretch, I was saying things like, 'the wind is coming against me and the rain is hitting my face, and I'm going up a hill.' It was verbalizing the barriers that I was currently facing, but I was pushing through it. [...] It's more about seeking the motivation within myself and being more in tune with how I'm doing?"

8.4.1.6 Vocalisation as meditative and therapeutic interventions

The vocalised thinking process also involves a therapeutic process of reducing worries. In the think-alouds, P8 verbally described the sensations in his legs and expressed uncertainty regarding the correct posture to use, expressing concern that he may have adopted an incorrect posture. In the interview, P8 indicated that verbalising or thinking-aloud had helped to release the worry and reduce his anxiety: *"It was almost like I got rid of that thought by talking about it. Because it is a bit like something I'm unsure about. Though saying it out loud helps to just process it or get it out of my head."*

Vocalizing thought processes may engender an immersive, inward-looking, self-focused, and holistic mental state. As P5 stated, *"It is more about how one is feeling in the moment. I am really tuning into myself. It is quite centring, because you have to consider how the body feels, where it is aching, and which points have more pressure than others. You have to think about if I'm happy, sad, or content. It is almost like being mindful, thinking about how one is feeling in the moment and not being influenced by external influences. So, it is more reflective and inward-looking, and in that sense, there is a bit more mindfulness and wholesomeness to it."*

8.4.1.7 Vocalising as an externalised version of experienced runners' self-monitoring strategy

Experienced runners found that vocalizing their self-monitoring (both thinking aloud and chatting to a chatbot) was similar to their existing habit of asking questions internally to reflect on while running. They did not find it to be beneficial to their performance, but suggested that newer runners who do not have existing self-monitoring skills may benefit from vocalizing their SARE. P4 stated, *"As I have been running for a long time, nothing is going to change the way I run. It may not bring anything new to me, [...] but it could bring something*

new to a new runner who might say, “Oh, I hadn't considered that”. Similarly, P6 explained, “It's just like self-monitoring, isn't it? Today, I have to do it verbally. I'm always doing it in my head anyway when I'm running, as I did today with commentating.”

8.4.2 Four types of prompts facilitate self-reflection

Based on the participants' comments about both the R-ESM and the chatbot, it was revealed that there were four types of prompts that were useful for facilitating self-reporting and aiding runners in reflection during the reporting process. These included proactive but contextually aware prompts, active follow-up prompts, retrospective prompts, and mindful introspective prompts. Table 8-5 provides an overview of the subthemes.

Table 8-5 :Four types of prompts facilitate self-reflection: sub-themes overview

Overview of sub-themes
1. Proactive prompts as an ice breaker but need to be contextually aware
2. Follow up prompts/active questioning facilitates deeper understanding
3. Retrospective prompts help with in-the-moment self-reporting
4. Body scanning prompts facilitate more comprehensive self-exploration and self-reporting

8.4.2.1 Proactive prompts as an ice breaker but need to be contextually aware

Participants acknowledged the value of both systems in prompting them to think before they reported their experiences. They noted that the chatbot was more contextually aware than the R-ESM system, because the chatbot was designed (as detailed in Chapter 6) to verify whether a participant was speaking before conversing and asking questions, thus eliminating any potential interruption issues.

The prompts offered by the R-ESM can be an effective icebreaker to initiate and facilitate runners' verbalization process. Runners found it useful to be prompted especially during the early stages of a run when a runner *“was not thinking about what to say”* (P2). It was noted that the prompts *“gave [P1] ideas of what to talk about, which helped to initiate the process, and it enabled [P1] to become accustomed to talking”*. As participants became more

comfortable with and accustomed to the process, they began to enjoy the experience of self-reporting and verbalizing; as P7 stated, *“I think it was more pleasant having just been prompted, and then I just went with the flow”* (P7).

However, the pre-determined scheduling of R-ESM prompts may be disruptive if utilized in an unsuitable context. For instance, P1 noted the prompts were presented even though they had already begun speaking, *“Towards the end of the run, the last two thirds, I did not need any prompts as I was already talking. It wasn't very necessary when you had already gotten into the swing of it”*. The inappropriate prompts can interrupt runners' verbalizing and thinking aloud processes. Additionally, P1 stated, *“It was getting in the way. When it popped up, I would pause and wait for it to finish before continuing my sentence. I eventually began to ignore the prompt and simply continued talking”*. Similarly, P5 experienced a similar situation, noting that *“It was a weird feeling to have someone talking in my ear while I was talking out loud, even though I knew exactly what it was going to say. It was difficult to talk over it. I still had an instinct to stop, wait, and then talk, which could interfere at times”* (P5).

8.4.2.2 Follow up prompts/active questioning facilitates deeper understanding

The follow-up questions posed by the chatbot were designed to be increasingly specific, which prompted the participants to think more deeply and resulted in a more in-depth self-reports data. As P1 noted, *“There was a two-way dynamic because it asked me a specific question and I answered it, and then it delved deeper, and then I answer that again. For example, one thing I hadn't thought about until it asked me was 'How is your emotional state?' I replied with something like 'good,' but without thinking about why. Then it asked, 'Can you explain what makes you feel this good emotion?' Then I had to think even further about why it is that I feel good. That definitely made me think more than I would have done naturally.”*(P1).

8.4.2.3 Retrospective prompts help with in-the-moment self-reporting

As demonstrated in Chapter 6, the chatbot was designed to initially repeat information that runners had provided regarding their exertion, pain, or discomfort, and then inquire on how they felt currently. Participants found such retrospective prompts to be beneficial, as they

served to remind them of their prior feelings and enabled them to compare them to their current ones making it easier to assess themselves. Participants perceived not only patterns of change in their feelings over time but also felt more profound when evaluating the current self-reports profound as they were built on the previous one. For example, Participant 6 commented, *“It was useful because maybe you forget what your previous feeling was. This gives you a comparison point to think, ‘am I feeling more or less exerted than before?’ So that was useful.”* Retrospective prompts could help runners gain a more comprehensive understanding of the fluctuations in their emotions, thus potentially reducing the potential for bias and assumptions in their subjectively perceived exertion. As P1 noted, *“I noticed that my exertion level went down later in the run, which was interesting. I wouldn't have noticed that if it didn't remind me. This was surprising because you would think it (exertion level) would go up over time, but instead, it went up and then came down. That was definitely good to know.”* (P1)

8.4.2.4 Body scanning prompts facilitate more comprehensive self-exploration and self-reporting

The guided body scanning provided by the chatbot was found to be particularly effective in helping runners to explore and investigate their feelings and bodily sensations in a more comprehensive and holistic manner. This enabled them to become more aware of their bodily sensations and to identify and articulate multiple pain feelings, thus bridging the gap in their thinking aloud. For instance, P5's attention focus was initially dominated by the pain in her knees. After following the body scanning exercise, she identified some discomfort and sensations on different body parts, which also enabled her to verbalise multiple pain feelings. Participant 5 stated, *“The body scan was more helpful. It's something more tangible, you can walk away being like, today, I felt more tension in my shoulders and in my legs, whereas yesterday (when thinking aloud), I was just blabbing on about my legs. But I didn't talk about any other body parts, because I wasn't in tune.”* (P5)

8.4.3 Self-reporting to a chatbot as a socially collaborative and empathetic experience

Chatbot was perceived as a conversational partner that provided a collaborative self-reporting experience. The ability of understanding the content of runners' speech input and

responding accordingly created a sense of bonding and empathy between runners and the chatbot. Runners felt being understood by the chatbot compared to the interacting with the R-ESM system. However, the limitation in the chatbot’ s natural language understanding capability also triggered anxiety to runners. Table 8-6 provides an overview of the subthemes.

Table 8-6: Self-reporting to a chatbot as a socially collaborative and empathetic experience: sub-themes overview

Overview of sub-themes
1. Chatbot as an interactive aware collaborator
2. Empathy is good but it should be at the appropriate level
3. Small talks rather than just self-report

8.4.3.1 Chatbot as an interactive aware collaborator

In comparison to the simple reporting process with the R-ESM, participants experienced a stronger sense of socially collaborative self-reporting when interacting with the chatbot. As one participant remarked, *“(when thinking aloud to the R-ESM system), you have to have the skills and feel comfortable with asking yourself those questions and doing reflective work, whereas the second one with the chatbot is more like a team thing. It's kind of guiding you toward reflective thinking while you're running”* (P5). Furthermore, P6 highlighted the interactive nature of the conversation, noting that the system was able to remember and learn from their responses, asking relevant questions based on their previous answers. They commented, *“It was interesting to see how it is interactive. It's not like you're just talking into a microphone, you're talking to something that is actually listening and responding to you. I don't know how it all works, obviously, but it must have a memory and it understands what I'm saying, because it asked me a question by quoting my previous answer. So, it's obviously the AI listening to you and remembering your answers, learning from my previous responses and then using that information straight away for its next question. That was quite interesting when I heard that.”*

8.4.3.2 Empathy is good but it should be at the appropriate level

Overall, the runners found the chatbot to be a helpful and reassuring companion during their runs, providing understanding and praise. For instance, P7 commented, *“I appreciated the chatbot's praise; it said something like, ‘I'm glad you're enjoying your run. You're not exerting yourself too much’”* (P7). P1 found the responses from the chatbot to be reassuring and comforting, stating, *“When I mentioned feeling low in motivation, it responded with ‘Don't worry, everyone feels that’. This was reassuring, as it demonstrated that the chatbot understood my situation. It was nice to receive this kind of feedback”* (P1).

It was observed that the social bond created between the runner and the chatbot through their collaborative experience. For instance, Participant 2 remarked that they felt lonely when the chatbot prototype ceased functioning for an unknown reason, saying, *“I was just getting used to talking to it and it was quite nice. But then it stopped and I had to run back on my own. I felt so lonely.”* (P2)

Runners were pleased to receive an empathetic response from an AI chatbot, but also found it to be inappropriate in certain circumstances. For instance, P1 found that the chatbot's response was overly dramatic when they expressed feeling *“a bit low”*. P1 stated, *“I said I was feeling a bit low at the moment. It reacted a little bit too extreme, as if I said I was very depressed or something. It was definitely nice to hear it, but it would be better if it answered in an appropriate way. I guess it doesn't get the full context that a human would get.”* (P1) It was suggested that AI chatbots should be programmed to respond in an appropriate manner in order to better reflect the context of a conversation.

8.4.3.3 Small talks rather than just self-report

Nevertheless, participants expressed a desire for the conversation to be more like a human-human chat, suggesting that the chatbot should engage in small talk and ask more personalised and humanised questions. Suggested topics included goals for the week, how their day had been, how they were feeling, and the weather. One participant said, *“I think it needs to be more personalised and humanised, asking me more questions and letting me talk. Maybe halfway through, it could ask things like, ‘What are your goals this week?’ ‘Why is this*

a training run?' 'How has your day been?' 'How are you feeling?' 'What's the weather like?' because it all contributes to how you're feeling, doesn't it? Yeah, all the things that are going on around.” (P7).

8.4.4 The role of unstructured vs structured conversation during self-reporting

Runners perceived a sense of structure when conversing with the chatbot and a sense of lack of it when engaging in the thinking aloud with the R-ESM system. The pros and cons of each of these aspects were revealed. Table 8-7 provides an overview of the subthemes.

Table 8-7: The role of unstructured vs structured conversation during self-reporting: sub-themes overview

Overview of sub-themes
1. Space for Thought Processing and Development, and Verbalization
2. Affective experience: unstructured free talking gives a sense of autonomy and freedom whereas the structured makes it restrictive
3. Focus facilitates reasoning, comparison and memorable report

8.4.4.1 Space for Thought Processing and Development, and Verbalization

Unstructured Think Aloud showed the advantage of offering the runner more time for thought processing, such as more thorough investigation and understanding of one thought, as well as more space for openly elaborating, such as more space for thoughts to go either as wide or deep as possible. P1 stated, *“In the think out loud, you can just let your thoughts go wild and take it as deep as you want. [...] You may be getting a deeper understanding of whatever you happen to be thinking about all the time. You probably think about it more deeply, because you get to process your thoughts all the way through.”*

Structured Chatbot Conversation conversely were seen as not providing sufficient time for preparing/formulating thoughts with expected answers that were restrictive and restricted to concrete outputs, thus lacking space to physically elaborate more or go deeper in self-understanding when reporting.

When conversing with the chatbot, runners experienced a lack of space to express/elaborate themselves, feeling as though they were being questioned and required to provide answers. P7 reported, *"I thought it was going to ask me another question, so I didn't elaborate; I just answered the question."* Participants reported a feeling of restriction due to the format of expected answers, as P1 stated, *"Chatbot said, 'Can you use a word to sum up your emotional state?' I had to summarize my emotions in one word, which is quite hard to do. That isn't an easy thing to think about. Yeah. You're sort of being a dictionary. [...] It is quite limited in terms of how deep you go. [...] It would be nice that either not restricted to few words or maybe to keep the conversation going a bit longer."* (P1).

Runners explicitly expressed their desire for more space to **talk and describe** beyond simpler giving short answers to the chatbot. As P7 said, *"Despite I had the time to think about what was going on, I didn't think I had the time to describe what was going on. The chatbot was questioning you, so you're just giving the answer. [...] I like that I was doing more of the talking last week, when you were left to describe."*

Describing one's emotional state is a slow and generative process that requires time for reflection, retrospection, and processing of subjective states before a definitive answer can be given. However, during the interactions with the chatbot, participants felt challenged and pressured to provide an answer quickly, as if they were being limited in the amount of time and space to think aloud openly and freely. P1 explained, *"Sometimes I haven't thought about it, and I need to dig into why I'm feeling good or something, and it takes longer than a few seconds to figure that out before giving answers. [...] but I don't know how long it's going to listen for. I felt like I had to give an answer quickly. Otherwise, it might... so I'm trying to rush to come up with an answer, but it's hard [...] and a bit probably stressful. Things like this, you don't have to account for in human conversation where you can take your time explaining something."* Such need to rush may have led to possibly untrue self-report. P1 desired more time and space to think aloud freely, which would help them to finalize their thoughts and answers. They noted, *"because maybe you haven't thought about it fully at that point in time, it's hard to get it fully processed it in your head before saying anything [...] if you can think aloud like before (in session 1) and then explain, then you don't need to do that processing, you can just let it spill out without needing to think about it in advance."*

8.4.4.2 Affective experience: unstructured free talking gives a sense of autonomy and freedom whereas the structured makes it restrictive.

Participants perceived the unstructured thinking aloud approach in running session 1 as a relaxing and liberating experience. P7 expressed a sentiment, stating, *“It was quite a spoke, quite liberating to talk about how I felt whilst I was actually doing my exercise. I felt relaxed because I could do the talking and just describe how I felt.”* This relaxing and liberating experience of expressing oneself made P7 feel as if she was engaging in a genuine conversation with a close person, saying, *“Last week, in my mind, it was like I was having a conversation with a friend or I was having a conversation with you. It felt like I wanted to say a lot more about what was going on.”* In contrast, with the chatbot that lacked space for free expression, P7 sentimentally felt that her feelings and words were not being wanted or appreciated, saying, *“This week, it just felt like it just wanted a short answer. Because it went on to the next question, it did understand the answer that I’ve given. But was it interested in my answer? I don’t know. It didn’t feel like it wanted to know my thoughts.”*

Besides, participants indicate their desire for more autonomy and proactive control in order to be more actively engaged in conversations, rather than allowing the chatbot to dictate when they could speak (in spite of the benefits of the chatbot's proactive interaction as reported in Section 8.4.2.1). P5 commented, *“Reflecting on yesterday's session, it was great that I was just chatting away the whole time. However, in the second session, I felt like it was more based on the times that the bot talked to me. Perhaps it would have been better if I had been able to talk more in the second session.”*

8.4.4.3 Focus facilitates reasoning, comparison and memorable report

Despite the drawback highlighted above, participants found that the structured Chatbot Conversation has the advantage of leading to more focused self-reports with more precise, specific and tangible information to report. In contrast, when engaging in unstructured think aloud, participants tended to generate too much unfocused and random information that they found at time difficult to remember and reflect upon either during or after the run. For example, P1 reported that his memory of their think-alouds information was *“less vivid”* than that of the chatbot conversation, and felt *“it would be highly challenging to utilize the think-*

alouds data for anything useful.” P7 also commented, “I was just throwing everything out there, about the tarmac over there, dog walkers, etc.” When participants were asked to reflect upon their think-aloud verbalisations during post-run interviews, they found it difficult to remember their utterances; P7 stated, “I completely forgot what I talked about, because I was talking so much”, and P5 remarked, “I was speaking a lot, I can't recall too much apart from a few words”.

Conversely, a structured chatbot conversation was associated with a higher memorability and easier interpretability that runners’ consider useful for the self-report task but also for themselves. They felt that with the structured conversation there was a more balanced attention to different aspects of their experience. Participant P5 explained, *“I was talking about everything in kind of similar or equal quantities”*. In addition, the structured chatbot conversation ensure the self-reports were specific, as said by P1, *“it gets you to think about specific things that you wouldn't think about on your own”*. The chatbot provided more precise, succinct, and tangible data, which users found easier to remember. P1 reported that his memory of the think-alouds information was *“less vivid”* than that of the chatbot conversation, and felt *“it would be highly challenging to utilize the think-alouds data for anything useful.”*

The structured and focused questions of the chatbot created opportunities for better reflection, pattern recognition, and semantic understanding of self, compared to unstructured thinking which produced random information that was harder to process and make sense of. As P1 noted, *“I would like the structured one because I would like to see the structured data, and I can get trends over time from this structured data, which I don't think you would get from the think out loud version. I don't remember noticing that as much yesterday, probably because you don't get directed to. Like, it's not asking you, what's your emotional state right now? You can say anything you want. So, you don't think about these specific questions. That is something I would definitely like for myself.”* Similarly, P5 explained, *“Interacting with the chatbot probably gave me more concrete and useful data, as it asked me to provide specific answers, explanations, or definite ratings. When talking freely about my feelings, I did not do that. Although it is a good process, it does not provide anything tangible to take away. I prefer information to be concise so I can retain it in my mind. The*

second run (with the chatbot) was more structured, with less information, but still as informative. So, it was probably easier for me to remember. The information I got from the second run is more applicable for me to reflect upon and consider for the following run. For instance, I would be wanting to reflect back and think, 'why was it different yesterday to today?'"

8.4.5 Factors or Barriers for self-reporting

During the interviews, we observed several barriers that could affect runners' self-reporting experience. These included variations in verbal and dialogical behaviours, contrasting attitudes towards self-consciousness, in public expression of feelings, self-attention to one's feeling, individuals' differences in ability to reflect, different running capabilities and goals, and the need for cognitive attention to external context (e.g. traffic, road conditions for safety). Table 8-8 provides an overview of subthemes.

Table 8-8: Factors or Barriers to self-reporting: sub-themes overview

Overview of sub-themes
1. Variations in Verbal and Dialogical Behaviours
2. Contrasting Attitudes Towards Emotions
3. Self-consciousness in Public Expression of Feelings
4. Need for Cognitive Attention to External Context (e.g. Traffic, Road Conditions for Safety)

8.4.5.1 Variations in Verbal and Dialogical Behaviours

The experience of runners vocalizing SARE can be linked to their habits of vocalizing their feelings and thoughts in everyday life. Some individuals find vocalizing to be effortless, as evidenced by P2's comment that they *"tend to talk a lot anyway"* and P7's remark that *"I find it quite easy and enjoyed talking to myself, and I'm renowned for it at work"*. On the other hand, some people find it difficult to verbalize their thoughts. As P6 noted that *"Thinking in my head is just a completely natural process,"* but it was *"very unnatural"* to try and verbalize what was going on in their head, saying, *"thinking out aloud is not something I can do very easily"*.

There is also evidence of individual differences in their dialogical preferences for open-ended or close-ended questions. For instance, P6 expressed a preference for close-ended questions, stating that they *“prefer almost closed questioning, what is this, and what is that, you know, and [they] can give an answer”* as opposed to expressing feelings or beliefs because they *“don't normally think deeply”*. Similarly, P4 expressed difficulty in figuring out what was expected to say when presented with open-ended prompts in the first running session where they were thinking aloud. In contrast, P7 expressed a preference for open-ended questions, noting that they *“like to verbalise things”* and that they didn't not like *“just saying 'yes' and 'no' ”* when presented with close-ended questions.

8.4.5.2 *Contrasting Attitudes Towards Emotions*

Individuals hold varied attitudes towards **attending** to their feelings. As P3 remarked, *“I generally do not reflect on the emotional aspect of running; I just think that it was a good or difficult run, but never about the emotion I was feeling - just about being hungry or tired, or wanting to walk.”* One participant shared that they had developed a tendency to disconnect from emotions and subjective feelings, stating *“I would say I'm somewhat autistic. I tend to be very logical and practical in many situations, rarely thinking about emotions. [...] As a police officer [...], I have the ability to detach myself from unpleasantness and continue with the task at hand - something I have learned to do throughout my life.”* (P6).

Individuals exhibit a range of attitudes towards the **disclosure** of emotions. P5 and P7 expressed comfort in doing so, with P5 noting, *“I'm quite open. I do talking therapy, and I've vocalized my thoughts and feelings a lot, so I don't find it intrusive or uncomfortable. I think for others, talking about how you feel, how you describe your body, how you describe your pain, and so on.. It's quite deep and personal.”* On the other hand, P4 stated, *“Do I need to talk about my emotions while running? No. No, no. (My brother does). I'm not as miserable as my brother because he has a degree in it [meaning his brother is miserable]”* (P4). P8, who was uncomfortable discussing their anxious feelings, commented, *“Sometimes I would filter what I would say. I was aware that you were going to listen to it later. [...] It felt like I was talking to myself, but also knowing that someone else was going to hear it. Even in my room if I'm by myself, I find it awkward to do something like that, so, yeah, I have this feeling, as if somebody could be listening or something”* (P8). These results indicate that personal

experiences have a considerable influence on the way individuals perceive and respond to the disclosure of emotions. At the same time, privacy is also an important issue.

8.4.5.3 Self-consciousness in Public Expression of Feelings

Participants reported a sense of self-consciousness and heightened anxiousness that led to potential restrictions on self-reporting while running. P7 stated, *"If there weren't people around, I probably would have talked more, but I stopped when I saw people."* This was associated with feelings of embarrassment, which interrupted their self-reporting behaviours. Another participant noted, *"I noticed I would stop talking as I was running past people, as it felt a bit embarrassing to share my private thoughts out loud"* (P1). Another participant shared a similar experience, stating, *"It was a bit strange to me, I was feeling a bit anxious doing it around other people. I was worried that somebody would hear me. When there were people around, I went a bit restricted, either wanting to find another path to avoid them or just make sure that I was quiet around them and didn't say anything. I was also worried that there would be someone nearby that I just didn't see, like maybe somebody was behind me or something"* (P8).

Although many people reported feeling self-conscious while verbalizing SARE in public, this was found to be possibly individual; one participant (P4) reported that they did not experience such self-consciousness or worry. When asked how they found thinking aloud their feelings in public, P4 stated, *"That didn't worry me at all. I don't actually think there was anyone around that could have overheard me anyway, and even if someone had been, that would not worry me either. That's just the way it was."*

A strategy used by participants to counterbalance the perceived embarrassments was to treat self-reporting as a contribution to research. For instance, P5 reminded themselves of the research purpose after experiencing such embarrassments, saying: *"I thought, 'no, it's for science. It's an experiment, and this is an experience so just don't let that distract me. Then, I just carried on."* (P5). Similarly, P6 expressed willingness to contribute to the research: *"if I was given the choice to not do it or do it, I would prefer probably not to do it. [...] But it's the research, I think about what I'm doing to benefit the research."* (P6).

8.4.5.4 Individuals' differences in reflectivity

Self-reporting of experiences can be impacted by an individual's capacity for self-reflection and introspection. Some participants are naturally adept at reflective thinking and analysing their feelings, exemplified by the statement: *"I'm quite a reflective individual. And I find that thinking and analysing how I'm feeling in the moment can be quite good. I have the ability to go deep quite quickly, where I would talk about feeling, like there's lots of things in my mind"* (P5). Self-reflection necessitates skills and practice; however, others may find difficulty engaging in full self-introspection, as expressed by the participant who remarked, *"I'm not very good at reflecting and doing things like this"* (P6). This difficulty may be due to a lack of descriptive words, as one participant noted, *"I kind of was struggling to find different descriptive words. So I felt like I was repeating myself a little bit. [...] kind of not finding enough words to kind of describe how you're feeling in the moment"* (P5). Similarly, when thinking aloud in session 1, P8 experienced cognitive challenge, as stated, *"It's very hard to reflect everything that goes on, because I can only mention one thing at a time. There might be several emotions going on pretty quickly. It probably didn't go into the perfect reflection about the emotions."* (P8)

8.4.5.5 Different Running Capabilities and Goals

Experienced runners may not need reflective thinking as they already possess adequate skills; however, reflective thinking may be beneficial for new runners. As P4_S2 stated, they have been confident in their running and suggested no need for reflection, noting, *"In a normal run, do I think 'Am I doing my arms correctly?' 'Am I thinking about parts of my body?' To be honest, no, I'm not. I have been running too long."*

Similarly, for P6, who is a very experienced runner with a mindset focused on training and achieving a desired level (P6). He claimed to have such habits of self-checking and self-questioning while running. P6 remarked, *"The questions asked by the Chatbot, such as 'How are you feeling?', 'Do you have any injuries?', and 'Is anything bothering you?' I'm talking about it today, but I'm just doing what I would have done anyway. It doesn't make any difference to my run today. It didn't make my performance any better because the Chatbot is asking me the questions that I might be asking myself anyway. [...] So, yes, it was just like*

putting that into a third party rather than my own head.” This appeared to be a result of their pre-existing practice of posing questions to themselves while running. As P6 stated, “I’m monitoring myself and how I feel, my performance, my energy levels, and my feelings. When I am running, I am also thinking about them particularly if I am recovering from an injury. I am asking myself how I am feeling if I do this or that, and how my fatigue is changing. I usually train alone, so I am more concentrated on what I am doing and these questions the chatbot was asking.”

P6 further suggested thinking about SARE could be a useful tool for inexperienced runners, enabling them to *“think about what they’re feeling and what effort they’re putting in and why they’re doing it”* (P6).

8.4.5.6 Need for Cognitive Attention to External Contexts

Despite being instructed to run in a confined area for ethical and safety reasons, some participants chose to cross a road in order to access another section of the park.

When running, participants had to pay attention to their external environment to observe traffic and road terrains, making it difficult for them to focus internally and discuss their subjective experience of the body and for self-reporting. P1 noted that *“it was hard to concentrate on it (chatbot), because I was running and you’ve got to look around. And maybe if you’re running next to a road, you’ve got to be careful if there’s a car come in. So it’s hard to be mindful on the body scanning while you’re running.”* Similarly, P4 noted that *“I’m quite conscious about traffic around, [...] On certain routes, you need to be conscious of surface. If I know there’s a hill coming up, I know I’ve got to concentrate.”*

Participants also raised concerns in they need to use the chatbot in places where they need to reserve attentional efforts in order to respond in a timely manner to the ever-changing environment. P3 noted that *“If you’re crossing the road and something’s talking in the background, you’ve got to be able to zone out to that rather than respond to the question, you’ve got to make sure you cross the road safely.”* P5 also noted that *“The chatbot obviously doesn’t know that you’re about to negotiate a roundabout or a group of people on the*

pavement, so it takes you away from being able to react instantly to what is going on with you.”

Another factor that impedes runners’ self-reporting (in either running session) is related to the use of headphones while running. Some runners may not feel comfortable wearing headphones in their ears, perceiving them to be burdensome, or unsafe. This can hinder the self-reporting process, as the conversational interaction with the chatbot requires the use of headphones. P2 stated, *“I don't generally run with headphones. I tend to run in nature”* (P2). Participant 4 expressed, *“I'm not into headphones at all. So as this one off that's fine but I actually think they're dangerous. Everyone knows you're not allowed to wear headphones in races anyway”* (P4). Participant 6 also noted, *“I'm not used to running with earpieces in my ears”* (P6).

8.5 Discussion

The research question being addressed in this chapter is ***RQ 3: How can chatbot conversation flow be designed to facilitate voice-based SARE self-reports while running?*** According to both quantitative results from the analysis of conversation logs (such as fallback rate, conversation smoothness, conversation pertinent, and task engagement) and qualitative feedback from interviews (e.g., the chatbot has facilitated self-reflection during running), the chatbot prototype designed for voice-based SARE self-reports has been found to provide runners with a reliable user experience.

The analysis revealed that vocalization or verbalization is a useful modality for self-reporting, and how different degrees of autonomy in the conversation and types of prompting can enhance the SARE self-reporting. These not only benefited the runners themselves, but also contributed to the richness of self-reported SARE data. Such findings suggested the following framework, Modality-Autonomy-Prompting (hereafter called MAP framework), which could provide a guide to the design of the chatbot to facilitate voice-based SARE self-reports while running.

Table 8-9: The Modality-Autonomy-Prompting (MAP) Framework: a guidance to design a chatbot to facilitate voice-based SARE self-reports while running

Dimension	Conversation Structures	For Runners (Experiences)	For Chatbots (Design Implications)	For Self-reports Data
Modality	Voice Interaction	speaking as reflective thinking process;	Adapting to the iterative nature of speaking behaviours	Extra contextual sound information
Autonomy Levels	Explorative Speaking via Focused Thinking Aloud to be Used Before Structured Self-Report	Facilitate free self-exploration, deep elaboration, and liberate reporting experience without language and time constraints	Identifying thought formation stage and completion stage; provide topic for follow-up conversation and context by identifying key concept (using GPT-3)	Enabled self-understanding for more reliable self-report responses; self-reports of reasoning process;
	Directed Conversation with a Chatbot	Facilitate concrete outcomes that are more memorable for easier pattern identification and reference	Easier to process and to develop context to provide sense of presence and companionship	Balanced, specific, concise, focused, and structured self-reports
Prompting	Context-Aware Proactive Prompting	Ice breaking and inspiration to start self-reporting	More intelligent context awareness	Increased self-reporting compliance
	Semi-Structured Follow-Up Prompting	Deeper and contextually relevant reflection and reporting	Adhere to the essential follow-up questions but with more dynamic utterances and optional follow-up questions	Dynamic but still focused self-reports
	Chatbot's Retrospective Prompting	Facilitate recall and sense-making for easier and more coherent self-reporting experience	Support user-friendly revisiting and information retrieval to assist self-reporting	Enhanced accuracy and consistency of self-reports
	Systematic and Mindful Scanning Prompts	Facilitate broader attention and vision with an increase in holistic self-reporting	Support holistic systematic inspection to guide self-reflection	Unbiased and holistic self-report with increased variety

Next, this section follows the MAP framework to discuss in detail the findings and how the chatbot could be designed to facilitate voice-based SARE self-reports while running.

8.5.1 Implication for leveraging vocalisation as a useful interaction modality for encouraging self-reflection and enhancing metacognition

Runner Experience: The findings of theme 8.4.1 suggested that vocalization of SARE for self-reporting can enhance self-reflection and metacognition. Participants reported that vocalizing their thoughts, either by self-reporting to the R-ESM or using the chatbot, improved their thought processing and formulation, as well as idea generation. Furthermore, externalizing their feelings helped regulate their attention and improve cognitive activity, such as pattern recognition and sense-making of their emotions. These findings are supported by existing literature in psychology, which suggests that one's inner thoughts are often highly erratic and expressed using discrete words or condensed sentences [269]. Speaking aloud can slow down the thought process, enabling a more manageable way of retrieving thoughts, complete with rhythm and intonation, which aids in restructuring messages and transforming obscure ideas into more meaningful concepts, ultimately leading to a better understanding and emphasizing pragmatic and argumentative meaning [24].

This thought formation process during speaking also means that reporting SARE is not a process of rapid thinking and responding, but rather a slow, exploratory and iterative process that requires time for runners to introspect and elaborate on their feelings. Feelings are not fixed or stable, but rather dynamic and compounding, evolving and changing over time and in different contexts. Consequently, when asked to describe this SARE, runners may not be able to offer a clear answer or a well-defined self-report initially. They might pause talking to think, or they might begin to articulate something and later revise, repeat, alter, or expand on what has been said.

Chatbot Design: Such findings in this chapter extend the existing literature on meditative thought processes and reflection while interacting with a GUI-based chatbot [160]. Recent research by [216] has demonstrated that answering questions posed by a chatbot during stock investigation can induce a slower thinking process and enhance metacognition for better behaviours awareness and decision-making. This chapter suggests that, in addition to

being prompted with reflective questions by the chatbot, vocalization itself is an essential interaction modality that can help users achieve enhanced awareness and thinking processes. This could inform the value of using voice-based chatbots for conversational interaction, particularly when the goal is related to self-reflection, such as in research about mental therapy.

Self-reports: Lastly, the vocalization process could also provide contextual sound information about how a person's thoughts were formed and their reasoning process. An analysis (outside the scope of this research) of such vocalizations could provide further understanding of how a runner felt.

8.5.2 Implication for chatbot with a mixture of both structured and unstructured space

The findings of this chapter suggest design implications for using a mixture of structured and unstructured conversation. We first discuss how this mixed-structured conversation structure differs from the existing literature, and then discuss how it could contribute to the runners' experiences as well as the gathering of self-reports data.

The proposed mixed-structure chatbot conversation flow designed to collect self-reports differs conceptually from the current literature on structured and unstructured chatbot conversations (see Table 8-10). As elaborated in Chapter 2.3's literature review, the traditional understanding of structured chatbot conversations primarily relies on closed-ended questionnaires with pre-determined response options (e.g., Likert-scale type of questions, binary questions (yes/no), and words (e.g., agree/disagree), or multi-optional questions with predetermined answers [70, 158-160, 272].), while unstructured chatbot conversations rely on open-ended questions [272] or [70]. In contrast, the structured conversation presented in this thesis incorporates both open-ended and close-ended questions to guide and prompt the user through a focused and systematic follow-up process, whereas the unstructured conversation gives users space to encourage users to express themselves around a topic in a more natural, exploratory, iterative and generative manner.

Table 8-10: Conceptual Extension to Chatbot’s structured & unstructured conversation

	Current literature	This Thesis
Structured Chatbot Conversation	Conversation using close-ended questions, such as Likert-scale, binary (yes/no) and words (e.g., agree/disagree) or multi-optional questions with predetermined answers [70, 158-160, 272]	Directed, guided & prompted conversation with the chatbot through a focused and systematic follow-up process
Unstructured Chatbot Conversation	Conversation using open-ended questions [272] or [70]	Unstructured conversation space that encourages users to express themselves around a topic in a more natural, exploratory, iterative and generative manner

Next, we explore how the aforementioned conversation structures can enhance runners' experiences and their implications for chatbot design as well as how they can facilitate the collection of self-report data from the runners.

Runner Experience: In our findings, participants perceived the explorative speaking via thinking aloud with R-ESM as unstructured, and the directed conversation with the chatbot as structured. While both structures have their own advantages and disadvantages, unstructured thinking aloud provides more time for thought processing and free elaboration, allowing for a thorough exploration of a single thought. The lack of control in conversations gives users more autonomy and comfort, but the absence of structure may hinder users from meta-analysing all verbalized thoughts.

The finding can be supported by the psychological knowledge about episodic and semantic memories [174, 260]. As shown earlier in findings, recalling and making sense of all the self-reported information discussed during the unstructured thinking aloud can be challenging. Runners noted that when speaking in an unstructured format, their memories of the running experience and self-reports were more akin to episodic memories [260], focused on a single

point of the event instead of the overall connected experience. On the other hand, the structured conversation with the chatbot facilitated the generation of focused and concise self-reports from users, enabling users to create a semantic memory of their verbalizations and overall running experience that was more memorable and meaningful. As a result of the structure, users were able to engage in a more effective self-reflection process, including the ability to recall how they felt during a previous run, identify the reasons behind their reactions, and learn how to react in the present.

Participant in this study indicated that predetermined conversation structures may make users feel constrained, leading to a sense of being dictated with no freedom to initiate or proactively talk. Some recent works on using voice-based chatbot for collecting users' survey response also reported restricted and pressured experience of participants and their desire for a space for a slow thinking and speaking. For instance, participants in [272] experienced pressure to give an answer quickly while using voice to self-report, as they were only given around 8 seconds to speak. This is distinct from the pressure reported by participants in this study, which was caused by the chatbot continuously prompting or pushing for an answer if they had not started, or by the chatbot's requirement for very structured answers that do not represent the complexity of how they felt. Similarly, users in [158] felt pressure to respond quickly enough to the chatbot, causing them to prepare their answer in advance in order to respond more smoothly. Additionally, one of their participants reported strategies to trick the chatbot to give them time for a slow thinking process, e.g., *"I tried different tricks, like, for instance, I tried to make some sound like 'hmmm leet meee thiiink', but it seems like my process was too slow for Sofia. So, at one point, I started giving very short answers so I won't be cut off yet could answer. (page 14)"*. These recent works confirm the importance of having space for slowed down self-reporting. But the findings in this chapter further explain why and argue it is essential to design a voice-based chatbot to support a slow, exploratory and iterative process of self-reporting during running rather than simply requesting quick final answers from runners.

Chatbot Design: The findings in this chapter suggest that the chatbot should allow runners sufficient time to think and accommodate the slow and explorative thought formation process. To address this issue, the chatbot could be designed with the following features.

Firstly, the chatbot should not assume that the user already knows what they want to say. It should recognize that users need time to think and avoid prompting or pushing them for answers. This would give runners the necessary space and freedom to explore their thoughts at their own pace. Secondly, the chatbot should be able to distinguish whether the user is in the process of slow thought-formation or providing reports of a well-formed final answer. If it is the former, the chatbot should wait until runners have finished their thought formation process before responding or transitioning the conversation to the next turn. This would help prevent interruptions and allow runners to complete their thought process. Furthermore, as observed in this study, runners engaging in slow thought formation may mumble or speak in a way that helps with internal thinking but is not easily understood by others. In such cases, the chatbot may not recognize such mumbled phrases, leading to conversational fallbacks that may restrict runners from providing rich self-reports and cause conversation interruptions.

The positive feedback on the structured chatbot conversations suggested the importance for chatbot to follow a structured conversation to further assist runners in finalizing and organizing their self-reports for a more structured experience. Therefore, we propose a combined structure for chatbot design that first employs unstructured conversations to support the slow, exploratory, and iterative process of self-reporting for more structure and autonomy (similar to thinking aloud with the R-ESM), and then provides structured guidance for more focused self-reporting through structured conversations. Nonetheless, it is important to consider individual differences in personality and preferences when designing exploratory and generative speech interactions. This is demonstrated by some runners' preference for short, concrete speech over lengthy open elaborations.

During the exploratory process, it would be beneficial for the chatbot to provide a certain level of responses to show empathy and facilitate the conversation by recognizing the topics that the runner is discussing. The study [159] that used chatbots to collect survey responses revealed that users often feel that their open-ended responses are not understood by chatbots because they do not respond specifically to their input. To address this issue, chatbots could be responsive to users' open-ended think-alouds by asking relevant questions, prompting for reflections, and providing detailed and sensible responses based on the user's

reported experiences for further continuous elicitation of reflections [158, 160]. However, the chatbots designed for this thesis are still limited in their ability to understand and support natural and exploratory SARE self-reporting. To enhance the chatbot's capabilities, future research should analyse the topics that runners discuss and how they describe their SARE experiences. This understanding can be used to design the chatbot to support both exploratory SARE self-reporting and guide the conversation by suggesting relevant questions in the structured format.

Self-reports: The chatbot's support of an iterative verbalization process not only facilitated the generation of final consolidated self-reports, but also provided valuable self-reports data users' reasoning processes. Our findings offer a compelling use case for recent research in speech and cognitive literature, which highlights the potential of heuristic think-while-speak processes for speech data analysis [24]. Overall, such a mixed-structure chatbot would have the potential to output rich, exploratory self-reports that have richness and variety, facilitate more profound and well-reasoned self-report data, and contribute to a more organized self-report dataset. This can result in more legitimate and justified self-reports of subjective experiences.

8.5.3 Implication for designing chatbot with four types prompting for self-reporting

The study's results indicate advantages of the various prompting structures that can be offered by the chatbot, which includes context-aware proactive prompting, semi-structured follow-up prompting, retrospective prompting, and systematic and mindful scanning prompts.

8.5.3.1 Context-aware proactive prompting

The chatbot should be designed with **context-aware proactive prompting**.

Runner Experience: As shown in 8.4.2.1, runners responded positively to proactive prompting, finding it motivating and inspiring. However, they also noted that the chatbot needs to understand the timing of prompts to avoid being intrusive.

Chatbot Design: While the chatbot in this study was capable of recognizing when a runner was speaking, it still lacks more advanced context-aware abilities. For instance, section 8.4.5

highlights that a runner's silence doesn't necessarily mean they're available to talk, as they may be concentrating on the road for safety. By detecting such attentional states, the chatbot can choose the optimal moment to avoid interrupting. There is prior work in HCI literature on designing ESM prompting systems to be more context-aware and predictive of the best moment [77, 79, 173]. The findings in this chapter indicate different contextual and personal factors to consider when designing context-aware prompting for running.

Self-reports: A chatbot with a higher capability of context-aware proactive prompting would not only improve users' self-reporting compliance rates and contribute to a larger SARE self-reports dataset but also supplement them with useful context-related information.

8.5.3.2 Semi-structured follow-up prompting

The chatbot could be designed with semi-structured follow-up prompting.

Runner Experience: Findings in our study indicated that the chatbot's follow-up questions were useful in bringing participants to enter a deeper level of introspection and self-reflection regarding their SARE.

Chatbot Design: However, designing such follow-up prompts for the chatbot required a significant amount of effort from researchers to develop detailed subtopics and precise dialogue scripts (as demonstrated in the Chapter 7). While predefined prompts offer predictability, certainty, and reliability for chatbot follow-up questions, the process can be time-consuming. To address this issue, fast-growing generative AI methods, such as the large language model GPT-3 [80], could replace researchers in generating exact dialogue scripts following a focused protocol defined by researchers based on research needs and domain knowledge. Thus, such semi-structured follow-up prompts would remain focused but also dynamic and diverse to offer runners with more engaging and personalized conversational interactions.

Self-report: Such semi-structured prompts would facilitate runners' self-reports to stay focused and structured, which has the potential for easy conversion for further analysis and use for computation.

8.5.3.3 Chatbot's retrospective prompting

The findings confirmed the utility of **chatbot's retrospective prompting**.

Runner Experience: Through prompting runners with their previously self-reported information, runners were able to identify trends and patterns in their experiences over time, deepening their understanding and self-awareness, and facilitating the development of strategies to address issues and make improvements.

Chatbot Design: Designing chatbots with retrospective prompts is an effective way to facilitate reflection and encourage users to revisit past experiences. Recent research confirms the value of using chatbots to prompt users to reflect on previous decision-making experiences before making new decisions [216]. Our findings show that the retrospective prompts can present the data as a reference point to elicit runners achieve a more coherent self-reporting behaviour.

Self-reports: Retrospective prompts helped runners compare their past experiences and consistently rate their current experience using the same standard, offering potential to enhance the accuracy and consistency of users' subjective reports and ratings. Previous research in affective computing supports this finding, suggesting that ranking and comparison are more reliable self-reporting methods compared to directly rating emotional states [166]. Therefore, such obtained self-reports facilitated by chatbot's retrospective prompts could lead to more reliable self-reports data for future machine learning models.

8.5.3.4 Systematic and mindful scanning prompts.

Findings suggested the importance of designing a chatbot with **systematic and mindful scanning prompts**.

Runner Experience: According to the participants, chatbots' mindful scanning prompts resulted in an increased attentional width. Runners reported shifting from a narrow focus on a single extreme sensation to a broader awareness of multiple body parts, sensations, and previously unnoticed thoughts and feelings. This could be an effective intervention to be used

in the situation that runners with pain tend to have a narrow attentional width (as showed in Sports psychology in Chapter 2).

Chatbot Design: The chatbot's systematic and mindful scanning prompts can be structured based on the attentional model described in 2.1.3. The model consists of four distinct quadrants of attentional focus (see Figure 2-2). The Broad-External zone can guide the runner to assess various cues from their environment, bodily and mental experiences. In the Broad-Internal zone, the chatbot can help the runner internalize and analyse this information. In the Narrow-Internal zone, the chatbot can assist the runner in selecting one critical thought to report. Finally, in the Narrow-External zone, the chatbot can guide the runner to externalize their self-report.

Self-reports: Systematic and mindful scanning prompts provided by a chatbot can be a useful tool to enhance runners' attentional width, offering opportunities for resulting in multi-dimensional self-reports that take into account not only pain but also other bodily states. This approach may also have the potential to generate an unbiased and holistic self-report dataset, which could be crucial for developing less biased machine learning systems in future. For instance, if a runner is primarily focused on throat pain, they may overlook other areas of tension or tightness that could be contributing to an incorrect running posture. If the machine learning model uses "*throat pain*" as a ground truth label for abnormal hip movement patterns, it will be biased and assume that throat pain is the sole cause of the incorrect hip movement.

8.6 Summary

This chapter presents the results of user studies conducted to investigate the design of a chatbot (with the contrast of R-ESM) for supporting voice-based SARE self-reports during running. The findings demonstrated that the chatbot creates a reliable and reflective experience through the use of voice for self-reporting SARE, and fosters a socially collaborative and empathetic experience, confirming the findings in Chapter 6. The study also reveals four types of prompts that could further facilitate reflection. The results show that runners appreciate both exploratory and unstructured self-reporting, as well as structured and directed but focused self-reporting. Such findings result in the Modality-Autonomy-

Prompting (MAP) framework that can be used to improve the design of the chatbot to gather richer and more thoughtful self-report data.

The next chapter will focus on investigating how runners self-report SARE with the help of provided MAP.

Chapter 9 Understanding the runners' self-reported SARE

The previous chapter's findings indicate that a combination of different levels of autonomy (exploratory but focused thinking aloud and structured conversation) can enhance the thoughtfulness of self-reporting. Furthermore, prior research in sports literature and our own investigations suggest that runners often engage in inner dialogue or think-aloud about their experiences while running. Findings in earlier chapters were based on post-run interviews of runners' experiences with various forms of chatbots. In this chapter, we will delve into the analysis of runners' conversation data while self-reporting during their runs. Our focus will be on addressing RQ4:

RQ 4: What are the dynamic, reflective elements of SARE self-reports, and what should SARE-ARS model?

The primary objectives of this chapter are twofold: (1) to understand the impact of the mixed-structured chatbot, proposed in Chapter 8, on runners' self-reporting processes; and (2) to examine how runners evaluate their SARE during focused think aloud. The first objective helps with understanding the implications of the proposed MAP framework for structuring a chatbot that supports self-reporting of SARE during running. The second objective is to identify further structures that characterize the way runners self-report their detailed SARE while running. While affective computing research typically use simple affect self-reports that can be captured by numbers or labels, recent work has suggested that a qualitative approach is necessary to fully understand the complex nature of these states [229]. However, such work has typically been done in non-ecological settings, such as lab-based interviews, and has predominantly relied on thematic analysis by observers of people's retrospective affective experiences, rather than directly exploring how people subjectively and proactively assess their states while experiencing them during the engagement of the primary activity.

To achieve these objectives, we took two steps. Firstly, we improved our chatbot to provide a mixed-structured approach for self-reporting SARE, and then we qualitatively analysed how this approach influenced runners' self-reporting processes. Secondly, we analysed runners' focused think-alouds data collected using the R-ESM (rather than the chatbot) in Chapter 8. We qualitatively examined the structure and content of this data in conjunction with the

think-alouds data collected using the chatbot in this chapter. It is important to note that the combination of the two types of think-alouds was not intended for a comparison of the two approaches, but rather to provide a larger dataset to understand runners' SARE processes. The following sections provide detailed information on the study methods, procedures, and findings.

9.1 Evaluating the impact of a mixed-structure chatbot on SARE self-reporting

9.1.1 Methods

9.1.1.1 Design of a Mixed-structure Chatbot (Focused Thinking Aloud + Directed Chatbot Conversations)

The chatbot used in Chapter 7 (called the **Structured Chatbot** hereafter) has been refined to create mixed level autonomy in the self-reporting. This new chatbot provides runners with a **Focused Thinking Aloud** space to exploratorily and iteratively elaborate on their subjective and affective running experiences. It is called focused because it specifically focuses on SARE, and is not intended for general thinking aloud. The focused think aloud was then followed by specific **Chatbot directed conversation** about the selected topic. Figure 9-1 illustrates the Mixed-structure Chatbot's conversation flow, with the focused thinking aloud space highlighted in blue and the structured conversation interactions highlighted in red. To improve the conversation, the chatbot has been redesigned to include more social and small talk, and to provide more supportive and empathetic feedback. A complete example of a conversation can be found in Table 9-2.

It should be noted that this research was done in conjunction with an MSc student who was under the supervision of the author of this PhD thesis. The PhD candidate designed and fully implemented the chatbot using the MAP framework whereas the MSc student improved the structured questions with more social and empathic language (not investigated in this thesis). After data collection using the chatbot, the PhD candidate analysed the runner's conversation data to study the impact of the design on self-reporting process, while the MSc student interviewed the runners to study the effect of emotional and social design for reflection and to confirm the results of Chapter 8 (MAP framework). The MSc student's results can be found in [67]. The results of the interviews and analysis done by the MSc student are not reported in this thesis but they confirmed the usefulness of the mix-approach proposed in Chapter 8.

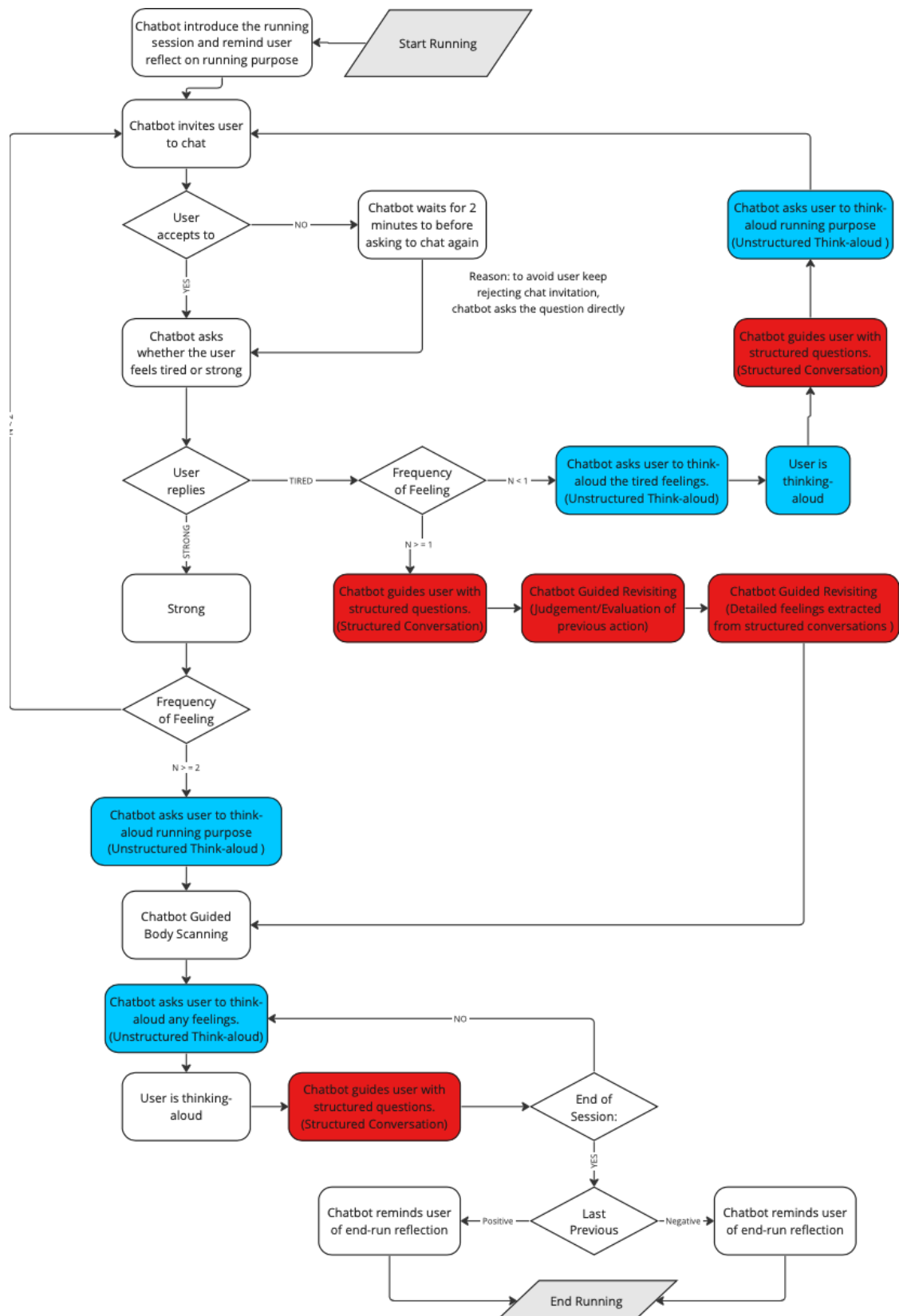


Figure 9-1: Overview of the Mixed-Structure Chatbot (Blue indicates the Focused Thinking aloud; Red indicates the Directed Chatbot Conversations)

9.1.1.2 Participants

Eight university students (5 females, 3 males, with an average age of 23) were recruited who were capable of running for at least 20 minutes at a self-selected pace. The group comprised three native English speakers and five who spoke English as a second language. Six participants ran outdoor parks, while two ran on an indoor treadmill. All participants were compensated with a £25 Amazon voucher. The average running duration was 27.12 minutes, with a standard deviation of 5.14.

Table 9-1: Participants' Profile (Indoor Treadmill (IT), Outdoor Track (OT), Outdoor Park (OP), M: Male, F: Female)

Participant	Gender	Time-Location	Running Experience Level
P1	F	27min-OP	Beginner
P2	F	32min-IT	Beginner
P3	F	22min-IT	Intermediate
P4	M	28min-OP	Beginner
P5	M	33min-OP	Intermediate
P6	F	26min-OP	Beginner
P7	F	18min-OP	Beginner
P8	M	31min-OP	Intermediate

9.1.1.3 Procedure

The study began by providing participants with a tutorial on operating the chatbot and practicing the thinking aloud technique, as outlined in Chapter 8. The conversations between the runners and the mixed-structure chatbot were recorded and logged as part of the data collection process.

In addition, the same participants took part in a second running session (order randomized) as part of an MSc student project [67]. The structured chatbot presented in Chapter 7 was used in this session, with the aim of comparing the two chatbots in the post-running interviews. Although this part of the work is not discussed in this thesis, the MSc students' interviews confirmed the results reported in Chapter 8 on the use of the Mixed-Approach

9.1.1.4 Data Analysis

The qualitative analysis of the conversation data gathered during running utilized a thematic approach, specifically focusing on comparing and identifying connections between the self-reported data gathered from the exploratory thinking aloud and the directed conversation structure to understand their impacts.

9.1.2 Results

Table 9-2 provides an example of mixed-structure conversation from P2 incorporating both focused thinking aloud (unstructured and exploratory) and directed conversation (structured). The following sections will provide an in-depth exploration of each theme, supported by evidence and relevant quotes from all participants.

The analysis of the conversation with the mixed structured chatbot identified three themes that demonstrate the impact of this mixed-structure on SARE self-reporting during running. These themes include a self-report structure as a learning process, a self-report tuning process, and a self-report personalization process.

Table 9-2: Examples of P2's Conversation Data

Conversation Session	Role	Conversation History
Start	Chatbot	<i>Hi! I am your Run Bot. I will be running with you today. During the run, you will be thinking aloud, and we will also be having conversations about your running experience.</i>
...	...	...
Directed Chatbot Conversation (Structured)	User	<i>Hey there! How are you doing? I remember you were feeling a bit tired earlier. You said you were going to forget it. Has this made you feel better?</i>
	User	<i>yeah better</i>
	Chatbot	<i>How amazing you are! Can you briefly tell me how?</i>
	User	<i>My breath becomes consistent, and my arms were initially feeling heavy, but now it's not anymore.</i>

Chatbot *How long has this been since you start to feel like this?*

User *2 minutes*

Chatbot *Oh, I also remember you had some heavy sensations on your core. Your overall exertion level is 4. You said your reaction to this tiredness was to forget it. How are you feeling right now? Can you pick a number on a scale of 1 - 5 to rate how much effort you perceive right now?*

User *the number is 3*

Chatbot *What factors have you investigated when you rate the effort level as 3?*

User *My breath*

Chatbot *If you are running too hard, you have the power to back off the pace until you find the right efforts, the efforts you feel comfortable while running. You know you are in control. By the way, I have one more question. How would you classify your emotional state right now? is it good, bad, or neutral?*

User *Good*

Chatbot *Can you describe your good feelings in more specific words?*

User *Excited because I'm feeling better*

Chatbot *Thanks for sharing you experience. Great efforts! You are doing great! You are certainly enjoying the run! Keep it going! I will chat to you later.*

Off

**Focused
thinking aloud
(Unstructured)**

Chatbot *Hey there! I'm back! How did everything go for you? How are you feeling at the moment? Take a minute to think aloud and describe your physical and emotional state. Don't worry, there's no right or wrong answer! When you're ready, just let me know and we can move on to some more specific questions.*

User *The guide was from head to feet. So firstly, I'm focused on my head. Because at the start of the exercise, I felt very heavy and not flexible on my upper body. But after about 20 minutes, I feel light. I don't know because it might be because my breath becomes consistent, and my arms were initially feeling heavy, but now it's not anymore. So, I feel my body becomes light as well. I feel my emotional state becomes happier and excited,*

		<i>and it makes me feel quite in a good mood. Then, it goes to my chest. But there's an awful sensation: the sweats over my chest. It makes me like not very happy. And, uh, then it goes on my legs. Fortunately, they are just feeling fine.</i>
Directed Chatbot Conversation (Structured)	Chatbot	<i>On a scale of 1 to 5, where would you rate your perceived exertion level right now? Let me break it down for you: 1 means it's super easy and comfortable, 2 means it's still pretty easy and you feel very comfortable, 3 means it's moderate and you can tell you're working but it's still within your comfort zone, 4 means it's hard and outside your comfort zone, and 5 means it's at your physical limit. So, what number are you feeling right now? Don't be shy, just let me know!</i>
	User	<i>3</i>
	Chatbot	<i>There are many factors that influence your perceived efforts. This could be due to muscle fatigue, stiffness or tightness, breathing difficulty, boredom or stress, uncomfortable sensations, lack of energy, or just simply not being in the right mood. Based on which factors, did you rate your efforts level as 3?</i>
	User	<i>sensation</i>
	Chatbot	<i>What sensations do you feel?</i>
	User	<i>Sweatiness</i>
	Chatbot	<i>Which part of your body feels such sweatiness sensation?</i>
	User	<i>My chest</i>
	Chatbot	<i>How long has this been going on? You can measure it in seconds or minutes.</i>
	User	<i>Around 3 minutes</i>
	Chatbot	<i>How would you like to deal with it? keep current status, slow down the speed, or power through?</i>
	User	<i>take it easier</i>
	Chatbot	<i>If you are running too hard, you have the power to back off the pace until you find the right efforts, the efforts you feel comfortable while running. You know you are in control. By the way, I have one more question. How would you classify your emotional state right now? is it good, bad, or neutral?</i>

	User	<i>neutral</i>
	Chatbot	<i>Can you please explain or specify more?</i>
	User	<i>not very happy or comfy</i>
	Chatbot	<i>Thanks for sharing you experience. This has been a good run so far. Good job. Keep it up! I will chat to you later.</i>
...		
END	Chatbot	<i>Thank you. Bye!</i>

9.1.2.1 A self-report structure as a learning process

First of all, the participants appeared to learn the language of the structured dialogue and incorporate it into their unstructured focused thinking aloud. For example, as shown in Table 9-2, the body scanning exercise, as a part of the structured dialogue, guided the runner to systematically scan their different body parts and their feelings during the unstructured focused think aloud bringing a structure to it over time. When P2 was asked to think aloud, they adopted this structure to organize their thinking aloud in a logical manner with not only reflections on multiple body parts (e.g., *"The guide was from head to feet. So firstly, I focus on my head Then, goes to my chest And, uh, then goes on my legs."*) and in scanning structurally and sequentially their body, she reports of various feelings and sensations (e.g., *"breath becomes consistent", "feeling heavy", "sweats", "good mood", etc.*).

Another example of learning from the structured conversation is reporting quantified information about their feelings during their thinking aloud. In the structured dialogue, the chatbot similarly asked participants to quantify their perceived exertion intensity using a number between 1 and 5. Although our participants did not particularly use a numerical scale in the think aloud session, they typically used more semantic and descriptive words (e.g., *"the exertion level is very low,"* P5) to provide specific explanations for the ratings they had provided in the structure dialogue (e.g., *"Just been walking for the past minute, so feeling fully recovered,"* P5). This also applied to the quantification of the duration of their feelings. For

example, P2 thought aloud, *“At the start of the exercise, I feel very heavy and not flexible on my upper body. But after like 20 minutes, I feel quite light.”*

In addition, the unstructured thinking aloud also showed that participants proactively supplemented their initial report of emotional state with further explanation, as if the chatbot had asked them to do so. As P3 said, *“I feel very very good right now,”* and then complemented with further elaboration - *“I feel my emotional state becoming happier and excited, and it makes me feel quite in a good mood”*.

Participants also used the chatbot’s phrases in their think-alouds. For instance, when the chatbot asked participants how they would like to respond to their feelings, it prompted users with such as *“push it harder,” “take it easier,”* or *“maintain the current status.”* These expressions were then utilized by the participants in their unstructured think-alouds. For example, P6 mentioned, *“I’m taking it easier [...] I probably want to run, but it feels really hot as soon as I’m in the sun, and then I just don’t want to run anymore. I feel pretty lazy. I’m just walking, but I think at least it means my body feels okay. I probably could run a bit more if I wanted to.”*

Overall, these findings showed that the format of questioning and answering during the structured dialogue contributed to a learning process in which runners learned to articulate their thoughts during the unstructured focused think-aloud. The focused part of the think aloud and the structured dialogues become hence more connected facilitating possibly deeper exploration and self-understanding.

9.1.2.2 A self-report tuning process

Following comparison of the data from the unstructured think-alouds and the structured dialogues, it was observed that, although participants voiced a variety of feelings and sensations, they tended to emphasize their most salient experiences when speaking with the chatbot in the structured conversation. For example, as shown in Table 9-2, during the exploratory thinking aloud, P2 mentioned how the sensations on their upper body (e.g., arms) changed from feeling heavy to a light sensation. At the same time, P2 also mentioned an uncomfortable sensation of sweating on their head and chest. However, during the structured

conversation with the chatbot, they only described the more unpleasant sensation of sweating when responding to the chatbot's question.

Similarly, P5 initially thought aloud about various feelings in their muscles and lungs, and a mild, swollen sensation in their legs and chest; yet, when interacting with the chatbot in a structured format, they only spoke of the more prominent sore sensation in their legs. Additionally, P1 described multiple sensations, such as hunger, fatigue, muscle sensations, and breathlessness, in their think-alouds; nevertheless, during the structured dialogue, they reported a *"headache sensation"* that had *"emerged 2 or 3 minutes ago"*. Although this headache sensation was not explicitly noted in P1's think-alouds, it is still connected to their acknowledgement of a lack of sleep, i.e., *"I only slept for about four or five hours last night"*. These findings indicate that the structured conversation with the chatbot can encourage participants to prioritize and concentrate on their most meaningful feelings when reporting, thus facilitating a self-report tuning process in the mixed-structured conversation.

9.1.2.3 A self-report personalisation process

The use of directed chatbot conversations can be beneficial in enabling runners to provide personalized categorization and interpretation of their SARE in their self-reports. For instance, within the structured conversation with the chatbot, participants were consistently prompted by the Chatbot to classify their emotional state as either *"good, bad, or neutral"* and to further *"explain and specify more"* with respect to their evaluation. Participants' responses during the directed chatbot conversation are demonstrated in Table 9-3. As shown, the classification labels of 'good/neutral/bad' includes a broad range of emotional experiences. For example, both P1 and P6 reported a *"neutral"* emotional state; P1 noted being *"a little tired"* physically but *"really calm and relaxed"* mentally, while P6 reported feeling physically *"energetic"* but sensationally fatigued in their legs. The interpretation of *"good"* as an emotional experience may also be linked to individual expectations, as shown in P7 statement of feeling *"energetic"*, *"clear-minded"*, and *"very strong"*, or P3's description of feeling *"good"* due to perceiving *"sweat"*. This shows that simple emotional terms might hide the complexity and the multifaceted nature of the SARE, indicating the requirements for not relying on simple labels of them but unpacking the rich elements of SARE when building automatic recognition

systems in the future. The conversation flow informed by the MAP framework could help runners to give more refined descriptions and reflection of their SARE.

Table 9-3: Example structured conversation for self-report personalisation (P # refers to Participants ID)

P#	Classified self-reports	Personalised self-reports
P1	neutral	Uh, it's a little tired, but I'm feeling really calm and relaxed.
P2	good	excited
P3	good	I sweat a lot so that makes me feel good.
P4	good	I'm feeling good even though my legs are very tired. My mind is very powerful and very clear, but my legs are very sore, so I can't run very fast anymore
P5	good	very excited even though a bit tired.
P6	good	Feel proud of myself. Feel okay.
	neutral	a bit annoyed
P7	good	I am energetic and I'm clear minded, and I feel very strong.
P8	great	I feel my mind situation is good. I feel very clear and powerful and dynamic. but my legs are tired

9.1.3 Summary

The analysis of the mixed-structure conversation data revealed a positive impact on the self-report data. After engaging in the focused thinking aloud space, where runners gave various self-reports about their feelings, they tended to refine their answers by concentrating on their most salient feelings during the structured conversation. This highlighted the importance of the order of the mixed-structure conversation. Providing a focused but exploratory space before the structured questions allowed for a natural process of formation and exploration, while the structured part served as a valuable tuning, highlighting and confirmation process.

The next section will focus on obtaining a deeper understanding of runners' natural exploration process while thinking aloud about SARE during running. This understanding will inform the improvement of the mixed-structured chatbot conversation design to better support this process and further enhance the runners' self-reporting experiences and the gathering of self-reports.

9.2 Understanding the nature of runners' exploratory and unguided SARE think-alouds

To gain a more comprehensive understanding of how runners naturally think about their experiences, we analysed their think-aloud data collected without the use of a chatbot but with the simple R-ESM described in Chapter 8 section 8.2.2.1. The findings from this analysis led to the development of a framework to guide the design of chatbot conversation topics that better support runners' self-reporting. The following sections will describe the methods used in this study, present the resulting framework, and discuss its design implications.

9.2.1 Method

The datasets consisting two parts of focused think-alouds data from runner participants. The primary dataset comprises think-alouds data collected in Chapter 8 section 8.2.2.1, using the R-ESM system (i.e., no chatbot), from nine participants (five females, four males) during their running sessions (Avg = 35.56 minutes, Std = 10.78). In addition, we also included the focused think-alouds from the conversation data collected using the mixed-structure chatbot in Section 9.1.1.

These data were fully transcribed and thematic analysis of such think-alouds data was conducted. Differently from section 9.1, here we focussed on how a runner would generally elaborate on their running experience and sensations and what further topics they would consider.

9.2.2 Results

Three themes were identified in the analysis of runners' natural think-alouds data about SARE: Description, Appraisal, and Management. To enhance readability, participant IDs were omitted from the quotes as they were brief and frequently paraphrased.

9.2.2.1 Description

The **description component** refers to how runners articulated their experiences of SARE in the moment, which includes four sub-components: descriptive elaboration, quantification, bodily location, and temporal information.

Descriptive elaboration: the runners tended to provide elaborated description in addition to their initial statement on how they felt. For instance, one participant reported a sensation of *“weird pins and needles”* when beginning to think aloud, and then further elaborated by providing additional sentences, such as, *“It’s the feeling you get when you come inside to the warm after you’ve been really cold.”* This additional elaboration not only completes the runner’s thoughts, but also provides valuable information that can be used to further characterize the sensation. For example, the cause of this sensation may be attributed to changes in bodily temperature.

Other elaborated sensations mentioned in their think aloud are listed here, such as *“sensation”, “cold”, “slight”, “tingle”, “cold”, “cold air”, “sort of chills your throat”, “tightness”, “weird pins and needles feeling”, “the cold air is hitting on”, “the feeling you get when you come inside to the warm after you’ve been really cold”, “sensation of the cold air”, “got the cold lungs”, “cold and wet”, “tighten up”, “painful”, “wet”, “warm”, “colder”, “sweater”, “aches”, “stiffness”, “tight”, “force”, “tightness”, “breathing heavier”, “hurt”, etc.*

“The elaborations related to emotions are also listed here, such as “bored,” “annoying,” “tired,” “desire,” “embarrassment,” “repetitive,” “variety,” “worrying,” “motivated,” “emotion,” “rewarding,” “struggling,” “good,” “weird,” “refreshed”, etc.

Quantification: Runner verbally quantifies the severity or intensity of their feelings; it is more common for them to use descriptive words, such as *“very”, “a little bit”, “quite”, “pretty”, “always”, “fairly”, “a lot”, “may”, “maybe”, “slight”, “so”, “very”, “not very”, “not too bad”, “most”, “too much”, etc.* rather than numeric numbers, such as 1, 3, or 5, when expressing the intensity of their feelings. This suggests, as also observed in the conversation of chapter 9, that the structure conversation could possibly be more natural and meaningful if such language were to be integrated in those expected answer rather than just expected numbers.

Bodily location: Runners might mention their body parts if their feelings are caused or related to sensations located on a specific part of their body. The commonly mentioned body parts as listed here, such as *“left”, “knee”, “throat”, “chest”, “ankles”, “lungs”, “ankles”, “Achilles”, “right ankle”, “legs”, “right calf”, “calf”, “back”, etc.*

Temporal information: Runners provide temporal information regarding their perceived SARE when verbalising them, such as explicitly specifying the time when it began, estimating its duration, or employing location or distance information to denote when the feeling occurs. As shown in the Table 9-4 below, category 1 includes the expression of stating the feelings that occur at the current moment, indicated by people using the present continuous tense, such as *“right now I’m feeling the aches in my legs”*, *“not feeling too many emotions at the moment”* and *“currently feeling still neutrally exerted”*. Category 2 includes the expression of stating a feeling that has just started at a point, with the keywords people often use being *“starting to get the XXX feeling now”* and *“just getting into”*. Category 3 includes the expression of stating a feeling using the complete tense to indicate a period of time, which can be either exact (e.g., *“2 minute”*) or ambiguous (e.g., *“a while ago”*). Category 4 includes the expression of how a person regulates and expects how they would like to feel in a near future time, such as *“gotta be a bit easier for a couple of minutes”*.

In addition to these four categories based on the tense of their statement, the temporality of a feeling can be indicated using a time duration (e.g., for 2 minutes, as shown in Category 5), a running distance (e.g., 100 meters, as shown in Category 6) or locations of running route (e.g., the corner, as shown in Category 7) as a temporal indicator. For location, a person could say *“gotta be a bit easier as I get to the tail end”*, *“gotta be a bit easier as I turn on the corner”* and *“not going to put a strain on because it’s the last one lap”*. For the running distance, a person could say, *“just got to get motivated to run into the wind in 50 yards”* and *“the next 300 yards should be very comfortable with the downwind”*.

Table 9-4: Temporality

Category ID	Category	Example quotes
1	The feeling is occurring presently, which is described using the present continuous tense.	<i>"Right now, I'm feeling the aches in my legs", "Not feeling too many emotions at the moment", "Currently feeling still neutrally exertion"</i>
2	The feeling has begun to appear.	<i>"Just getting into XXX", "Starting to get into", "Starting to get XXX feeling now"</i>
3	The feeling has been present for a period of time, which is depicted using the present perfect progressive tense.	<i>"I've been feeling breathless for a while"</i>
4	The feeling will be self-regulated or managed in the near future.	<i>"Turn on the corner, gotta be a bit easier for a couple of minutes"</i>
5	The feeling's temporality is described by its temporal duration.	<i>"Going to be a little bit easier for the next minute or two"</i>
6	The feeling's temporality is described by its distance of the run.	<i>"Just got to get motivated to run into the wind in 50 yards", "The next 300 yards should be very comfortable with the downwind"</i>
7	The feeling's temporality is described by its geographical position of the running route.	<i>"Gotta be a bit easier as I get to the tail end", "Gotta be a bit easier as I turn on the corner", "Not going to put a strain on because it's the last one lap"</i>

9.2.2.2 Appraisal

The **Appraisal component** refers to how runners evaluate their feelings and hence the action to take in response. Five appraisal checks were identified in the conversation: Familiarity, Severity& Tolerance, Affective Association, Revisiting & Comparing, and Future Prediction.

Familiarity: Runners tended to assess the familiarity of their body experiences. They might for example judge the occurred feelings using phrases like *“not unexpected”, “should be used to it in the UK”, “nothing unusual, even for such a cold day”, “nothing unusual there”, “that was unexpected”*.

Severity & Tolerance: Runners tended to assess the situation by considering its severity and impact on them. For example, one reported a pins and needles sensation and assessed it as *“not very comfortable but not too bad,”* and another who experienced a slight tingle on their left knee may evaluate it as *“nothing bad really.”* Some other examples include *“I would worry about it too much”, “it’s not too difficult to deal with at the moment”*.

Affective Association: Runners also gave their affective associations to their perceived sensations, such as *“It’s an uncomfortable thing”, “[the sweats over my head and my chest] just makes me like not very happy”*.

Revisiting & Comparing: The appraisal also involves revisiting and recalling previously reported sensations, indicated with words such as *“still”* and *“again”*. Instances include: *“still feeling a little bit cold”, “still having a runny nose which is a bit annoying”, “still feeling cold now”, “still not feeling very tired”,* and *“still having that tightness in my calf now”*. They also give direct comparisons to a previous state, such as *“not as bad as it was before,”* as well as a relative ranking, e.g., *“it’s slightly getting worse”, “feeling a bit easier now”,* or *“my breathing is heavier now”*.

Future Estimation: Additionally, individuals may provide forecasts or estimations of future states, such as *“I am sure it will improve shortly”, “it’ll probably get more annoying after I get more tired”*

9.2.2.3 Management

The **Management component** refers to the when runners articulated how they intend to react to a certain type of SARE and how they act. It has the following elements: intention, action and its re-evaluation.

Intention refers to what runners intend to do to regulate their feelings and sensations, as identified in the conversation (See Table 9-5). For instance, one runner stated, *“I want to warm up, so I am just focusing on warming up,”* when experiencing the uncomfortable sensation of *“now feeling a little bit cold”*. In another instance, a runner showed an intention to change their boredom *“a little bit bored right now, maybe I should run a bit faster.”*

Table 9-5: Examples of Runners’ intentions

Intentions	Quotes	Context	Quotes	Sentiment
Intention to change sensations.	<i>“I want to warm up so just focusing on warming up”</i>	Cold weather or body stiffness discomfort	<i>“Now feeling a little bit cold”</i>	Negative
Intention to change boredom by increasing their running effort	<i>“Maybe I should run a bit faster. Increasing speed”</i>	Boredom from repetitive running route or lack of challenge	<i>“A little bit bored right now”</i>	Negative
Intention to change the annoying effect by clearing one’s nose	<i>“Would be nice to blow my nose”</i>	Annoying emotion caused by running nose (congestion, allergy)	<i>“Still got that runny nose which is a bit annoying”</i>	Negative
Intention of increasing efforts	<i>“Thinking about doing some sprinting”</i>	Wanting to push oneself to the next level	<i>“Make things up a little bit to tire me out a bit more”</i>	Positive
Intention of changing running route	<i>“It would be nice to change location”</i>	Boredom from repetitive running route or desire to explore new areas	<i>“I’ve been running around the same loop for the whole time it’s getting a little bit repetitive ...(it) gets you a bit more bored”</i>	Negative

Intention of reducing effort	<i>"I should just slow down instead"</i>	Heaviness in breathing or muscle fatigue	<i>"My breathing is heavier now"</i>	Negative
Intention of increasing effort	<i>"Going faster"</i>	Feeling good or wanting to challenge oneself	<i>"It feels good nice to run"</i>	Positive
Intention of adjusting temperature	<i>"I'm going to take my hat off"</i>	Uncomfortable warmth or needing to regulate body temperature	<i>"I'm pretty warm"</i>	Negative
Intention of reducing efforts	<i>"Gotta be a bit easier for the next minute or two"</i>	Excessive exertion or need to conserve energy	N/A	N/A
Intention of self-motivating	<i>"Just got to get motivated to run into the wind in 50 yards"</i>	Hardship and difficulty caused by weather or mental state	<i>"Downwind" and "don't want to talk to, just concentration"</i>	Negative
Intention of finding running flow	<i>"So, trying to fit into a routine. Turn on the corner, gotta be a bit easier for a couple of minutes"</i>	Uncomfortable sensation or desire to establish a rhythm	<i>"Weather got a bit colder"</i>	Negative

The **Action** element refers to participant's actions of making changes as a result of their intention for the sensations they experience (see Table 9-6). For instance, one participant first reported the uncomfortable sensation of *"cold lungs feeling"* and then verbalized the intention of *"probably breathing through the nose"* and the action of *"I am now trying to do that purposefully"*. Similarly, another participant verbalized their actions of *"moving feet more along the ground instead of upwards"* after feeling *"using too much force in each drive"*. Besides, participants also evaluate the effects of the chosen actions. For instance, one participant verbalized their action of *"changing to a different location"* after reporting the

feeling of boredom. Then, they re-evaluated their action saying *“when you run somewhere else, it takes your mind off the fact that you're running.”* Similarly, P5 reevaluated that their action of taking a recovery helped, saying *“I have just been walking for the past two minutes, so feeling fully recovered.”*

Table 9-6: Example of Runners' Actions

Actions	Quotes	Context	Quotes	Sentiment
Trying out new running route	<i>“Changing routes slightly onto the footpath”</i>	When experiencing the boredom of a repetitive running route	<i>“The running trucks getting a bit repetitive at this point”</i>	Neutral
Adjusting breathing pattern	<i>“Trying to breathe in my nose instead of my mouth”</i>	When experiencing the sensation of cold air in the throat	<i>“Cold air coming into my throat. that feeling you get when it sorts of chills your throat a lot”</i>	Negative
Slowing down	<i>“So, slowing down a little bit”</i>	When experiencing discomfort, particularly from the cold air in the lungs	<i>“The cold air in my lungs is the most uncomfortable thing at the moment”</i>	Negative
Adjusting breathing pattern	<i>“Probably breathing through your nose so I'm trying to do that purposefully”</i>	When experiencing the sensation of cold lungs	<i>“Cold lungs feeling”</i>	Negative
Adjusting breathing	<i>“Breathing now; breathing in from the nose not from the mouth”</i>	N/A	N/A	Negative
Reacting to bad weather	<i>“Putting my hood now”</i>	When experiencing bad weather, specifically hailstone	<i>“Oh, it's actually hailstone”</i>	Negative

Adjusting running posture	<i>"Moving feet more along the ground instead of upwards"</i>	When experiencing the sensation of heaviness in the legs	<i>"Using too much force in each drive".</i>	Negative
Adjusting breathing	<i>"Breathing through your nose"</i>	When experiencing uncomfortable sensation of cold lungs	<i>"This cold lung feeling".</i>	Negative

9.2.3 Summary

This section aims to deepen our understanding of runners' natural process of exploring and thinking aloud about their subjective and affective running experiences (SARE) during running. This understanding will inform the improvement of the mixed-structured chatbot conversation design to better support this process, enhance runners' self-reporting experiences. In the following sections, we will discuss the findings and explore how they can help answer the **RQ 4: What are the dynamic, reflective elements of SARE self-reports, and what should SARE-ARS model?**

9.3 Discussion

9.3.1 Impact of Modality-Autonomy-Prompting (MAP) on runners' SARE self-reporting

The findings showed that the proposed mixed-structure chatbot enabled a learning, tuning and personalization process in runners' SARE self-reporting during running. Such results confirm the efficacy of the proposed MAP framework as a design guideline, and provide explanations and evidence how the chatbot designed with MAP framework can facilitate runners' self-reflections in their self-reporting process.

Overall, runners' self-reporting experience is enhanced with an enhanced self-reflection. The entire process, from learning, tuning to personalisation, is a rigorous self-reflective exercise, as participants proactively and progressively work on their thinking. Therefore, the mixed-structure approach to conversation design proposed in this study represents an extension to the one used in [216] for improving meta-cognition enhancement, where the chatbot was designed with simple questions that led users to revisit previous decision-making processes and aid in current ones. More specifically, the process of learning, tuning, and personalization

in our study provided conceptual validation for the use of MAP Framework to design a mixed-structure chatbot for self-report data. This was reflected in several ways. First, we found that the chatbot's prompts and reflective questions in the directed conversation inspired runners in their thinking aloud with relevant languages and structures, which could result in more structured and linear think-alouds data.

Second, the mixed-structure conversation facilitated nuanced self-reporting, which created possibilities for revealing and elucidating some types of SARE that may have previously been implicit or hidden in runners' exploratory think-alouds. Additionally, the structured conversation also encouraged runners to give refined answers by prioritize their most salient and significant experiences to self-report, in contrast to everything as they said in the think-alouds. The order of the exploratory think-aloud and structured conversation was hence important, as the thinking aloud space allowed for exploration and formation, while the structured part served as a valuable tuning and confirmation process. In addition, personalization in self-reporting allowed runners to interpret and define their SARE state according to their own experiences and perspectives, which is essential to account for the individual differences often observed in self-reported emotional states in affective computing research [30] .

9.3.2 Description-Appraisal-Management (DAM) Framework for designing Chatbot for richer SARE self-reports

Based on the results in Section 9.2.2, the Description-Appraisal-Management (DAM) framework has been derived to conceptualise how runners self-reflect and report their subjective affective states while running. As illustrated in Figure 9-2, the DAM framework comprises three components: description, appraisal, management. Description encompasses descriptive elaboration, quantification, bodily location, and temporal information (e.g., time, location and duration, etc.). Appraisal involves the evaluation of familiarity, severity& tolerance, affective association, revisiting & comparing, and future prediction. Management encompasses their intentions and its context, actions and its context and re-evaluation. This model could capture all the elements of the self-report, providing a better understanding of the individual's state and think processes to guide their actions in response.

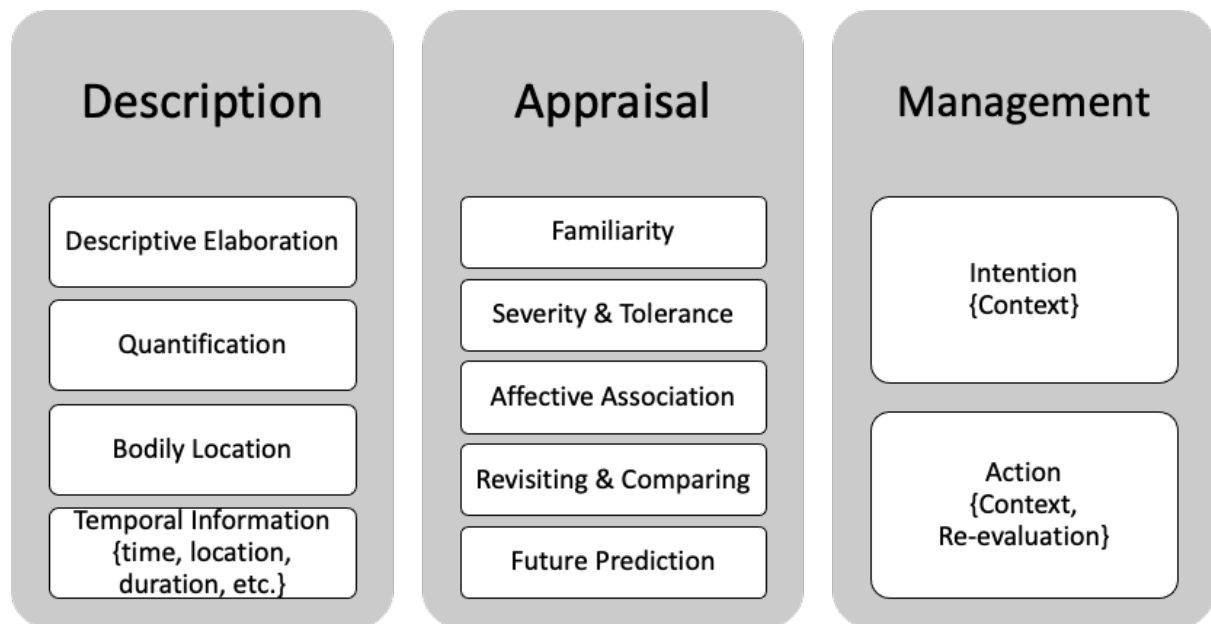


Figure 9-2: The Description-Appraisal-Management (DAM) Framework about how Runners Naturally Reflect and Report Subjective Affective States

The chatbot can be further developed with the guidance of the DAM framework. Figure 11-8 shows the refined version of the general conversation flowchart (see Appendix 11.6), whereas Figure 9-3 and Figure 9-4 demonstrate what questions and how the chatbot could ask SARE related questions. Such a chatbot is capable of extracting complex ground truth information through structured questioning. However, during the exploratory thinking-aloud process, the runner may have already mentioned some of this information, as demonstrated in Section 9.2.2. Therefore, training the chatbot with specific NLP capabilities based on the runner's think-aloud data can enable it to identify and extract ground truth information directly during this stage, reducing the need for multiple questions during structured conversations.



Figure 9-3: Example of in-depth structured Chatbot conversation questions for the topics of sensation, discomfort, or pain



Figure 9-4: Example of Example of in-depth structured Chatbot conversation for topics of Emotion, Affect, Motivation, Psychological State, etc.

The proposed DAM framework suggests a shift from a simply rating and category based SARE self-reports to DAM-based self-reports that include rich detailed and descriptive elaborations of SARE states. This rich description allows for differentiation and contextualization of SARE states, such as what type of feelings, what the intensity of it, and where it is located on their body, and how long it has been present. Similarly, specifying the cause of a sensation as changes in bodily temperature, might inform not only the design of the structure of chatbot conversations to aid in such specification, but also inform the design of wearable devices and sensors to capture the relevant behavioural signals (e.g., using body temperature sensor) to infer specific aspects of the runner's state.

Description of the location of one experienced sensation could also be useful to differentiate between sensations and their effects on behaviours. For example, the level of desire-to-stop caused by a pain sensation in the right knee may differ from the same level of desire-to-stop caused by pain sensation in their shoulder. Furthermore, datasets of such elaborations could possibly be explored to train NLP processing with SARE-specific language to support the chatbot to directly extract key information from people's think-alouds. Thus, such elaborative elements, beyond the discrete, categorical and numeric values, could be used to represent more nuanced affective states in the affective computing research community [59, 198-202].

The DAM framework also indicates the importance of considering appraisal and action-tendency dimensions when collecting SARE self-reports. Appraisal-based self-report data may provide insights on whether a runner is familiar with the experience, its severity and tolerance levels, affective associations, anticipated changes, etc. Similarly, self-report data regarding one's action tendency (e.g., what the person is ready to do) may encompass the runner's preferred approach to manage the experience, their actual responses, and the effectiveness of their actions. This information can be useful for teach SARE-ARS to not only learn a runners' state of SARE but also better understand how a runner think and infer the runner's needs. For example, in LDRE (as discussed in Chapter 4), SARE-ARS can not only identify the location and intensity of a runner's pain but also recognize their preferred approach to address it, thereby enabling a spectator to provide targeted support. Future work could explore the possibility of using such information not only to build SARE-ARS but also using machine learning to train systems to provide support and recommendations on what actions could be considered.

Additionally, developing such chatbot systems for long-term data collection could involve runners as citizen scientists [130, 210]. Our interview results (e.g., from the studies in Section 8.4) suggest that runners were interested in taking on such a role in contributing data (e.g., conversation data with the chatbot and running data via sensors) for research purposes. They also recognized the value of the developed systems and using the chatbot to gain insights into their running. However, understanding how to engage runners as citizen scientists and how the industry can utilize this engagement requires further research outside the scope of this thesis.

9.4 Summary

In this chapter, we examined the impact of a mixed-structure chatbot on runners' self-reported SARE and validated the effectiveness of the MAP framework for chatbot design. The results showed a learning, tuning, and personalization process in runners' self-reporting, suggesting the opportunities for enabling the identification of implicit SARE self-reports, providing accurate and well-tuned SARE self-reports. Furthermore, we investigated how runners naturally express their thoughts during SARE and developed the DAM Framework to suggest more comprehensive self-reports data. The framework not only captures detailed descriptions of SARE states but also includes information about runners' appraisals and management strategies. This additional information gathered in the self-reports could be potentially used as richer ground truth in the future to enhance the automatic recognition system by providing insight into runners' thinking and reasoning processes while reporting their SARE states.

Chapter 10 General discussion and conclusions

This thesis seeks to offer novel insights on the design of technology for running experiences that go beyond practice, competition, and performance. The thesis builds upon the existing psychology literature which suggests that the running experience is not limited to objective metrics such as distance, speed and physiological measurements, but instead comprise multiple interacting components including physiological, affective, cognitive, and social states and their dynamic interactions and processes.

Therefore, the focus of this thesis was to understand the importance of capturing these various aspects of the running experience and how chatbots could facilitate the necessary data collection to create systems that can automatically recognize them. Rather than relying solely on the typical self-reports gathering used in affective computing where affective states are simplified into numerical ratings, the thesis explored the possibility to capture the richness and complexity of the SARE . As a result, this thesis led to propose a novel approach to gather SARE during running, opening new opportunities for the modelling of SARE when building SARE-ARS. At the same time the proposed approach for SARE self-reporting emerges to be of use in itself to runners. Hence, while the original focus of the second part of the thesis was on methods for SARE data collection processes, the findings suggest further uses of the proposed chatbots-conversation design frameworks such as: SARE self-report for self-coaching and SARE-self-report for direct sharing with friends and family without the need for an automatic recognition system.

We discuss below the different contributions made by this thesis in addressing our main research questions. We then discuss limitation and future research as well as the ethical issues that our proposed approach raises.

10.1 Summary of Research Questions, Methods and Discussion of Findings and Contributions

Our main research question was ***RQ: How can technology be designed to support runners in self-reporting their subjective and affective running experience (SARE) while running?*** To achieve this goal, the thesis posed three specific research sub-questions discussed below.

RQ1: What are the components of SARE that matter when sharing, and why?

The Chapter 4 tackled RQ1 by analysing a range of qualitative data, including semi-structured interviews with runners and spectators, online LDRE articles, and onsite observations of marathons and user reviews and comments on LDRE applications. The findings provided a rich and detailed understanding of runners' subjective and affective experiences during long-distance running events in relation to their interactions with spectators. It proposed the DePECS framework to inform the design of technology for sharing between runners and spectators during long distance running. The DePECS framework structures the multifaceted SARE, sensational experience, discomfort and pain, cognitive and emotional, interactions among different types of experiences. In addition, it goes beyond the current literature in sport psychology and in HCI by identifying the phases and complexity of sharing and how this informs the support the runner needs.

Our framework enhanced the current understanding of runners' thoughts and desires regarding sharing their experiences. It demonstrated that existing HCI technology for sharing running experiences, which focuses primarily on physical performance metrics [133, 169] or simple forms of encouragement [219, 276], fails to adequately support both the runners' and spectators' journey. Our results showed that runners found value in sharing their discomfort and hardships with spectators during long-distance running events, as it was considered a part of the beauty of the experience. At the same time, rather than simply looking for cheering [276], runners' needs varied at different stages of their run. In fact, cheering was often not simply what they want as it may be disrupting and even leads to injury when they are tired or concentrated on running. They often seek for a silent emotional support to be delivered during moments of desperation and self-doubt and want their experiences to be empathetically understood. To better understand runners' needs and address them, technology have to be able to capture and share the various SARE rather than simple performance metrics.

Hence, the second question addressed in this thesis was to explore the feasibility and challenges of constructing SARE Automatic Recognition Systems (SARE-ARS). In particular, we explored how SARE self-reports can be continuously gathered and mapped to the related physiological and behavioural sensors data.

RQ2: How can a ubiquitous ESM based on voice and wearables be designed to gather in-situ self-reports and running data for the purpose of creating SARE-ARS?

In order to explore the feasibility of building SARE-ARS, in Chapter 5 and Chapter 6, we utilized a mixed-method approach to investigate the continuous gathering of the SARE through voice self-reports. Additionally, we explored the feasibility of mapping these self-reports to behavioural and physiological sensor data collected during running. To achieve this, we first developed a wearable R-ESM system to be used as a research probe. Subsequently, we conducted a user study to gain insights into the design of a voice-based data collection system. . The a voice- and wearables-based ESM was implemented as a mobile app to wirelessly deliver speech-based ESM prompts every minute and record runners' voice-based self-reports (rating and Likert scale based self-reports for perceived exertion, pain, desire-to-stop, and emotional valence, supplemented with open think-alouds for running experiences understanding) via earbuds, while simultaneously collecting physiological and behavioural data that relates to runners' SARE by using wearable sensors.

Standard machine learning techniques were employed to validate the feasibility of mapping the ground truth to the physiological and behavioural data collected from the runners to build the SARE-ARS. The machine learning recognition performances well above chance level. In particular, the best model based on the combination of all sensor signals obtained a weighted average F1 of 0.79 for discriminating between three exertion levels, 0.84 for three pain levels, 0.78 for two desire-to-stop levels, and 0.85 for three emotion valence levels. This study represents the first study demonstrating how continuous SARE self-reports, collected in real-world settings, can be effectively gathered and correlated with physiological and behavioural data. Previous studies were limited to controlled laboratory studies focused solely on induced affective states (e.g., perceived exertion) during treadmill exercises [99, 128] or relied on self-reports collected at discrete laps of a circuit [165].

Furthermore, the qualitative studies shed light on the feasibility and the benefits, specifically for the runners, of utilizing continuous voice-based ESM self-reports. Post-running interviews were gathered and analysed to gain insight into runners' self-reporting experiences and inform the design of R-ESM. The R-ESM system was overall accepted by runners as a non-intrusive, easy, and relevant companion to their mental state during running. Runners

reported a more mindful running experience while self-reporting SARE to the R-ESM system. Our findings suggested that the R-ESM system could be designed as a conversational agent to facilitate mindfulness and self-reflection, provide emotional and social companionship, and offer context-adaptive prompts. For example, such a conversational system could assist runners to increase self-awareness and reflection while reporting; it could also postpone prompts during high physical or cognitive load but retrieve them during periods of ease. This is in strong contrast with typical questionnaire-based methods [92, 122], where self-reporting is often seen as a disruptive activity that interrupts the primary task of running, rather than contributing to its experience.

Such findings hence lead to the next research question:

RQ 3: How can chatbot conversation flow be designed to facilitate voice-based SARE self-reports while running?

In this part of the thesis, we fully focused on exploring the design of a companion as a way to facilitate self-reflection and self-reporting. Such approach had not been previously explored in this context. To answer this question, Chapter 7 detailed the iterative design and implementation of the chatbot prototype. The primary emphasis of this stage shifted on understanding the essential conversational components that should be integrated into a chatbot to facilitate self-reporting of SARE. The iterative design started with an initial proposal for the chatbot conversation structure and scripts informed by sports literature and running articles. This was followed by a co-design study and the final technical implementation of the chatbot prototype. The chatbot welcome runners and discusses their purpose and plan. It proactively invited runners to conversation sessions on predetermined topics, including exertion, pain, desire to stop, and emotion. The chatbot remembered user responses and provides personalized and empathetic replies. It guided users through mindful scanning exercises to increase self-awareness of their feelings.

Chapter 8 presents a user study (with in-the-wild running sessions) that contrasted the chatbot probe to the R-ESM enhanced with a focused thinking aloud between ESM prompts. The findings showed that the voice-based chatbot could create a reliable and reflective self-reporting and that foster a socially collaborative and empathetic experience, confirming the

findings in Chapter 6. Moreover, the study contributed the types of conversation structures and components that facilitate self-reflection and promoted more reliable self-reporting. These components were summarised in the novel Modality-Autonomy-Prompting (MAP) framework to drive the design of chatbots for continuous SARE-self-reflections and self-report during running.

The MAP framework suggested chatbot leveraging speaking as an interactive modality could facilitate reflection while self-reporting rather than afterward as in [128]. By capturing speech data during the heuristic think-while-speak process, the chatbot could provide additional source of self-reported SARE data to learn users' thoughts formation and reasoning process to account for the complexities of users' subjective and affective feelings. Furthermore, the MAP framework emphasized the significance of incorporating a mixture of structured and unstructured conversations. This would not only provide runners with the freedom to provide flexible and exploratory self-reports, but also generate focused and precise self-report and further result in more consolidated self-reports data with prioritized and important information on runners' SARE. This could be useful to address the issues of imprecise and uninterpretable self-reports collected via traditional open-ended questions asked by a chatbot [159].

The MAP framework introduced four types of prompts to enhance self-reflection and improve the quality, consistency, and comprehensiveness of self-reports. These prompts include context-aware proactive prompts, semi-structured prompts, retrospective prompts, and systematic and mindful scanning prompts. Context-aware proactive prompts could encourage runners to provide relevant context-related information when reporting their SARE. Semi-structured prompts could enable conversational dynamics while maintaining focus in self-reports, making it easier to convert qualitative self-report data into structured formats for computational purposes. Retrospective prompts could help runners compare their past experiences, ensuring consistent ratings of their current experiences using standardized criteria. This process could enhance the trustworthiness and consistency of self-report data. Lastly, systematic and mindful scanning could prompt facilitate the inclusion of multiple dimensions in self-reports, resulting in an unbiased and holistic self-report dataset.

As such, our MAP framework addresses a significant gap in the current literature by specifically focusing on the design of chatbot conversations to facilitate self-reflection and the collection of self-reports related to SARE in real-world running scenarios. The framework introduced innovative conversation structures and prompting components that differ from existing approaches in the literature, which predominantly relied on questionnaires for chatbot conversation design in areas such as work reflection [135] stock investment reflection [216], or affect self-reports collection [142, 159].

The MAP framework offered design guidelines for the conversation structure and prompts components of chatbots. Next, we take an additional step to gain a deeper understanding of the specific elements that can be incorporated into SARE self-reports. This understanding informs the design of chatbots aimed at facilitating the self-reporting of SARE. Hence, the next question is:

RQ 4: What are the dynamic, reflective elements of SARE self-reports, and what should SARE-ARS model?

The chatbot was revised using the MAP framework, providing a mixed-structured conversational space for runners to self-reflect and self-report their SARE. At this stage of our research, we confirmed the effect of using the MAP-guided chatbot on runners' self-reporting processes. A user study was conducted with the revised chatbot, and the analysis of runners' conversation data showed that runners appreciated both exploratory and unstructured self-reporting, as well as structured but focused self-reporting. Engaging in such mixed-structure conversations with the chatbot could lead to further learning, self-tuning, and personalization when runners self-reporting their SARE data. The prompts used in the structured conversation for self-reporting were well-received by the runners. They found them effective in bringing focus to the think-aloud reflection phase and enhancing the formation and consolidation of their SARE self-reports. This confirmed that our MAP framework guided design could not only benefit runners own experience but also allow for the identification of implicit SARE self-reports and provide accurate and well-tuned SARE self-reports.

In addition to validating the effectiveness of the MAP framework, our analysis contributed the Description-Appraisal-Management (DAM) framework. Our findings indicated that

runners underwent a reflective process in which they reflected on the details and contexts of events and feelings related to their SARE, which went beyond simple dimensional or discrete emotion approaches in capturing SARE. This involves detailed elaborations of their SARE states such as the intensity, location on the body, and duration of the feelings. The framework also emphasized the importance of incorporating appraisal-based self-report data that focused on factors like familiarity with the experience, severity, tolerance levels, affective associations, and anticipated changes as they help better characterise the state of the person and understand them from their perspectives. Moreover, management component suggested to capture the runner's action tendency and preferred strategies for managing the experience, their actual responses, and the effectiveness of those actions. These findings aligned with the view of appraisal process in emotion [226, 227], which considered appraisal as a process of multilevel sequential checking in emotional process.

Our work contributed also by critiquing and highlighting the limitations of the current approach to emotion or affect-related self-report collection to various fields. For instance, current approaches used in affective computing [92], psychology, mental health research[66], and sports psychology [65]) were deemed overly simplistic as they relied on affective dimensions or discrete affective terms. Our DAM framework proposed the appraisal-like process for collecting SARE self-reports, in which self-reports become enriched with contextual information, enabling a more nuanced analysis and facilitating clearer self-reflection. The appraisal process enhances the depth and accuracy of the self-reports, providing a more comprehensive understanding of the subjective and affective experience. While previous studies have also recognized the importance of appraisal in understanding affective states (e.g., using post-task appraisals and analysis during post-interview analysis [229]), our approach could enable individuals to appraise their SARE in real-time with a more naturalistic experience offered by the chatbot, empowering individuals to autonomously engage in the appraisal process and enhancing the accuracy and immediacy of the collected data.

The DAM framework also guides to improve the design of chatbot conversational prompts and questions to elicit runners with more structured self-reports. Moreover, the framework also sheds light on the future direction for a chatbot to improve its natural language

processing abilities to directly extract self-reports during the exploratory think-aloud process, eliminating the need for further structured conversations on already expressed self-reports.

10.2 Ethical considerations

Our research has made important contributions to the design of chatbots that aid runners in self-reflection and self-reporting of their SARE. However, it is crucial to consider ethical concerns when creating these chatbots, specifically regarding the potential risks of emotional meltdowns when users disclose vulnerable experiences, pain, or incapability [81]. As discussed in psychotherapy research, a person may develop fear or become overwhelmed by the intensity of their negative emotions during a disclosure stage, perceiving them as dangerous or bad. This might lead to emotional meltdowns and a sense of being overwhelmed by vulnerability [163].

This is a critical and general problem for chatbots used for emotion regulation and mental health [126] since even if they maybe designed to be supportive and empathic [52, 148, 171], they are not yet capable of offering the same level of professional guidance as qualified therapists in professional therapy settings. It is hence important that participants pre-screening is considered before deploying the chatbots for data collection to train affect recognition systems. It is also important to consider the impact that this may have as a support to self-coaching. Indeed, many people use running to address stress or more serious mental issues. In such cases, people should be informed of the risks and possibly discouraged from using such SARE deep self-reflection processes unless under the supervision of clinical or carer support. Such considerations should also be adopted when the system is used for SARE sharing during competition. However, in such situation, one could consider the opportunity of leveraging their friends/family spectators to provide the right support needed during the competition and afterwards. As technology for affect recognition and for natural language processing get better and better, future research could not just acknowledge but also explore the possibility to address the potential risks of triggering users' vulnerability and explore protocols for professional support throughout the interaction process to possibly deliver new types of intervention.

Another ethical issue derives from the demand that our chatbot conversation may impose on the person. For instance, when runners experience high exertion, pain, or a heightened heart rate, which often hinders their ability to articulate thoughts effectively, it is crucial for the chatbot to adapt accordingly. The feedback provided by the chatbot should be designed thoughtfully to avoid adding to the existing physical, mental or cognitive load of the runner. The chatbot may have to give brief and concise responses that minimize the amount of information conveyed. In such situations, it can be beneficial for the chatbot to employ a self-paced and open-structured think-aloud approach, allowing the runner to share their feelings at their own pace and providing them with flexibility in expressing their experiences without the pressure of adhering to a specific time frame. In addition, participants must be pre-screened to ensure they are comfortable talking while running, and the chatbot must respect users' right to choose when to engage and disengage from speaking.

It is also important to be aware of issues regarding data privacy [147] and data breaches risks [177]. The chatbot must be designed to ensure that runners' emotional, physical, and affective states are not uploaded to the cloud, and their information is not shared without their consent.

The chatbot should be designed without bias, respecting diversity and not discriminating against runners based on experience level. For instance, the chatbot trained on data from experienced runners should not pose discriminative questions when interacting with beginner runners. Transparency in the formation of dialogue and access to conversation content is crucial for users' trust in chatbots [177].

The chatbot must be designed with safety concerns and prevention considerations. It should not be used when runners are in public places that are unsafe, and it should never encourage users to push through physical boundaries that could result in injuries. Moreover, the chatbot must not embarrass users in public contexts or expose them to unnecessary risks.

Overall, the chatbot's design should prioritize the well-being, privacy, ethics and safety of its users when being used to collect SARE related data or affect data in other use cases, ensuring that the interactions are positive and not harmful. As technology advances, the future development and adaptation of more powerful AI generative models for chatbots in

emotional data collection must be regulated and controlled to prevent harm and ensure compliance with ethical guidelines. It is crucial to take an “*ethical by design*” [81] adapting and refining the chatbot’s design at every design stage to adhere to evolving ethical guidelines in the field of HCI.

10.3 Limitations and future directions

Although this thesis mainly focused on the design of a chatbot to facilitate the collection of SARE, it sheds light on one important research direction regarding examining the chatbot collected SARE self-reports for use in the computational process of building ARS. A recent work presented in [229], which proposes to map the multiple sources of qualitative information about the emotion as ground truth in a Bayesian network, could be direction to explore. The qualitative results in this thesis could help extend such proposal regarding how emotion ground truth labels can be enriched.

From technical perspective, this thesis only used simple ML techniques in Chapter 5, and future work could investigate more advanced methods, including different feature selections, windowing techniques, ML models, and validation approaches. The chatbot used in this thesis had limitations in its natural language understanding ability and speech recognition performance, which were subject to noise. The generation of chatbot responses was limited and dependent on researchers' predesigned prompts. To address these limitations, future studies could incorporate improved speech recognition APIs, such as Whisper²⁵, and more advanced natural language understanding and generation models, such as GPT-4²⁶. Privacy and ethics should be considered when using such APIs.

Future work could also explore how the chatbot can better recognize the context and users' state in multiple modalities, such as physical and behavioural cues (e.g., high heart rate, abnormal heart rate, wrong posture), to prompt personalized conversations accordingly.

²⁵ <https://openai.com/research/whisper>

²⁶ <https://openai.com/research/gpt-4>

Moreover, the self-report data collected through the chatbot in this thesis was limited in scale. To improve the DAM framework's accuracy, future work could focus on enhancing data collection methods to obtain larger datasets. Such larger datasets could also help to train the Chatbot with improved natural language processing capability.

This research did not conduct longitudinal studies to investigate how runners' experiences and needs with the chatbot for reflection self-reporting may change over time. Understanding these long-term changes via longitudinal studies is crucial in informing the design of the chatbot for adoption in real-life scenarios and long-term usage. Future research could also explore and evaluate the validity of the proposed measurements (such as Fallback Rate, Fallback Success, Conversation Smoothness, and Conversation Usefulness as introduced in Section 8.3) for chatbots' conversation interactions.

Additionally, the results of this thesis suggest individual differences in dialogical preferences, reflection ability, and runners' experience levels and goals, which may influence their preferences regarding what to reflect on. Thinking aloud as a data collection approach might also have altered runners' original running behaviours, which could be captured through the wearable sensors. Some more exploration of possible effects should be considered in future research.

10.4 Conclusion

To conclude, this thesis presented a series of mixed-method research studies that identified the necessity of capturing Subjective and Affective Running Experience (SARE) and underscored the significance of runners' perceived exertion, pain, desire-to-stop, and emotional valence in capturing and sharing their experiences. Subsequently, the thesis investigated the effectiveness of using the developed R-ESM system to collect SARE self-reports in real-world settings, aiming to construct SARE Automatic Recognition Systems (SARE-ARS). The user studies conducted suggested that the R-ESM system could be further developed into a chatbot, providing social and emotional companionship during the self-reporting process and promoting mindfulness to facilitate self-reflection. The thesis then showcased the design process of the chatbot and presented valuable findings from our user

studies, which have contributed to the development of both the MAP framework and DAM framework.

The thesis proposed that a chatbot leveraging different structures recommended by the MAP and DAM frameworks could facilitate an appraisal-like process while enabling self-reports of SARE. This approach yielded more detailed and nuanced self-report data, leading to a comprehensive understanding of runners' subjective and affective experiences. Our frameworks contributed to the design of a chatbot that utilized conversations to overcome the limitations present in current approaches used in sports psychology and in the affective computing community, which primarily focused on single dimensions or discrete emotions that only captured static measures of affect, lacking the capability to capture the dynamic and multifaceted nature of affective experiences. Building on this thesis, future research should explore how the self-reports enhanced with the MAP and DAM framework can be leveraged in machine learning for the development of SARE-ARS.

Bibliography

1. *E4 User Manual*. [cited 2023 1 Jan]; Available from: <https://www.empatica.com/en-gb/manuals/>.
2. *Mens Marathon RIO 2016 Replay*. [cited 2018 1 June]; Available from: <https://www.olympicchannel.com/en/playback/mens-marathon-rio-2016-replay/>.
3. *Olympic Channel goes live with 2018 Boston Marathon stream*. [cited 2019 3 Sep]; Available from: <https://www.svgeurope.org/blog/news-roundup/olympic-channel-goes-live-with-2018-boston-marathon-stream/>.
4. *Perceived Exertion (Borg Rating of Perceived Exertion Scale)*. [cited 2020 10 Jan]; Available from: <https://www.cdc.gov/physicalactivity/basics/measuring/exertion.htm>.
5. *ReTiSense Stridalyzer Insight*. [cited 2018 1 Nov]; Available from: <https://www.retisense.com/product/stridalyzer-insight-sensor-insoles/>.
6. *scikit-learn*. [cited 2022 10 Aug]; Available from: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning.
7. *sklearn.cluster*. [cited 2023 1 April]; Available from: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>.
8. *sklearn.preprocessing.MinMaxScaler*. [cited 2023 1 April]; Available from: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>.
9. *Virgin Money London Marathon App*. 2018 [cited 2018 Dec]; Available from: <https://www.virginmoneylondonmarathon.com/en-gb/news-media/latest-news/item/download-the-virgin-money-london-marathon-app-2/>.
10. Abdul-Kader, S.A. and J. Woods, *Survey on chatbot design techniques in speech conversation systems*. International Journal of Advanced Computer Science and Applications, 2015. **6**(7).
11. Allen-Collinson, J., *'Weather work': embodiment and weather learning in a national outdoor exercise programme*. Qualitative Research in Sport, Exercise and Health, 2018. **10**(1): p. 63-74.
12. Allen-Collinson, J. and J. Hockey, *Feeling the way: Notes toward a haptic phenomenology of distance running and scuba diving*. International review for the sociology of sport, 2011. **46**(3): p. 330-345.

13. Allen-Collinson, J. and J. Hockey, *From a certain point of view: sensory phenomenological envisionings of running space and place*. Journal of Contemporary Ethnography, 2015. **44**(1): p. 63-83.
14. Allen-Collinson, J. and A. Leledaki, *Sensing the outdoors: a visual and haptic phenomenology of outdoor exercise embodiment*. Leisure Studies, 2015. **34**(4): p. 457-470.
15. Allen-Collinson, J. and H. Owton, *Take a deep breath: Asthma, sporting embodiment, the senses and 'auditory work'*. International review for the sociology of sport, 2014. **49**(5): p. 592-608.
16. Allen-Collinson, J., et al., *Exploring Lived Heat, "Temperature Work," and Embodiment: Novel Auto/Ethnographic Insights from Physical Cultures*. Journal of contemporary ethnography, 2018. **47**(3): p. 283-305.
17. Almzayyen, A., et al., *Voice-Based Conversational Agents for self-reporting fluid consumption and sleep quality*. arXiv preprint arXiv:2202.02186, 2022.
18. Alohal, A., K. Kunze, and R. Earle, *Run with me-Designing Storytelling Tools for Runners*. Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16, 2016: p. 5-8.
19. Altmann, J., *Observational study of behavior: sampling methods*. Behaviour, 1974. **49**(3-4): p. 227-266.
20. Aminuddin, R., et al. *Framework for a mobile application with a chatbot to self-report injuries and carry out contact tracing for athletes and sports trainers*. in *2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*. 2021. IEEE.
21. Anbarian, M., H.J.G. Esmaili, and posture, *Effects of running-induced fatigue on plantar pressure distribution in novice runners with different foot types*. 2016. **48**: p. 52-56.
22. Antonio, N., et al., *Fatigue effects on body balance*. *Electroencephalography and Clin. Neurophysiology*, 1997. **105**: p. 309-320.
23. Aragonés, D., et al., *Fluctuating dynamics of perceived exertion in constant-power exercise*. Psychology of Sport and Exercise, 2013. **14**(6): p. 796-803.
24. Ariel, N., *Don't think before you speak: on the gradual formation of thoughts during speech*. Pedagogy, Culture & Society, 2022: p. 1-13.
25. Audacity. *Audacity*. [cited 2022 13 Sep]; Available from: <https://www.audacityteam.org>.
26. Audacity. *Noise Reduction in Audacity*. [cited 2022 13 Sep]; Available from: https://manual.audacityteam.org/man/noise_reduction.html.

27. Aung, M.S., et al., *The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal EmoPain dataset*. IEEE transactions on affective computing, 2015. **7**(4): p. 435-451.
28. Baer, R.A., et al., *Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples*. Assessment, 2008. **15**(3): p. 329-342.
29. Balagué, N., et al., *Dynamics of perceived exertion in constant-power cycling: time-and workload-dependent thresholds*. Research Quarterly for Exercise and Sport, 2015. **86**(4): p. 371-378.
30. Barrett, L.F., *Feelings or words? Understanding the content in self-report ratings of experienced emotion*. Journal of personality and social psychology, 2004. **87**(2): p. 266.
31. Barrett, L.F., *Psychological construction: The Darwinian approach to the science of emotion*. Emotion review, 2013. **5**(4): p. 379-389.
32. Barrett, L.F., *How emotions are made: The secret life of the brain*. 2017: Pan Macmillan.
33. Barrett, L.F. and J.A. Russell, *The psychological construction of emotion*. 2014: Guilford Publications.
34. BBC, *London Marathon Live Coverage*.
35. Berger, B.G. and R.W. Motl, *Exercise and mood: A selective review and synthesis of research employing the profile of mood states*. Journal of applied sport psychology, 2000. **12**(1): p. 69-92.
36. Bi, T., *Supporting Sharing Experience between Runners and Spectators during Long-distance Running Event*, in *UCL Interaction Centre*. 2017, University College London.
37. Bianchi-Berthouze, N., *Understanding the role of body movement in player engagement*. Human-Computer Interaction, 2013. **28**(1): p. 40-75.
38. Bicknell, K., *Sport, Entertainment and the Live(d) Experience of Cheering*. Popular Entertainment Studies, 2011. **2**(1): p. 96-111.
39. Bishop, S.R., et al., *Mindfulness: A proposed operational definition*. Clinical psychology: Science and practice, 2004. **11**(3): p. 230.
40. Borg, G., *Borg's perceived exertion and pain scales*. 1998: Human kinetics.
41. Borg, G.A., *Perceived exertion: a note on "history" and methods*. Medicine and science in sports, 1973. **5**(2): p. 90-93.
42. Bradley, M.M. and P.J. Lang, *Measuring emotion: the self-assessment manikin and the semantic differential*. Journal of behavior therapy and experimental psychiatry, 1994. **25**(1): p. 49-59.

43. Braun, V. and V. Clarke, *Using thematic analysis in psychology*. Qualitative research in psychology, 2006. **3**(2).
44. Braun, V., V. Clarke, and P. Weate, *Using thematic analysis in sport and exercise research*. Routledge handbook of qualitative research in sport and exercise, 2016: p. 191-205.
45. Brown, K.W. and R.M. Ryan, *The benefits of being present: mindfulness and its role in psychological well-being*. Journal of personality and social psychology, 2003. **84**(4): p. 822.
46. Buckley, C., et al. *Binary classification of running fatigue using a single inertial measurement unit*. in *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. 2017. IEEE.
47. Cambria, E., *Affective computing and sentiment analysis*. IEEE intelligent systems, 2016. **31**(2): p. 102-107.
48. Cameron, J.E., *Self and Identity A Three-Factor Model of Social Identity A Three-Factor Model of Social Identity*. 2017.
49. Chang, H.-C., et al. *Combining Participatory and ESM: A Hybrid Approach to Collecting Annotated Mobility Data*. in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020.
50. Chang, W.H., et al., *Mass gathering emergency medicine: a review of the taiwan experience of long-distance swimming across sun-moon lake*. International Journal of Gerontology, 2010. **4**(2): p. 53-68.
51. Chatfield, S.L., *Recommendations for secondary analysis of qualitative data*. The Qualitative Report, 2020. **25**(3): p. 833-842.
52. Chaves, A.P. and M.A. Gerosa, *How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design*. International Journal of Human–Computer Interaction, 2021. **37**(8): p. 729-758.
53. Cheron, G., et al., *Brain oscillations in sport: toward EEG biomarkers of performance*. Frontiers in psychology, 2016. **7**: p. 246.
54. Cho, Y., N. Bianchi-Berthouze, and S.J. Julier. *DeepBreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings*. in *2017 Seventh international conference on affective computing and intelligent interaction (acii)*. 2017. IEEE.
55. Cho, Y., et al. *Deep Thermal Imaging: Proximate Material Type Recognition in the Wild through Deep Learning of Spatial Surface Temperature Patterns*. in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018. ACM.

56. Choe, E.K., et al. *Understanding self-reflection: how people reflect on personal data through visual data exploration*. in *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 2017.
57. Cloud, G. **Text-to-Speech client libraries**. 2022; Available from: <https://cloud.google.com/text-to-speech/docs/libraries>.
58. Consolvo, S. and M. Walker, *Using the experience sampling method to evaluate ubicomp applications*. IEEE pervasive computing, 2003. **2**(2): p. 24-31.
59. Constantine, L. and H. Hajj. *A survey of ground-truth in emotion data annotation*. in *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. 2012. IEEE.
60. Cowie, R., et al., *Issues in Data Labelling*, in *Emotion-Oriented Systems: The Humaine Handbook*, R. Cowie, C. Pelachaud, and P. Petta, Editors. 2011, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 213-241.
61. Cruse, D., et al., *Bedside detection of awareness in the vegetative state: a cohort study*. The Lancet, 2011. **378**(9809): p. 2088-2094.
62. Curmi, F., et al., *HeartLink: open broadcast of live biometric data to social networks*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13, 2013: p. 1749-1749.
63. Curmi, F., et al. *HeartLink: open broadcast of live biometric data to social networks*. in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 2013.
64. Curmi, F., et al., *Crowdsourcing Synchronous Spectator Support: (go on, go on, you're the best)ⁿ⁻¹*. Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems, 2015. **1**: p. 757-766.
65. Dickens, Y.L., J. Van Raalte, and R.T. Hurlburt, *On investigating self-talk: A descriptive experience sampling study of inner experience during golf performance*. The Sport Psychologist, 2018. **32**(1): p. 66-73.
66. Doherty, K., et al., *A mobile app for the self-report of psychological well-being during pregnancy (BrightSelf): qualitative design study*. JMIR mental health, 2018. **5**(4): p. e10007.
67. Dong, Y., *How chatbot conversation should be designed to facilitate introspection in long distance running experience*, in *UCL Interaction Centre*. 2022, University College London. p. 37.
68. Doppelmayr, M., et al., *Changes in EEG during ultralong running*. Journal of Human Performance in Extreme Environments, 2012. **10**(1): p. 4.

69. Doyle, J.P., et al., *Exploring PERMA in spectator sport: Applying positive psychology to examine the individual-level benefits of sport consumption*. Sport Management Review, 2016. **19**(5): p. 506-519.
70. Draxler, F., et al. *Flexibility and Social Disconnectedness: Assessing University Students' Well-Being Using an Experience Sampling Chatbot and Surveys Over Two Years of COVID-19*. in *Designing Interactive Systems Conference*. 2022.
71. Ekkekakis, P., *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. 2013: Cambridge University Press.
72. Ekman, P., *Basic emotions*. Handbook of cognition and emotion, 1999. **98**(45-60): p. 16.
73. Enders, H., et al., *Changes in cortical activity measured with EEG during a high-intensity cycling exercise*. Journal of neurophysiology, 2015. **115**(1): p. 379-388.
74. Ericsson, K.A. and H.A. Simon, *Verbal reports as data*. Psychological review, 1980. **87**(3): p. 215.
75. Ericsson, K.A. and H.A. Simon, *Verbal reports on thinking*. 1987.
76. Eslambolchilar, P., M. Bødker, and A. Chamberlain, *Ways of walking: understanding walking's implications for the design of handheld technology via a humanistic ethnographic approach*. Human Technology, 2016. **12**.
77. Exler, A., *Investigating the Perceptibility of Smartphone Notifications and Methods for Context-Aware Data Assessment in Experience Sampling Studies*. 2019, Karlsruhe Institute of Technology, Germany.
78. Fedosov, A., et al. *Towards understanding digital sharing practices in outdoor sports*. in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 2016.
79. Fischer, J.E., et al. *Effects of content and time of delivery on receptivity to mobile interruptions*. in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*. 2010.
80. Floridi, L. and M. Chiriatti, *GPT-3: Its nature, scope, limits, and consequences*. Minds and Machines, 2020. **30**: p. 681-694.
81. Følstad, A., et al., *Future directions for chatbot research: an interdisciplinary research agenda*. Computing, 2021. **103**(12): p. 2915-2942.
82. Fontaine, J.R., et al., *The world of emotions is not two-dimensional*. Psychological science, 2007. **18**(12): p. 1050-1057.

83. Fox, M.C., K.A. Ericsson, and R. Best, *Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods*. Psychological bulletin, 2011. **137**(2): p. 316.
84. Frijda, N.H., P. Kuipers, and E. Ter Schure, *Relations among emotion, appraisal, and emotional action readiness*. Journal of personality and social psychology, 1989. **57**(2): p. 212.
85. Frommel, J. and R.L. Mandryk, *Modeling Behaviour to Predict User State: Self-Reports as Ground Truth*. arXiv preprint arXiv:2007.14461, 2020.
86. Fuentes, C., et al., *A systematic literature review about technologies for self-reporting emotional information*. Journal of Ambient Intelligence and Humanized Computing, 2017. **8**(4): p. 593-606.
87. Ganganwar, V., *An overview of classification algorithms for imbalanced datasets*. International Journal of Emerging Technology and Advanced Engineering, 2012. **2**(4): p. 42-47.
88. Gau, L.S., *Development of a model connecting self-directive value and satisfaction of sociability needs in sport spectators*. Social Behavior and Personality, 2013. **41**(5): p. 795-804.
89. Gebhard, A., et al. *Towards Heart Rate Categorisation from Speech in Outdoor Running Conditions*. in *2022 E-Health and Bioengineering Conference (EHB)*. 2022. IEEE.
90. Gebhard, A., et al. *Improving Exertion and Wellbeing Prediction in Outdoor Running Conditions using Audio-based Surface Recognition*. in *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*. 2022.
91. Ghosh, S., et al., *Designing an experience sampling method for smartphone based emotion detection*. IEEE Transactions on Affective Computing, 2019.
92. Ghosh, S., et al., *Designing an experience sampling method for smartphone based emotion detection*. IEEE Transactions on Affective Computing, 2019. **12**(4): p. 913-927.
93. Ghosh, S., B. Mitra, and P. De. *Towards Improving Emotion Self-report Collection using Self-reflection*. in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020.
94. Giulianotti, R., *Supporters, Followers, Fans, and Flaneurs*. Journal of Sport and Social Issues, 2002. **26**(1): p. 25-46.
95. Gould, D., C. Greenleaf, and V. Krane, *Arousal-anxiety and sport behavior*. 2002.
96. Graether, E. and F. Mueller, *Joggobot: a flying robot as jogging companion*, in *CHI '12 Extended Abstracts on Human Factors in Computing Systems*. 2012, ACM: Austin, Texas, USA. p. 1063-1066.

97. Groth, S. and Y. Krahn, *Sensing athletes: Sensory dimensions of recreational endurance sports*. Journal of Ethnology and Folkloristics, 2017. **11**(2): p. 3-23.
98. Guérin, E., M.S. Fortier, and S.N. Sweet, *An experience sampling study of physical activity and positive affect: investigating the role of situational motivation and perceived intensity across time*. Health psychology research, 2013. **1**(2).
99. Halkiadakis, Y., H.M. Alzakerin, and K.D. Morgan. *Classification Model for Discriminating Trunk Fatigue During Running*. in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2021. IEEE.
100. Harada, S., et al. *VoiceLabel: using speech to label mobile sensor data*. in *Proceedings of the 10th international conference on Multimodal interfaces*. 2008.
101. Hassan, M., et al., *Footstriker: An EMS-based foot strike assistant for running*. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2017. **1**(1): p. 1-18.
102. Hernandez, J., et al., *Wearable ESM: differences in the experience sampling method across wearable devices*, in *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 2016, Association for Computing Machinery: Florence, Italy. p. 195–205.
103. Hertzum, M. and K.D. Holmegaard, *Thinking aloud influences perceived time*. Human Factors, 2015. **57**(1): p. 101-109.
104. Hockey, J., *Sensing the run: The senses and distance running*. The Senses and Society, 2006. **1**(2): p. 183-201.
105. Hockey, J., *Knowing the 'going': The sensory evaluation of distance running*. Qualitative research in sport, exercise and health, 2013. **5**(1): p. 127-141.
106. Iekura, M.-S., et al., *SMASH: synchronization media of athletes and spectator through haptic*, in *SIGGRAPH Asia 2015 Mobile Graphics and Interactive Applications*. 2015, ACM: Kobe, Japan. p. 1-2.
107. Ingold, T., *Culture on the ground: the world perceived through the feet*. Journal of Material Culture, 2004. **9**(3): p. 315-340.
108. Ingold, T. and J.L. Vergunst, *Ways of walking: Ethnography and practice on foot*. 2008: Ashgate Publishing, Ltd.
109. Inkster, B., S. Sarda, and V. Subramanian, *An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study*. JMIR mHealth and uHealth, 2018. **6**(11): p. e12106.
110. Intille, S., et al., *μEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch*, in *Proceedings of the 2016 ACM International Joint*

- Conference on Pervasive and Ubiquitous Computing*. 2016, Association for Computing Machinery: Heidelberg, Germany. p. 1124–1128.
111. Jack, R.E., et al., *Four not six: Revealing culturally common facial expressions of emotion*. *Journal of Experimental Psychology: General*, 2016. **145**(6): p. 708.
 112. Jacucci, G., et al. *Comedia: mobile group media for active spectatorship*. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2007. ACM.
 113. Jacucci, G., A. Oulasvirta, and A. Salovaara, *Active construction of experience through mobile media: a field study with implications for recording and sharing*. *Personal and Ubiquitous Computing*, 2007. **11**(4): p. 215-234.
 114. Jacucci, G., et al. *Supporting the shared experience of spectators through mobile group media*. in *Proceedings of the 2005 ACM International Conference on Supporting Group Work*. 2005.
 115. Jakob, C., et al. *Estimation of the knee flexion-extension angle during dynamic sport motions using body-worn inertial sensors*. in *Proceedings of the 8th International Conference on Body Area Networks*. 2013. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).
 116. John, N.A., *The social logics of sharing*. *The Communication Review*, 2013. **16**(3): p. 113-131.
 117. Kabat-Zinn, J., *Mindfulness-based interventions in context: past, present, and future*. 2003.
 118. Kächele, M., et al., *Adaptive confidence learning for the personalization of pain intensity estimation systems*. *Evolving Systems*, 2017. **8**(1): p. 71-83.
 119. Kächele, M., et al., *Methods for person-centered continuous pain intensity assessment from bio-physiological channels*. *IEEE Journal of Selected Topics in Signal Processing*, 2016. **10**(5): p. 854-864.
 120. Kächele, M., et al. *Multimodal data fusion for person-independent, continuous estimation of pain intensity*. in *International Conference on Engineering Applications of Neural Networks*. 2015. Springer.
 121. Kächele, M., et al. *Bio-visual fusion for person-independent recognition of pain intensity*. in *International Workshop on Multiple Classifier Systems*. 2015. Springer.
 122. Kang, S., et al. *Understanding Emotion Changes in Mobile Experience Sampling*. in *CHI Conference on Human Factors in Computing Systems*. 2022.
 123. Karg, M., et al., *Body movements for affective expression: A survey of automatic recognition and generation*. *IEEE Transactions on Affective Computing*, 2013. **4**(4): p. 341-359.

124. Karkar, R., et al. *Tummytrials: a feasibility study of using self-experimentation to detect individualized food triggers*. in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017.
125. Kathan, A., et al. *Investigating Individual-and Group-Level Model Adaptation for Self-Reported Runner Exertion Prediction from Biomechanics*. in *2022 E-Health and Bioengineering Conference (EHB)*. 2022. IEEE.
126. Kawasaki, M., et al. *Assessing users' mental status from their journaling behavior through chatbots*. in *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 2020.
127. Kawsar, F., et al. *eSense: Open Earable Platform for Human Sensing*. in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 2018. ACM.
128. Khan, T., et al., *A novel method for classification of running fatigue using change-point segmentation*. *Sensors*, 2019. **19**(21): p. 4729.
129. Kim, S., J. Lee, and G. Gweon. *Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality*. in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019.
130. Kim, S., J. Mankoff, and E. Paulos. *Sensr: evaluating a flexible framework for authoring mobile data-collection tools for citizen science*. in *Proceedings of the 2013 conference on Computer supported cooperative work*. 2013.
131. Kleinsmith, A. and N. Bianchi-Berthouze. *Recognizing affective dimensions from body posture*. in *International conference on affective computing and intelligent interaction*. 2007. Springer.
132. Kleinsmith, A. and N. Bianchi-Berthouze, *Affective body expression perception and recognition: A survey*. *IEEE Transactions on Affective Computing*, 2013. **4**(1): p. 15-33.
133. Knaving, K. and P. Wozniak. *Ticktockrun: Towards enhancing communication in runner families*. in *Proceedings of the 19th acm conference on computer supported cooperative work and social computing companion*. 2016.
134. Knaving, K., et al. *Flow is not enough: Understanding the needs of advanced amateur runners to design motivation technology*. in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015.
135. Kocielnik, R., et al. *Designing for workplace reflection: a chat and voice-based conversational agent*. in *Proceedings of the 2018 Designing Interactive Systems Conference*. 2018.
136. Kocielnik, R., et al., *Reflection companion: A conversational system for engaging users in reflection on physical activity*. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018. **2**(2): p. 1-26.

137. Kowatsch, T., et al., *The impact of interpersonal closeness cues in text-based healthcare chatbots on attachment bond and the desire to continue interacting: an experimental design*. 2018.
138. Kowatsch, T., et al., *Text-based healthcare chatbots supporting patient and health professional teams: preliminary results of a randomized controlled trial on childhood obesity*. 2017.
139. Kremer, J.M., et al., *Anxiety and arousal*, in *The SAGE key Concepts: Key concepts in sport psychology*. 2012, SAGE. p. 27-31.
140. Kruger, M. and M. Saayman, *Creating a memorable spectator experience at the Two Oceans Marathon*. *Journal of Sport & Tourism*, 2012. **17**(1): p. 63-77.
141. Kuru, A., *Exploring experience of runners with sports tracking technology*. *International Journal of Human-Computer Interaction*, 2016. **32**(11): p. 847-860.
142. Kwakkel, D., *Measuring well-being with a conversational agent*. 2022, University of Twente.
143. LaCaille, R.A., K.S. Masters, and E.M. Heath, *Effects of cognitive strategy and exercise setting on running performance, perceived exertion, affect, and satisfaction*. *Psychology of Sport and Exercise*, 2004. **5**(4): p. 461-476.
144. Lane, A.M. and M. Wilson, *Emotions and trait emotional intelligence among ultra-endurance runners*. *Journal of Science and Medicine in Sport*, 2011. **14**(4): p. 358-362.
145. Law, E.L.-C., A. Følstad, and N. Van As. *Effects of Humanlikeness and Conversational Breakdown on Trust in Chatbots for Customer Service*. in *Nordic Human-Computer Interaction Conference*. 2022.
146. Le Bris, R., et al., *Effect of fatigue on stride pattern continuously measured by an accelerometric gait recorder in middle distance runners*. *Journal of sports medicine and physical fitness*, 2006. **46**(2): p. 227.
147. Lee, Y.-C., N. Yamashita, and Y. Huang, *Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional*. *Proceedings of the ACM on Human-Computer Interaction*, 2020. **4**(CSCW1): p. 1-27.
148. Lee, Y.-C., et al. *" I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot*. in *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.
149. Leunes, A. and J. Burger, *Profile of Mood States Research in Sport and Exercise Psychology: Past, Present, and Future*. *Journal of Applied Sport Psychology*, 2000. **12**(1): p. 5-15.

150. Levenson, R.W., *The autonomic nervous system and emotion*. Emotion review, 2014. **6**(2): p. 100-112.
151. Li, I., A.K. Dey, and J. Forlizzi. *Understanding my data, myself: supporting self-reflection with ubicomp technologies*. in *Proceedings of the 13th international conference on Ubiquitous computing*. 2011.
152. Liao, Q.V., et al. *All Work and No Play?* in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018.
153. Lopatovska, I. and H. Oropeza, *User interactions with "Alexa" in public academic space*. Proceedings of the Association for Information Science and Technology, 2018. **55**(1): p. 309-318.
154. López, V., et al., *An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics*. Information sciences, 2013. **250**: p. 113-141.
155. Lu, Z., et al., *Adaptive adjustment strategy for walking characteristics of single-legged exoskeleton robots*. Machines, 2022. **10**(2): p. 134.
156. Ludvigsen, M. and R. Veerasawmy. *Designing technology for active spectator experiences at sporting events*. in *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*. 2010. ACM.
157. Maas, E., et al., *Novice runners show greater changes in kinematics with fatigue compared with competitive runners*. Sports Biomechanics, 2018. **17**(3): p. 350-360.
158. Maharjan, R., et al., *Experiences of a Speech-enabled Conversational Agent for the Self-report of Well-being among People Living with Affective Disorders: An In-the-Wild Study*. ACM Transactions on Interactive Intelligent Systems (TiiS), 2022. **12**(2): p. 1-29.
159. Maharjan, R., et al. *Can we talk? design implications for the questionnaire-driven self-report of health and wellbeing via conversational agent*. in *Proceedings of the 3rd Conference on Conversational User Interfaces*. 2021.
160. Maharjan, R., et al., *What Is the Difference? Investigating the Self-Report of Wellbeing via Conversational Agent and Web App*. IEEE Pervasive Computing, 2022. **21**(2): p. 60-68.
161. Makantasis, K., et al., *The Invariant Ground Truth of Affect*. arXiv preprint arXiv:2210.07630, 2022.
162. Mann, R., et al., *The effect of shoe type and fatigue on strike index and spatiotemporal parameters of running*. 2015. **42**(1): p. 91-95.

163. Markowitz, J.C. and B.L. Milrod, *The importance of responding to negative affect in psychotherapies*. American Journal of Psychiatry, 2011. **168**(2): p. 124-128.
164. Marlatt, G.A. and J.L. Kristeller, *Mindfulness and meditation*. 1999.
165. Marotta, L., et al., *Towards machine learning-based detection of running-induced fatigue in real-world scenarios: Evaluation of IMU sensor configurations to reduce intrusiveness*. Sensors, 2021. **21**(10): p. 3451.
166. Martinez, H.P., G.N. Yannakakis, and J. Hallam, *Don't classify ratings of affect; rank them!* IEEE transactions on affective computing, 2014. **5**(3): p. 314-326.
167. Masters, R. and J. Maxwell, *The theory of reinvestment*. International Review of Sport and Exercise Psychology, 2008. **1**(2): p. 160-183.
168. Matthews, G., D.M. Jones, and A.G. Chamberlain, *Refining the measurement of mood: The UWIST mood adjective checklist*. British journal of psychology, 1990. **81**(1): p. 17-42.
169. Mauriello, M., M. Gubbels, and J.E. Froehlich. *Social fabric fitness: the design and evaluation of wearable E-textile displays to support group running*. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2014.
170. McNair, D., M. Lorr, and L. Droppleman, *Profile of mood state manual*. San Diego (CA): Educational and Industrial Testing Service, 1971.
171. Medeiros, L., T. Bosse, and C. Gerritsen, *Can a chatbot comfort humans? studying the impact of a supportive chatbot on users' self-perceived stress*. IEEE Transactions on Human-Machine Systems, 2021. **52**(3): p. 343-353.
172. Mehrabian, A., *Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament*. Current Psychology, 1996. **14**: p. 261-292.
173. Mehrotra, A., et al. *Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling*. in *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 2015.
174. MEMORY, T.H. *Episodic and Semantic Memory*. 2020 [cited 2023 6 March]; Available from: <https://human-memory.net/episodic-semantic-memory/>.
175. Mendiburu, B., *3D movie making: stereoscopic digital cinema from script to screen*. 2012: Focal press.
176. Mikos, V., et al. *Optimal window lengths, features and subsets thereof for freezing of gait classification*. in *2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*. 2017. IEEE.

177. Moilanen, J., et al., *Supporting mental health self-care discovery through a chatbot*. Frontiers in Digital Health, 2023. **5**.
178. Moors, A., et al., *Appraisal theories of emotion: State of the art and future development*. Emotion Review, 2013. **5**(2): p. 119-124.
179. Morgan, W.P., *Hypnosis and muscular performance*. Ergogenic aids and muscular performance, 1972: p. 193-233.
180. Morgan, W.P., *Psychological factors influencing perceived exertion*. Medicine and science in sports, 1973. **5**(2): p. 97-103.
181. Morgan, W.P., et al., *Hypnotic perturbation of perceived exertion: ventilatory consequences*. American Journal of Clinical Hypnosis, 1976. **18**(3): p. 182-190.
182. Morgan, W.P., et al., *Facilitation of physical performance by means of a cognitive strategy*. Cognitive Therapy and Research, 1983. **7**(3): p. 251-264.
183. Morgan, W.P. and M.L. Pollock, *Psychologic characterization of the elite distance runner*. Annals of the New York Academy of Sciences, 1977. **301**(1): p. 382-403.
184. Morgan, W.P., et al., *Perceptual and metabolic responsivity to standard bicycle ergometry following various hypnotic suggestions*. International Journal of clinical and experimental hypnosis, 1973. **21**(2): p. 86-101.
185. Mueller, F.F., S. O'Brien, and A. Thorogood, *Jogging over a Distance – Supporting a “Jogging Together” Experience Although Being Apart*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts(CHI EA), 2007: p. 1989-1994.
186. Nardone, A., et al., *Fatigue effects on body balance*. Electroencephalography and Clinical Neurophysiology/Electromyography and Motor Control, 1997. **105**(4): p. 309-320.
187. Nasukawa, T. and J. Yi. *Sentiment analysis: Capturing favorability using natural language processing*. in *Proceedings of the 2nd international conference on Knowledge capture*. 2003.
188. Newland, B.L. and T.J. Aicher, *Exploring sport participants’ event and destination choices*. Journal of Sport & Tourism, 2018. **22**(2): p. 131-149.
189. Niculescu, A., et al., *Making social robots more attractive: the effects of voice pitch, humor and empathy*. International journal of social robotics, 2013. **5**: p. 171-191.
190. Nideffer, R.M., *Test of attentional and interpersonal style*. Journal of personality and social psychology, 1976. **34**(3): p. 394.

191. Nideffer, R.M. and M.-S. Sagal, *Concentration and attention control training*. Applied sport psychology: Personal growth to peak performance, 1993. **2**: p. 243-61.
192. Nietfeld, J., *An examination of metacognitive strategy use and monitoring skills by competitive middle distance runners*. Journal of Applied Sport Psychology, 2003. **15**(4): p. 307-320.
193. Nijholt, A., D.P.-O. Bos, and B. Reuderink, *Turning shortcomings into challenges: Brain–computer interfaces for games*. Entertainment computing, 2009. **1**(2): p. 85-94.
194. Nylander, S., M. Jacobsson, and J. Tholander, *Runright: real-time visual and audio feedback on running*, in *CHI '14 Extended Abstracts on Human Factors in Computing Systems*. 2014, ACM: Toronto, Ontario, Canada. p. 583-586.
195. O'connor, P., et al., *Selected Psychoendocrine Responses To Overtraining*. Medicine and Science in Sports and Exercise, 1980. **21**(2): p. S50.
196. Ojala, J., *Personal content in online sports communities: **motivations to capture and share personal exercise data***. International Journal of Social and Humanistic Computing, 2013. **2**(1-2): p. 68-85.
197. Oliver, A., P.J. McCarthy, and L. Burns, *Using a “Think Aloud” protocol to understand meta-attention in club-level golfers*. International Journal of Sport and Exercise Psychology, 2020: p. 1-14.
198. Olugbade, T., et al., *Human Observer and Automatic Assessment of Movement Related Self-Efficacy in Chronic Pain: from Exercise to Functional Activity*. IEEE Transactions on Affective Computing, 2018.
199. Olugbade, T., N. Bianchi-Berthouze, and A.C.d.C. Williams, *The relationship between guarding, pain, and emotion*. Pain reports, 2019. **4**(4).
200. Olugbade, T., et al. *What cognitive and affective states should technology monitor to support learning?* in *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*. 2017. ACM.
201. Olugbade, T.A., et al. *Pain level recognition using kinematics and muscle activity for physical rehabilitation in chronic pain*. in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2015. IEEE.
202. Olugbade, T.A., et al., *How Can Affect Be Detected and Represented in Technological Support for Physical Rehabilitation?* ACM Transactions on Computer-Human Interaction (TOCHI), 2019. **26**(1): p. 1.
203. Op De Beéck, T., et al. *Fatigue Prediction in Outdoor Runners Via Machine Learning and Sensor Fusion*. in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018. ACM.

204. Ouergui, I., et al., *Changes in perceived exertion, well-being, and recovery during specific judo training: Impact of training period and exercise modality*. Frontiers in Physiology, 2020. **11**: p. 931.
205. pandas. *pandas.DataFrame.interpolate*. [cited 2023 28 Feb]; Available from: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html>.
206. pandas. *pandas.DataFrame.resample*. [cited 2023 Feb 28]; Available from: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.resample.html?highlight=resample#pandas.DataFrame.resample>.
207. Picard, R.W., E. Vyzas, and J. Healey, *Toward machine emotional intelligence: Analysis of affective physiological state*. IEEE transactions on pattern analysis and machine intelligence, 2001. **23**(10): p. 1175-1191.
208. Plutchik, R., *The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice*. American scientist, 2001. **89**(4): p. 344-350.
209. Prapavessis, H., *The POMS and sports performance: A review*. Journal of Applied Sport Psychology, 2000. **12**(1): p. 34-48.
210. Preece, J. and A. Bowser, *What HCI can do for citizen science*, in *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 2014. p. 1059-1060.
211. Qiao-Tasserit, E., et al., *Transient emotional events and individual affective traits affect emotion recognition in a perceptual decision-making task*. PloS one, 2017. **12**(2): p. e0171375-e0171375.
212. Quigley, K.S., K.A. Lindquist, and L.F. Barrett, *Inducing and measuring emotion and affect: Tips, tricks, and secrets*. 2014.
213. Rahman, M.M. and D.N. Davis, *Addressing the class imbalance problem in medical datasets*. International Journal of Machine Learning and Computing, 2013. **3**(2): p. 224.
214. Raij, A., et al. *Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment*. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2011. ACM.
215. Reed, J. and D.S. Ones, *The effect of acute aerobic exercise on positive activated affect: A meta-analysis*. Psychology of Sport and Exercise, 2006. **7**(5): p. 477-514.
216. Reicherts, L., G.W. Park, and Y. Rogers. *Extending Chatbots to probe users: Enhancing complex decision-making through probing conversations*. in *Proceedings of the 4th Conference on Conversational User Interfaces*. 2022.
217. ROBERTSON, R.J., *Central signals of perceived exertion during dynamic exercise*. Medicine & Science in Sports & Exercise, 1982. **14**(5): p. 390-396.

218. Rogers, Y., H. Sharp, and J. Preece, *Interaction design: beyond human-computer interaction*. 2023: John Wiley & Sons.
219. Romanowski, A., et al. *Towards supporting remote cheering during running races with drone technology*. in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2017.
220. Russell, J.A., *A circumplex model of affect*. Journal of personality and social psychology, 1980. **39**(6): p. 1161.
221. Russell, J.A., A. Weiss, and G.A. Mendelsohn, *Affect grid: a single-item scale of pleasure and arousal*. Journal of personality and social psychology, 1989. **57**(3): p. 493.
222. Samson, A., et al., *Think aloud: An examination of distance runners' thought processes*. International Journal of Sport Exercise Psychology, 2017. **15**(2): p. 176-189.
223. Sasaki, T., *Recipient orientation in verbal report protocols: Methodological issues in concurrent think-aloud*. Second Language Studies, 2003. **22**(1): p. 1-54.
224. Savva, N. and N. Bianchi-Berthouze. *Automatic recognition of affective body movement in a video game scenario*. in *International Conference on Intelligent Technologies for interactive entertainment*. 2011. Springer.
225. Savva, N., A. Scarinzi, and N. Bianchi-Berthouze, *Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience*. IEEE Transactions on Computational Intelligence AI in games, 2012. **4**(3): p. 199-212.
226. Scherer, K.R., *Appraisal considered as a process of multilevel sequential checking*. Appraisal processes in emotion: Theory, methods, research, 2001. **92**(120): p. 57.
227. Scherer, K.R., A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research*. 2001: Oxford University Press.
228. Scherer, K.R., et al., *The GRID meets the Wheel: Assessing emotional feeling via self-report*. Components of emotional meaning: A sourcebook, 2013. **53**: p. 1689-1699.
229. Schneeberger, T., et al. *Towards a deeper modeling of emotions: The Deep method and its application on shame*. in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2021. IEEE.
230. Schücker, L., et al., *The effect of attentional focus on running economy*. Journal of sports sciences, 2009. **27**(12): p. 1241-1248.
231. Schücker, L., L. Schmeing, and N. Hagemann, *"Look around while running!" Attentional focus effects in inexperienced runners*. Psychology of Sport and Exercise, 2016. **27**: p. 205-212.

232. Schütte, K.H., et al., *Wireless tri-axial trunk accelerometry detects deviations in dynamic center of mass motion due to running-induced fatigue*. PloS one, 2015. **10**(10): p. e0141957.
233. Segal, Z.V., et al., *The mindfulness-based cognitive therapy adherence scale: Inter-rater reliability, adherence to protocol and treatment distinctiveness*. Clinical psychology & psychotherapy, 2002. **9**(2): p. 131-138.
234. Seuter, M., et al., *Live-feedback from the IMUs: animated 3D visualization for everyday-exercising*, in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 2016, ACM: Heidelberg, Germany. p. 904-907.
235. Shacham, S., *A shortened version of the Profile of Mood States*. Journal of personality assessment, 1983.
236. Shaji, R. and A. Singh. *Stridalyzer insight smart insoles: A clinical grade gait analysis system*. in *2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*. 2019. IEEE.
237. Shipway, R., I. Holloway, and I. Jones, *Organisations, practices, actors, and events: Exploring inside the distance running social world*. International Review for the Sociology of Sport, 2013. **48**(3): p. 259-276.
238. Siegel, E.H., et al., *Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories*. Psychological bulletin, 2018. **144**(4): p. 343.
239. Silvia, P.J., et al., *Missed beeps and missing data: Dispositional and situational predictors of nonresponse in experience sampling research*. Social Science Computer Review, 2013. **31**(4): p. 471-481.
240. Simpson, D., et al., *"It's not about taking the easy road": The experiences of ultramarathon runners*. The Sport Psychologist, 2014. **28**(2): p. 176-185.
241. Singh, H., et al., *The brain's response to pleasant touch: an EEG investigation of tactile caressing*. Frontiers in human neuroscience, 2014. **8**: p. 893.
242. Sloboda, J.A., S.A. O'Neill, and A. Ivaldi, *Functions of music in everyday life: An exploratory study using the Experience Sampling Method*. Musicae scientiae, 2001. **5**(1): p. 9-32.
243. Smith, C.A. and R.S. Lazarus, *Emotion and adaptation*. Handbook of personality: Theory and research, 1990. **21**: p. 609-637.
244. Soash, R.L., *MEDIA EQUATION: HOW PEOPLE TREAT COMPUTERS, TELEVISION, AND NEW MEDIA LIKE REAL PEOPLE AND PLACES*. 1999, Taylor & Francis.

245. Stephen, M., *Effect of Wind on Sound Transmission*. 2018: sciencing.com.
246. Stevinson, C.D. and S. Biddle, *Cognitive orientations in marathon running and "hitting the wall"*. British journal of sports medicine, 1998. **32**(3): p. 229-234.
247. Steward, C., *ATTENTIONAL FOCUS IN ENDURANCE RUNNING*.
248. Stieler, M. and C.C. Germelmann. *Fan Experience in Spectator Sports and the Feeling of Social Connectedness*. 2016. Cham: Springer International Publishing.
249. Stone, A.A. and S. Shiffman, *Capturing momentary, self-report data: A proposal for reporting guidelines*. Annals of Behavioral Medicine, 2002. **24**(3): p. 236-243.
250. Strohrmann, C., et al., *Monitoring kinematic changes with fatigue in running using body-worn sensors*. IEEE transactions on information technology in biomedicine, 2012. **16**(5): p. 983-990.
251. Strohrmann, C., H. Harms, and G. Troster. *What do sensors know about your running performance?* in *Wearable Computers (ISWC), 2011 15th Annual International Symposium on*. 2011. IEEE.
252. Strohrmann, C., et al., *Can smartphones help with running technique?* Procedia Computer Science, 2013. **19**: p. 902-907.
253. Summers, J.J., et al., *Middle-aged, non-elite marathon runners: A profile*. Perceptual and Motor Skills, 1982. **54**(3): p. 963-969.
254. Tajadura-Jiménez, A., et al. *As light as your footsteps: altering walking sounds to change perceived body weight, emotional state and gait*. in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015. ACM.
255. Tenenbaum, G., *A social-cognitive perspective of perceived exertion and exertion tolerance*. Handbook of sport psychology, 2001. **2**: p. 810-820.
256. Terry, P.C., A.M. Lane, and G.J. Fogarty, *Construct validity of the Profile of Mood States—Adolescents for use with adults*. Psychology of sport and exercise, 2003. **4**(2): p. 125-139.
257. Thin, N. *Happiness or wellbeing? Difficulties with definitions*. [cited 2023 27 Jan]; Available from: <https://www.happiness-lens.ed.ac.uk/happiness-or-wellbeing-difficulties-with-definitions/>.
258. Timmermann, J., W. Heuten, and S. Boll. *Input methods for the Borg-RPE-scale on smartwatches*. in *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. 2015.
259. Triantafyllopoulos, A., et al., *Fatigue Prediction in Outdoor Running Conditions using Audio Data*. arXiv preprint arXiv:2205.04343, 2022.

260. Tulving, E., *What is episodic memory?* Current directions in psychological science, 1993. **2**(3): p. 67-70.
261. Vallerand, R.J., *A hierarchical model of intrinsic and extrinsic motivation for sport and physical activity*, in *Intrinsic motivation and self-determination in exercise and sport*. 2007, Human Kinetics: Champaign, IL, US. p. 255-279,356-363.
262. Vallerand, R.J., *Intrinsic and extrinsic motivation in sport and physical activity*. Handbook of sport psychology, 2007. **3**: p. 59-83.
263. Valsted, F.M., et al. *Strive: exploring assistive haptic feedback on the run*. in *Proceedings of the 29th Australian Conference on Computer-Human Interaction*. 2017.
264. Van Berkel, N., D. Ferreira, and V. Kostakos, *The experience sampling method on mobile devices*. ACM Computing Surveys (CSUR), 2017. **50**(6): p. 1-40.
265. Van Leeuwen, L., S. Quick, and K. Daniel, *The Sport Spectator Satisfacion Model: A Conceptual Framework for Understanding the Satisfaction of Spectators*. Sport Management Review, 2002. **5**: p. 99-128.
266. Van Oost, E. and D. Reed. *Towards a sociological understanding of robots as companions*. in *International Conference on Human-Robot Personal Relationship*. 2010. Springer.
267. Veerasawmy, R., *Designing technology for spectator experiences–Beyond the passive spectator*. DHRS 2009, 2009: p. 55.
268. Velloso, E., et al., *The feet in human--computer interaction: A survey of foot-based interaction*. ACM Computing Surveys (CSUR), 2015. **48**(2): p. 21.
269. Vigotsky, L.S., *Thought and speech*. Psychiatry, 1939. **2**(1): p. 29-54.
270. Walter, S., et al., *Automatic pain quantification using autonomic parameters*. Psychology & Neuroscience, 2014. **7**(3): p. 363.
271. Wei, J., T. Dingler, and V. Kostakos, *Understanding User Perceptions of Proactive Smart Speakers*. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021. **5**(4): p. 1-28.
272. Wei, J., et al., *Understanding How to Administer Voice Surveys through Smart Speakers*. Proceedings of the ACM on Human-Computer Interaction, 2022. **6**(CSCW2): p. 1-32.
273. Wei, J., et al. *What Could Possibly Go Wrong When Interacting with Proactive Smart Speakers? A Case Study Using an ESM Application*. in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022.

274. Woźniak, P.W., et al., *Soil, Rock, and Snow: On Designing for Information Sharing in Outdoor Sports*, in *Proceedings of the 2017 Conference on Designing Interactive Systems*. 2017, ACM: Edinburgh, United Kingdom. p. 611-623.
275. Wollseiffen, P., et al., *The effect of 6 h of running on brain activity, mood, and cognitive performance*. Experimental brain research, 2016. **234**(7): p. 1829-1836.
276. Woźniak, P., et al. *RUFUS: remote supporter feedback for long-distance runners*. in *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 2015.
277. Woźniak, P., et al., *RUFUS: Remote Supporter Feedback for Long-Distance Runners*. Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services, 2015: p. 115-124.
278. Woźniak, P., et al., *Untangling running: designing for real-life runner experiences*. interactions, 2015. **22**(2): p. 40–43.
279. Wulf, G. and W. Prinz, *Directing attention to movement effects enhances learning: A review*. Psychonomic bulletin & review, 2001. **8**(4): p. 648-660.
280. Yan, X., et al., *Toward Lightweight In-situ Self-reporting: An Exploratory Study of Alternative Smartwatch Interface Designs in Context*. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020. **4**(4): p. 1-22.
281. Yannakakis, G.N., J. Hallam, and H.H. Lund, *Entertainment capture through heart rate activity in physical interactive playgrounds*. User Modeling and User-Adapted Interaction, 2008. **18**(1-2): p. 207-243.
282. Zeng, Z., et al., *A survey of affect recognition methods: Audio, visual, and spontaneous expressions*. IEEE transactions on pattern analysis and machine intelligence, 2009. **31**(1): p. 39-58.
283. Zuckerman, M., B. Lubin, and C.M. Rinck, *Multiple Affect Adjective Check List--Revised*. Journal of Behavioral Assessment, 1985.
284. 박소현, *Designing Conversational Agents to Encourage User Narrative and Self-Reflection in Mental Wellbeing*. 2020, 서울대학교 대학원.

Chapter 11 Appendices

11.1 Appendix A-Chapter 3

UCL INTERACTION CENTRE
RESEARCH•CONSULTANCY•SEMINARS•COURSES

UCL Interaction Centre
2nd floor, 66 – 72 Gower St
London WC1E 6EA



INFORMATION SHEET FOR REMOTE PARTICIPANT

Project Title: EnTimeMent

Department: University College London Interaction Centre

Researcher: Tao Bi, t.bi t.bi@ucl.ac.uk

Principal Researcher: Prof Nadia Berthouze, nadia.berthouze@ucl.ac.uk, 0203 108 7067

Please note that the phones attached to the numbers above are not manned during the period of the COVID-19 lockdown. During this period, kindly use the emails provided instead.

1. Invitation Paragraph

You are being invited to take part in a research project. Before you decide whether to take part, it is important for you to understand why the research is being done and what participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

2. What is the project's purpose?

The project extends previous studies where we developed scientific knowledge and technological tools toward supporting engagement in everyday physical activity for people with chronic pain and engagement in fitness activities for the general population. In the current project, we aim to further advance understanding of how to design a Chatbot to help people communicate and document their experience while running. This is part of a larger project called H2020 FET EnTimeMent (<https://entiment.dibris.unige.it>) funded by the European community and aimed at developing new movement sensing technologies that better define movement at different time scales.

3. Why have I been chosen?

We are asking people from the general population to volunteer for this study, if they are at least 18 years of age. You must also normally engage in running as we will not ask you to make any change to your normal plan of physical activity but only to record and reflect on your experience during it. We are looking for participants that use a novel research prototype designed by our researchers in their normal running routine.

4. Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part, you will be given this information sheet to keep and be asked to sign a consent form. You can withdraw from the study at any time and without giving a reason. You will not in any way be penalised for withdrawing from the study.



Approved Version No.: 1.0c Approved Date: 29 May 2019 UCL REC Project ID.: 5095/001

You can additionally withdraw data that you have provided to us, up till 14 days after the last data you send to us, without giving a reason. To withdraw from the study or withdraw your data, you should contact the researcher, using the contact details at the top of this information sheet. We will no longer use data that you withdraw if it is withdrawn no more than 14 days after the last data you send to us.

If you choose to withdraw from the study, you will be asked to revoke your original consent by signing a consent change form.

5. What will happen to me if I take part?

You will be asked to dialoguing with an existing running app or prototype. You may be asked to complete online questionnaires. You may additionally be asked to wear sensors that will capture and record your body movements and related physiological data (e.g. muscle activity, heart rate, electric skin resistance) while you run. The extent to which you do each exercise will be up to you, and you would be able to take breaks when you feel the need to. You should simply follow your normal exercise plan. We are not interested in performance. We only ask you to do the study when you run in safe place, ie., during the day, in parks (not street) and where other people are around.

We will further have an interview to discuss with you about your movements, related experiences, and/or your experiences with the engagement in reflecting and or reporting your experience while running through the prototype we provide.

6. Will I be recorded and how will the recorded media be used?

Your Name and Contact Information - NO ONE outside of our research team will be allowed access to any written or verbal information which can be used to personally identify you that you provide in your communication with us and in giving your consent.

What we will record:

- a. Textual data - Your responses to our questions in text form and the text diaries that you complete will be recorded in digital format either as online forms or other electronic documents, e.g. pdf forms.
- b. Video/photo/audio diary data - You will be asked to document your running experience and related thoughts/feelings using videos or photographs or audio. These forms of diary must NOT contain anyone that is not you. Should any videos or photographs or audio you send us contain someone who is not you or indecent content, we will edit out the third party and the indecent content before further processing of the data. You are also welcome to share with us some of your running data (e.g., route, speed) that you track with your usual Apps. If you would like to share your own running data, we ask you that you upload them through the UCL-Drobox (<https://www.ucl.ac.uk/dropbox/>) after encrypting and password protecting the files. This is for your security. We will show how to do so during our video conference meeting.
- c. Running data (movement, pace, distance, GPS) tracked by a running App on one of our study phones.

- d. Interview - Our interview with you will be done physically following social distance, or remotely via a telephone call or secure video/audio conferencing. For accurate detailing, the discussions will be video- or audio-recorded.
- e. All of the photos video, and audio data and any other data about your running will be stored securely and only accessed by members of our research team.

How we will use what we record:

The data we record (s text/video/photo diary, running data, discussion transcript and discussion audio/video recording) will be used for analysis. We will further use extracts which CANNOT be used to identify you from the running data and text (text responses, text diary, discussion transcript) data in scientific publications and presentations.

ADDITIONAL USE OF DATA THAT CAN BE USED TO IDENTIFY YOU - With your permission, we will use extracts from the video/photo/audio diary or discussion audio/video recording in presentations or publications, for illustration purposes. We will never disclose your name or contact details with these data. Note that you could be recognised from these photo, video. In addition, the audience may record the presentation or make copies of the video/photo included in the publication and we would not have control over the use of these recordings/copies. Nevertheless, these presentations and publications can enable the research community and the general public better understand the results of our research and contribute to scientific and/or technological advancement.

Also, if you give us permission to, we would use the data we record (running data, text responses, text/video/photo/audio diary, discussion transcript and discussion audio/video recording) for secondary analysis in other studies that our research team is involved in. This will support advance in better understanding of movement and the design of related technology.

ADDITIONAL USE OF DATA THAT CANNOT BE USED TO IDENTIFY YOU - If you give us permission, we will share anonymised data (sensor data, text responses, text diary, diary transcript, discussion transcript), which CANNOT be used to identify you, with the larger EnTimeMent project partners and the wider research community, without your names or contact details included. This will further support advance in better understanding of movement and the design of related technology.

7. What are the possible benefits of taking part?

Whilst there are no immediate benefits for those people participating in the project, it is hoped that this work will shape future research and contribute to the development of advanced technologies for supporting engagement in physical activity for people with chronic pain and for fitness activities for the general population.

8. What if something goes wrong?

Extreme care will be taken in this research. However, if you wish to complain or have any concerns that are not addressed by the researcher, you should contact Prof Nadia Berthouze (nadia.berthouze@ucl.ac.uk) who is the Principal Researcher on the project. If you further feel that your complaint has not been handled to your satisfaction, you can contact the Chair of the UCL Research Ethics Committee (ethics@ucl.ac.uk).

9. Will my taking part in this project be kept confidential?

Your name and contact details will be kept strictly confidential. The video/audio/images that we recorded of you will also be kept strictly confidential and only used as outlined in (6) above. We would only use examples from these data in ensuing publications or presentations (e.g. academic journals, teaching presentation) if you give us the permission to do so as detailed above in (6). We would further only share DE-IDENTIFIED text and sensor data with other researchers, and only if you give us permission to.

10. Limits to confidentiality

Confidentiality will be respected unless there are compelling and legitimate reasons for this to be breached. If this happens, we will inform you of any decisions that might limit your confidentiality.

11. What will happen to the results of the research project?

The findings of our analysis of the data collected from the participants of the research project will be published in reports and articles and presented at public engagement and research talk venue. You will be able to access academic publications of these findings on the project website: <https://entiment.dibris.unige.it>. You will not be identifiable in these publications and presentations except you give us permission to include examples from the videos/images we collect from you. Your name and contact details will never be included in publications and presentations.

The DE-IDENTIFIED text and sensor data we collect from you, which CANNOT be used to identify you, will be made open for use by other researchers, for the benefit of scientific and technology development; but only if you give us permission to do so.

12. Local Data Protection Privacy Notice

Notice:

The controller for this project will be University College London (UCL). The UCL Data Protection Officer provides oversight of UCL activities involving the processing of personal data, and can be contacted at data-protection@ucl.ac.uk

This 'local' privacy notice sets out the information that applies to this particular study. Further information on how UCL uses participant information can be found in our 'general' privacy notice: <https://www.ucl.ac.uk/legal-services/privacy/ucl-general-research-participant-privacy-notice>

The categories of personal data used will be as follows:

- Name
- Contact details
- Body movement data
- Physiological signals
- Photos and Videos and Voice in video/audio recordings
- Physical activity information and information about related experiences

The lawful basis that would be used to process your *personal data* will be performance of a task in the public interest. The lawful basis used to process *special category personal data* will be for scientific and historical research or statistical purposes.



CONSENT FORM FOR REMOTE PARTICIPANT

Project Title: EnTimeMent

Department: University College London Interaction Centre

Researcher: Tao Bi, t.bi t.bi@ucl.ac.uk

Principal Researcher: Prof Nadia Berthouze, nadia.berthouze@ucl.ac.uk

This study has been approved by the UCL Research Ethics Committee: Project ID number 5095/001

Thank you for considering taking part in this research. Please complete this after you have read the Information Sheet provided by the researcher.

1. I confirm that I have read and understood the Information Sheet for the above study. I have had an opportunity to consider the information and what will be expected of me. I have also had the opportunity to ask questions which have been answered to my satisfaction.	Yes/No
2. I confirm that I am at least 18 years old.	Yes/No
3. I consent to participate in the study and I understand that the discussion with me will be audio recorded to enable accurate record. I understand that my personal information (text question responses, running data (route, speed, etc.), interview transcript, audio recordings) will be used for the purposes explained to me. I understand that according to data protection legislation, 'public task' will be the lawful basis for processing for my personal data and the lawful basis used to process any special category personal data will be for scientific and historical research or statistical purposes.	Yes/No
4. I understand that my participation is voluntary and that I am free to withdraw from the study at any time. I understand that if I decide to withdraw from the study, I can revoke my consent by signing a consent change form. I understand that I can also withdraw data I have provided, up till the 14th day after the last data I submitted, without giving a reason. If I choose to also withdraw my data, the data will no longer be used.	Yes/No
5. I am aware of who I should contact if I wish to lodge a complaint.	Yes/No

6. I understand that identifiable information collected will remain confidential and that all efforts will be made to ensure that I cannot be identified except where explicit consent is given below. Yes/No

FURTHER USE OF MY DATA

7. I agree for my running data, text question responses, and interview transcripts, which CANNOT be used to identify me, to be made open to other researchers and for future research or secondary analysis, to facilitate research and innovation. Yes/No
8. I agree for PHOTOS recorded from me to be included in *written publications used to disseminate the project findings*. I understand that I could be recognised from such images and the audience may make copies of the image(s) and that the researcher will not have control over such copies. Yes/No
9. I agree for PHOTOS/VIDEOS/AUDIO recorded from me to be used in *presentations used to disseminate the project findings*. I understand that I could be recognised from such photos/videos/audio and the audience may record the presentation and that the researcher will not have control over such recordings. Yes/No
10. I agree for my personal information (running data, text question responses, text diaries, discussion transcript,) to be used for secondary analysis in other studies that the researchers of this study are involved in, *to facilitate research and innovation*. Yes/No
11. I agree to be contacted to participate in follow up studies to this project, or in future studies of a similar nature. I understand that my name and contact details will not be shared with anyone who is not a member of this research team. Yes/No

Preferred Contact Details (if 'Yes' to 11): _____

FULL NAME OF PARTICIPANT: _____

PARTICIPANT INFORMATION SHEET FOR ADDED PROCEDURES FOR REDUCING COVID-19 TRANSMISSION RISKS

This information sheet details the procedures and adjustments put in place to reduce risks related to COVID-19 transmission during the experimental procedure. We are following UCL and Government guidelines and will update these procedures regularly as the situation evolves.

We have implemented strict controls and regulations, because your safety, and the safety of the experimenters are our highest priority. Therefore, things may differ from any previous visit you may have made.

After reading this information sheet, you can contact the researchers to ask any questions you may have about participation in the study, and to ask for more information if anything is unclear.

The proposed set of procedures is designed to minimize the known risks of COVID-19 infection, but cannot abolish all risks. We encourage you to consider your participation carefully. It is up to you to decide whether or not to take part. If you choose not to participate, you won't incur any penalty or loss of benefits to which you are otherwise entitled. If you do decide to take part, you will be asked to read and answer a series of consent questions in order to document your agreement to participate.

The purpose of this document: This document should be read together with the additional information sheet detailing the procedures specific to the experiment in which you are participating. Here we will focus on COVID-19 related procedures. The other document will explain procedures specific to the research study.

Personal details: when you sign up to participate in this experiment you will be required to provide your contact details including your address, mobile phone number and email address. We will use this information to contact you for a health screening prior to the experiment. We may also share these details with government officials if contact-tracing becomes necessary, according to current Public Health regulations. This information will be kept secure, adhering to General Data Protection Regulation (2018) requirements and will be safely deleted after an appropriate amount of time (currently 3 weeks after the date of your visit to the lab).

Before your appointment: We will contact you 24 hours before your appointment to go over the details of your visit and to confirm your current state of health using the NHS COVID-19 symptom questionnaire (<https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/>). If you fall within an at-risk group, as defined by the NHS, or are sharing a household with an at-risk individual, we encourage you to carefully consider participating. You may also be requested to take a lateral flow test prior to your appointment and provide evidence of a negative result on arrival.

On arrival:

You will be required to sanitize your hands regularly as instructed by the experimenter.

During the experiment: We have adjusted the experimental procedures to minimize face-to-face interaction and maximize social distancing. Therefore:

- Strict disinfection procedures will be followed before, during, and after the experimental session.
- Instructions, explanation and interviews will follow social distance when it is outdoor. They may be delivered by computer or video rather than face-to-face.
- If the experimenter's presence in your vicinity is required, they will try to position themselves behind you to minimize face-to-face contact, and will maintain social distancing where possible.
- Even when not physically present, the experimenter will be available for the remainder of the session in case you need them. You will always be able to communicate with the experimenter by appropriate means (e.g. video chat, intercom)

Your responsibility:

- As part of the informed consent procedures we will be asking you to formally confirm that you understand and will adhere to these regulations.
- You must let us know if you test positive for COVID-19 within 1 week of your visit to the lab, so we can trace and inform any contacts that you may have had during your visit, such as the experimenter(s).

Potential risks:

Immediate risk: The procedures we have set up are designed to minimize the risk of COVID-19 infection. However, a certain level of risk is unavoidable. If you fall within an at-risk group, as defined by the NHS, or are sharing a household with an at-risk individual, we encourage you to carefully consider participating.

Secondary considerations:

- Your contact information will be shared with the Government contact tracing service if required.
- If you choose to travel to and from the site using public transport you must consider the risk of COVID-19 through exposure to other passengers.

**CONSENT FORM FOR HEALTHY ADULT VOLUNTEERS TAKING PART IN RESEARCH
STUDIES FOR ADDED PROCEDURES FOR REDUCING COVID-19 TRANSMISSION RISKS**

Please complete this form after you have read the Information Sheet for procedures in place for reducing COVID-19 transmission risk in experiments involving human participants.

If you are happy to take part in the study, we will need you to give your consent. To do this, please read the statements below and tick the appropriate boxes.

		Tick Box
1.	I confirm that I understand that by ticking each box below I am consenting to this element of the study. I understand that it will be assumed that unticked boxes mean that I DO NOT consent to that part of the study. I understand that by not giving consent for any one element that I may be deemed ineligible for the study.	
2.	I confirm that I have read and understood the Information Sheet for procedures in place for reducing COVID-19 transmission risks. I have had an opportunity to consider the information and what will be expected of me. I have also had the opportunity to ask questions which have been answered to my satisfaction.	
3.	<u>Personal Information</u> I understand that: - For purposes of Government-led contact tracing, I will need to supply my mobile phone number, address and email address. These data will be stored in accordance with General Data Protection Regulation (2018) that protects the privacy of personal information. - No one other than the researchers (and authorized government officials, if necessary) will have access to this data. - This data will be securely deleted 3 weeks after the participation date as shown on this form.	
4.	<u>Right to withdraw:</u> I understand that my participation in this study is voluntary and that I am free to withdraw from the study without giving a reason.	
5.	<u>COVID-19 transmission risk:</u> I understand that whilst every effort is made to reduce COVID-19 transmission risks down to manageable levels, the risk cannot be abolished completely.	
6.	<u>My responsibility:</u> - I confirm that I will adhere to the instructions detailed in the information sheet including those that relate to face covering, hand sanitizing and social distancing. - I understand that my contact details will be shared with Government officials in case contact tracing is necessary.	

11.2 Appendix B-Chapter 4

11.2.1 Supporting Spectator's Runner-Identity Process (Not relevant to SARE-ARS)

These design implications are not directly related to SARE-ARS or RQ1. However, they have been included in this report to provide a complete and thorough discussion.

Like the literature on team-identity [48, 94], findings in this chapter show that spectators experienced a kind of team-identity or more precisely a spectator-runner identity (SRI). The literature review showed that building an identity with the group of supporters and the team they support is critical to fully enjoy the event [88, 265] including LDRE [134, 276, 278]. This chapter's findings add to this literature by describing the process of building the SRI. In sports literature, the process of building a team identity is based on long term exposure to the professional sports team and the supporter group, enhanced by purposely created media representations of the team and sold merchandise [94]. This chapter's results show a different process of building SRI with friend/family runners, starting with the training period and continuing during LDRE.

Spectators are treated as passive consumers of information and to deliver cheering in existing LDRE technology. The findings in this chapter suggest a new space for design: to enhance spectators' engagement [276] [38] [156], satisfaction [88] [140] [248] and wellbeing [69]. This space is conducive to the building of the runner-identity. First, technology design could support spectators in their identity of Runner-Enabler and Supporter both during training and the LDRE. As discussed in 6.1 and 6.2, providing a better understanding of runners' 4-C is critical to building better knowledge of running and its needs and sensing technology could be leveraged for this purpose.

Another important design space, to support the process of building runner-identity (Section 5.5), is role recognition. Spectators were involved since the training period, e.g. taking care of children or cooking to provide free time to the runner (enabling), or tracking the runner to prepare food for them on time (supporting) [276]. Technology design should start recognize spectators' involvement since early stages (training) even before the journey in DyPECS. Furthermore, the 4-C effort during LDRE should also be recognized. For example, one of the authors ran 4.5 km to reach 4 viewing spots during her spouse's marathon. These unseen

efforts could be tracked using sensors currently designed only for runners and then represented on the spectator's annotated map (complementary to the runner's map) or sent to the runner at run-time for bidirectional sharing of the journey.

11.3 Appendix C-Chapter 5

11.3.1 Study Protocols

Participant ID	
Date	
Researcher	

INTRODUCTION & CONSENT

Hello!	
Describe study	
Give information sheet	
Get consent	

PRE-RUN QUESTIONS (PROFILE QUESTIONS)

How many times do you run? Per week? Per month? Per year?	
How far do you typically run?	
How far are you planning to run today? Why?	
Why are you running today? Training for a marathon? Running for fitness?	
What are your goals/expectations for this run? Pace? Anything else? Why?	

SELF-REPORT TRAINING

Think aloud training (See attachment 1)	
Exertion Scale + Pain Scale + Motivation Scale + Emotion	

SET UP SENSORS

Put on Insoles; connect insoles	

Input information in eSense (android phone);	
Connect left earbud to android phone Bluetooth	
Connect left earbud with eSense App; turn on playing question and recording;	
Connect iPhone with wired headphone; turn on recording	
iPhone recorder + eSense Recorder clapped 10 times	
Turn on the running app on apple watch (connected with iPhone);	
Lower down the screen light	

SYNCHRONIZATION & Trial

Sensor sync protocol (10s standing still)	
Arrange meeting at run end	
Check connection	
Check recording: eSenseIMU + InsolePreProcessing + InsoleProcessing (without disconnection)	

MAIN

Sensor sync protocol (10s standing still)	
Run (on the track)	
Meeting at run end (planned meet-point)	
Remove sensors	

DEBRIEF

Keep audio recorder on	
Interview questions Any feedback about the Scales? think aloud? Sensors? Distraction and awareness of sensors. Can you describe if you have your own definition of exertion and pain?	

Can you mark the exertion level you were planning to reach in the beginning for this run?	
Can you mark the overall exertion you have reached? Is the exertion level you want? below or above?	
How do you feel about this run? How your feelings vary? What are your most memorable sensations?	
Goodbye!	

11.3.2 Plots and visualizations

Figure 11-1, **Error! Reference source not found.** Figure 11-2, Figure 11-3, Figure 11-4, Figure 11-5, Figure 11-6 and Figure 11-7 show the line charts visualizing the time series pressure data for each sensor's (Hallux, OtherToes, Front, Front2, Mid, Arch, Heel2, respectively) during the acted data collections (see Chapter 5), in the case of P1's left foot insole, during the four acted emotional runs. The X-axis indicates the pressure value (kPa). The Y-axis indicates the number of time frames captured in each time series. It was observed that the number of data frames was smallest during runs simulating a state of wanted to run, followed by enjoyable runs, which were followed by tired and painful running despite the distance covered during each of this run was exactly the same distance of 40 meters. This could mean that in such states runners run may tend to be faster. The pressure value for the painful run was smaller than that of other conditions, which could indicate that the runner is engaging in a protective behaviour, avoiding putting too much pressure.

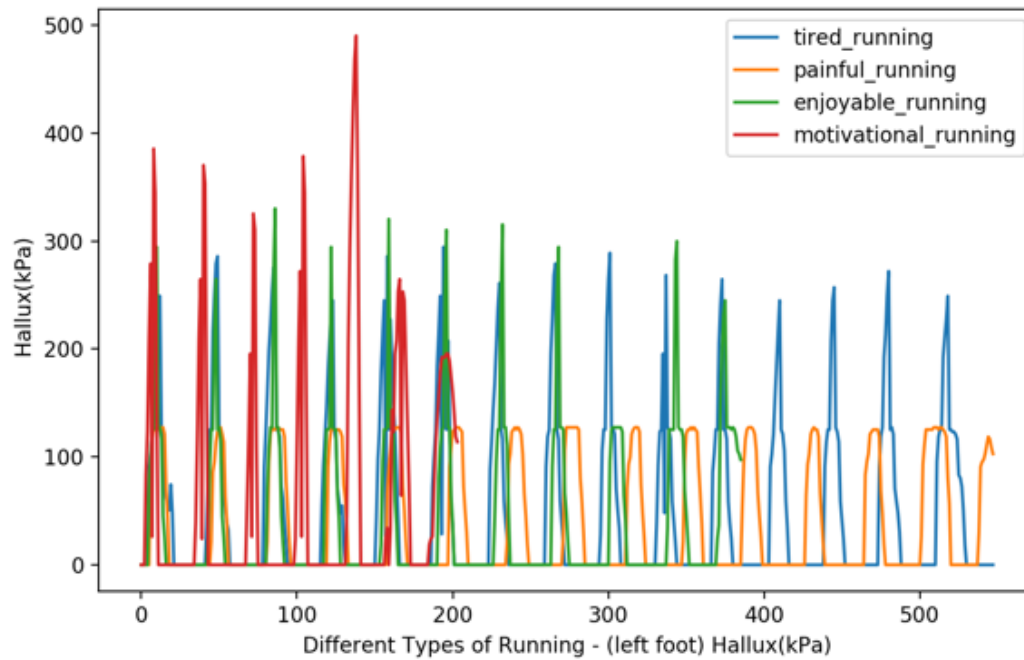


Figure 11-1: P1's Hallux's Pressure Changes under Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series)

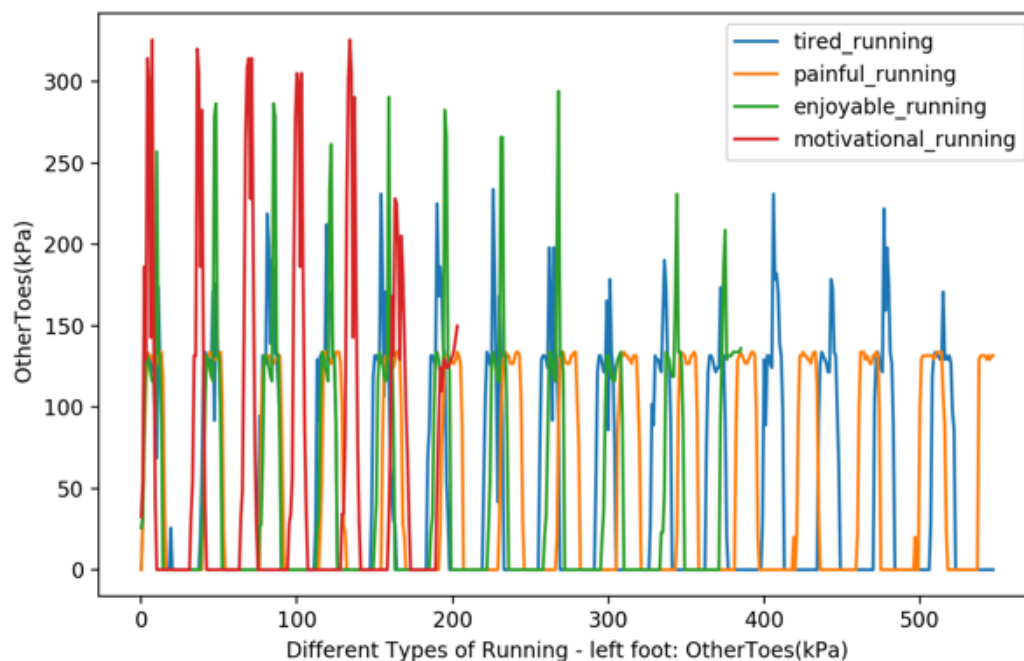


Figure 11-2: Line Chart of OtherToes' Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series)

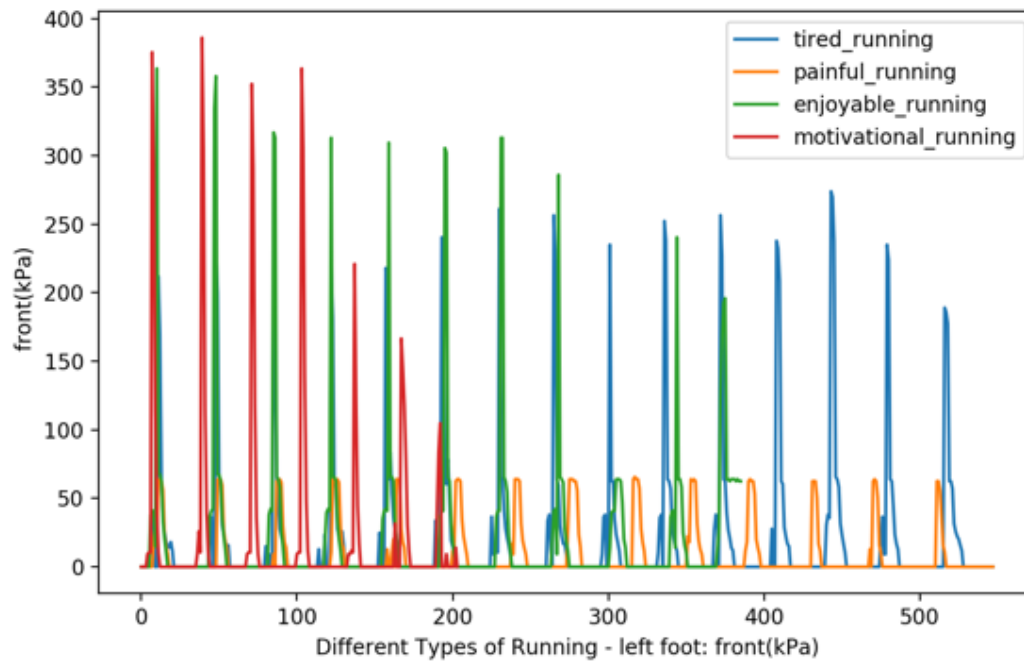


Figure 11-3: Line Chart of Front's Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series)

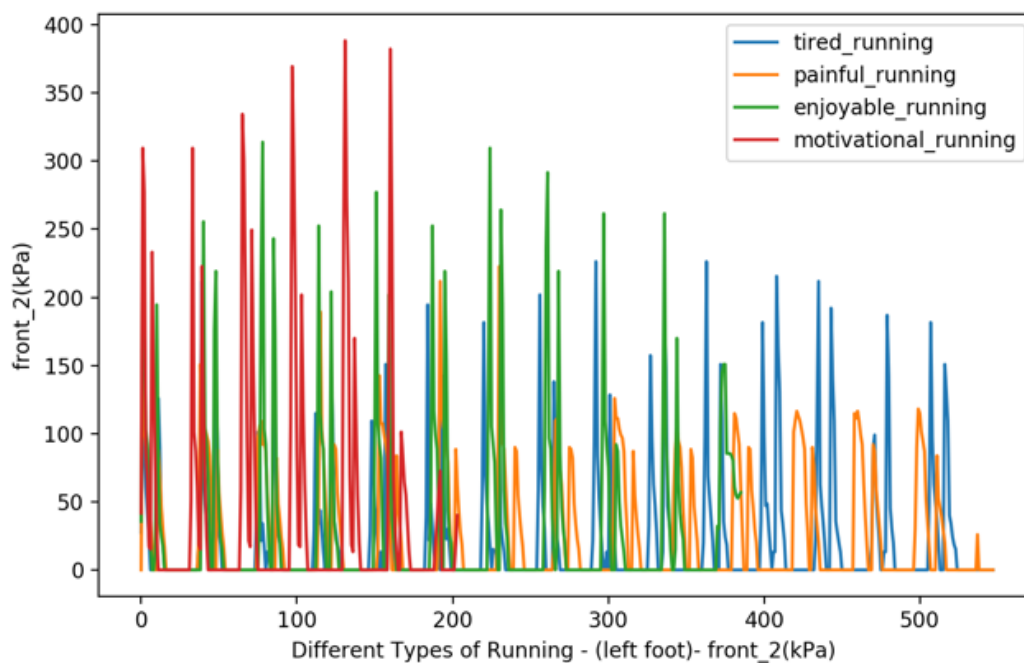


Figure 11-4: Line Chart of Front_2's Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series)

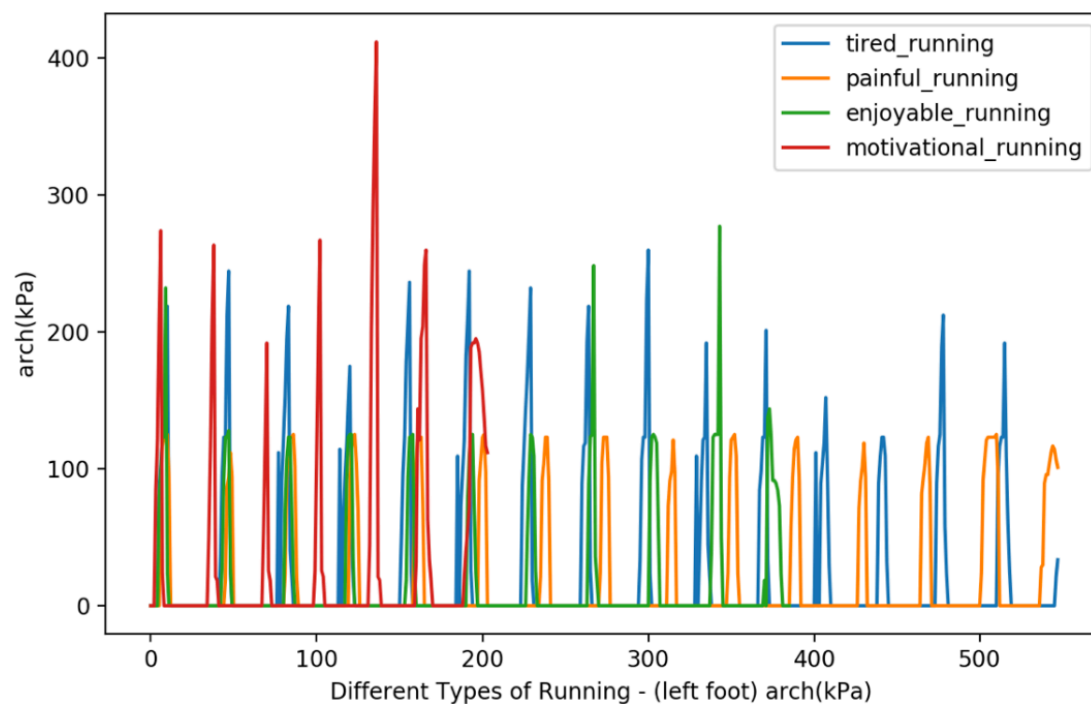


Figure 11-5: Line Chart of Arch's Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series)

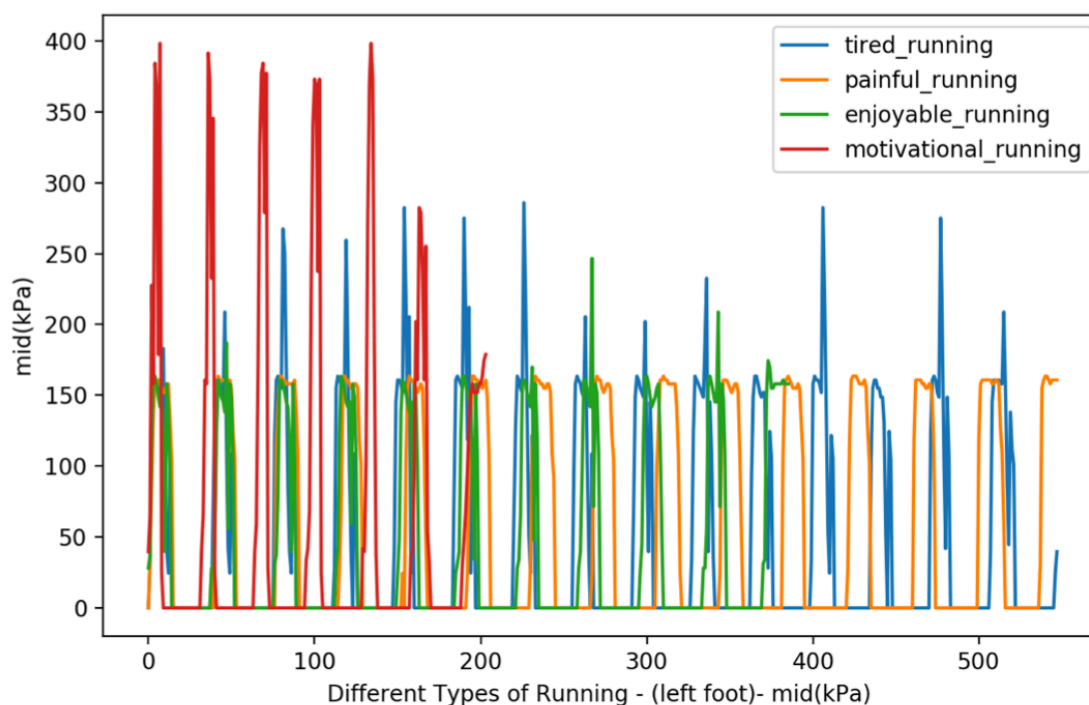


Figure 11-6: Line Chart of Mid's Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series)

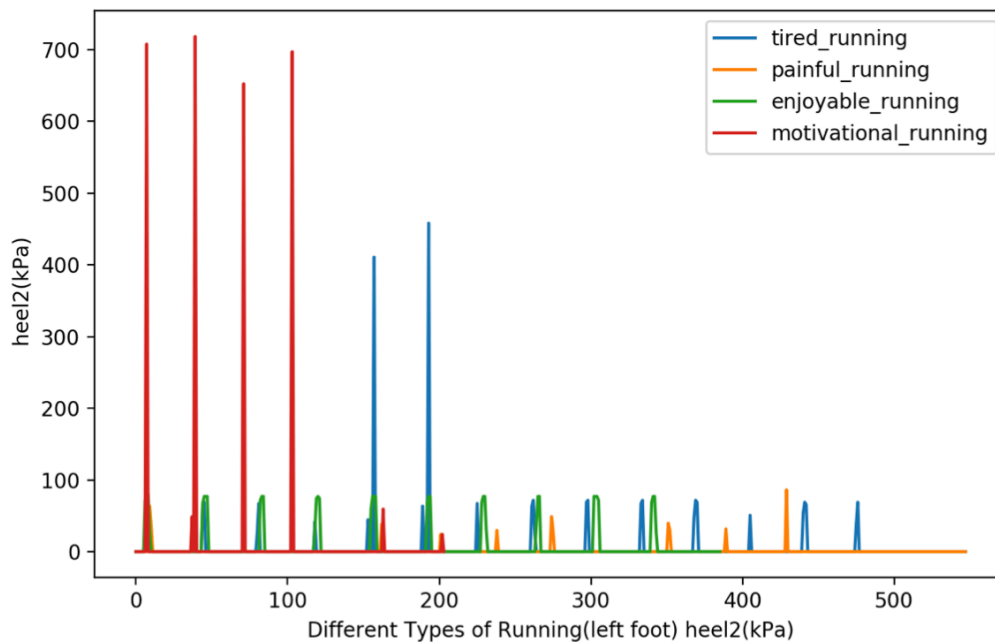
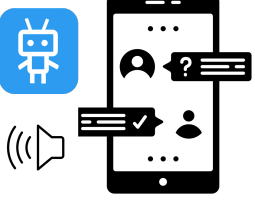


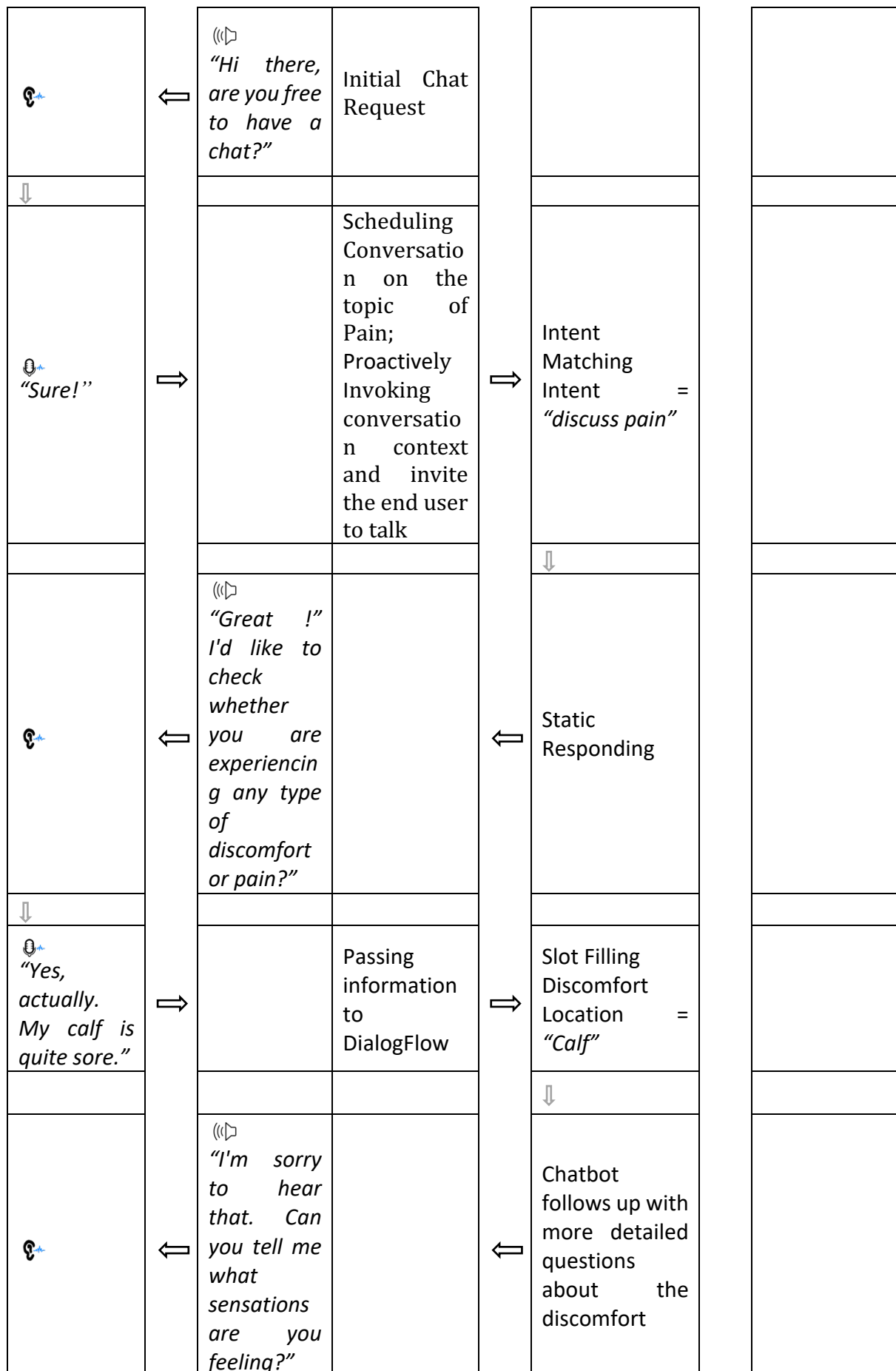
Figure 11-7: Line Chart of Time Series of Heel2's Pressure Time Series during Different Emotional Runs (Y-axis: pressure in Kpa; X-axis: number of data frames in the time series)

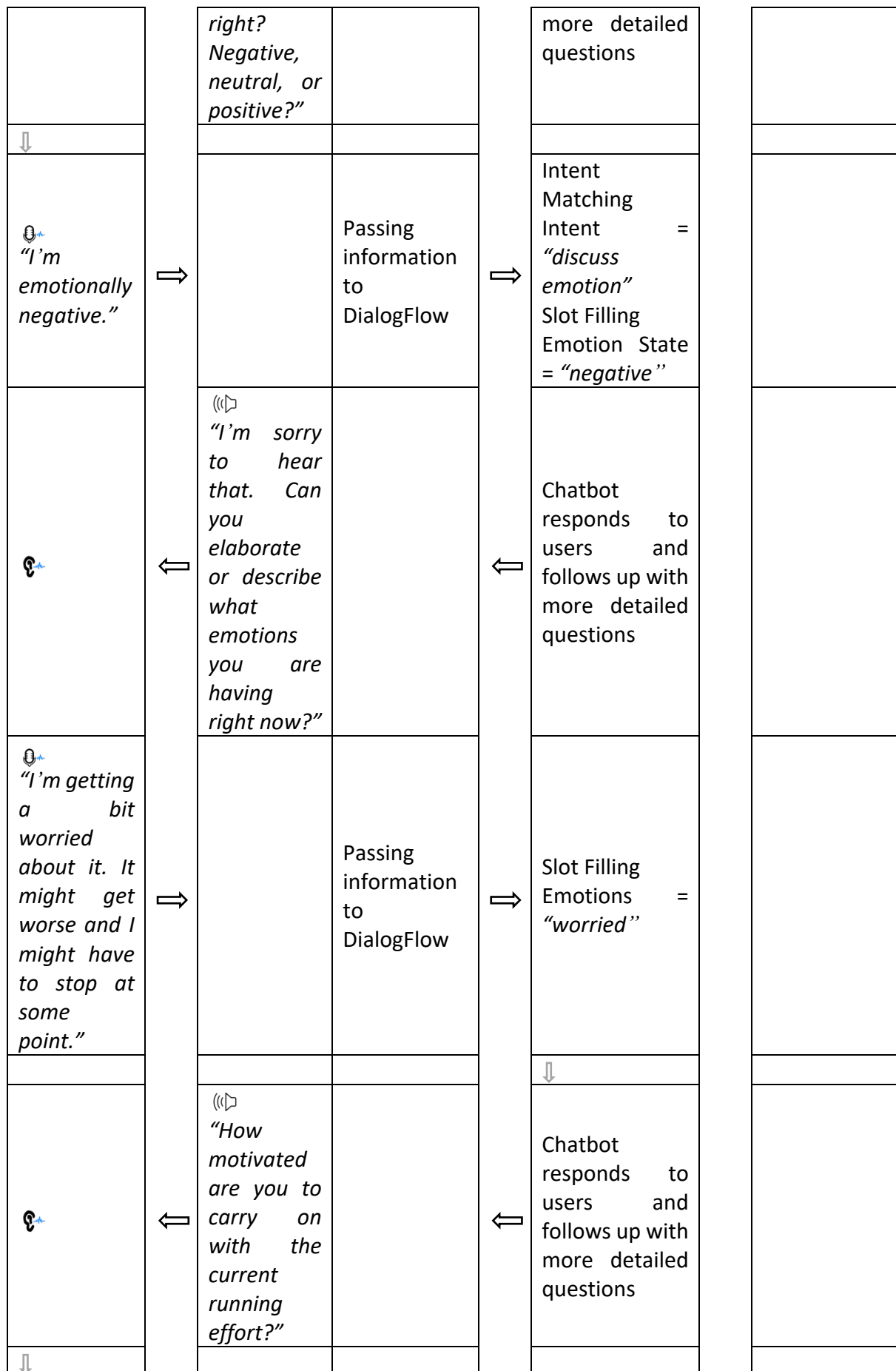
11.4 Appendix E-Chapter 7

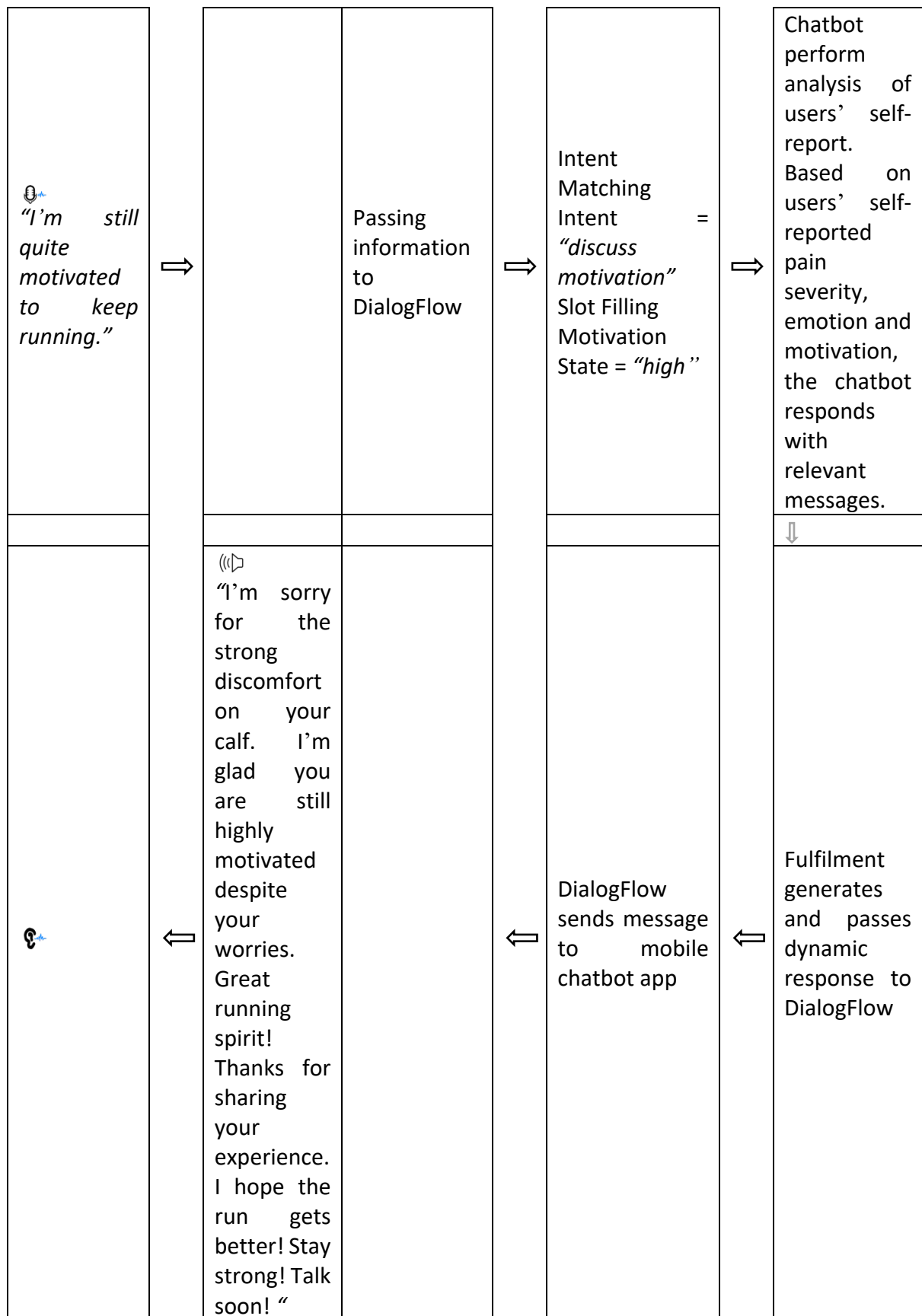
11.4.1 Demonstration of the conversation flow following the architecture

Table 11-1: Demonstration of the conversation flow following the architecture

				
Runner	Chatbot Voice Interface	Macro-CP with PIM	Micro-CP	Fulfilment







11.4.2 Documentation for Chatbot's Intent, Entity, Context, and Fulfilment Implementation

This table summarise the main intents implemented in our chatbot, their type (proactive, follow-up, etc.), triggering phrases, entities, input and output context, static responses and fulfilment functions.

Table 11-2 Documentation for Chatbot's Intent, Entity, Context, and Fulfilment Implementation

Intent Name	Intent Type (Chatbot's proactive Intent)	Root-level Intent? Follow-up Intent?	Settings	Invocation Keyword	End-user Expressions	Parameter Name	Entity	PROMPTS	Input Context	Output Context (life cycle)	Static Text Response	Dynamic Fulfillment Function	Pausing time before next intent	Next intent	End of conversation
"chatInvitation-Pain"	Chatbot's proactive Intent	Root-level Intent	Disabled ML	"1stChat-Pain"						chatPain-followup (2)	"Hi, would you like to have a quick chat"				
"acceptChat"	User-triggered	Follow-up Intent of	Enabled ML		"Yes"				chatPain-followup	chatPain-yes-	"Are you				
"yesPain"	User-triggered				"Yes"				chatPain-yes-	yes_Pain (5)	"Can you tell				
"noPain"	User-triggered				"No"				chatPain-yes-	No_Pain (5)	"Great. I'm				
"rejectChat"	User-triggered	Follow-up Intent of	Enabled ML		"No"				chatPain-followup	chatPain-yes-	"Oh! Sorry				
"rejectReason"	User-triggered	Follow-up Intent of	Enabled ML		Any						"Thank s for				
"chatPain"	User-triggered intent				XXXX	pain_disco	@human_b	"Can you tell me a bit more	yes_Pain	chatPain_followup (2)	None	Yes	2min		No
						mfort_locat	ody_parts	"I'm sorry to hear that. Can							
						pain_disco	@body_sensations	"Has this \$pain_discom							
						mfort_famil	@familiarity	"Can you pick a number							
					XXXX	mfort_inte	@sys.number	"Would you classify your			None	Yes			End
						emotion_valence	@emotion_valence								

"chatPain_affect"	User-triggered intent					emotional_words	@emotiona l_words	"What words would you use"	chatPain_followup	Pain_affect_info (5)					
						motivation_level	@motivatio n_level	"How motivated are"							
"repeatChat_pain"	Chatbot's proactive	Root-level Intent	Disabled ML	"2ndChat-Pain"						repeatChat-Pain-			None	Yes	
"no_previous_pain"	User-triggered				"No"				repeatChat-Pain-	noPreviousPain (5)			None	Yes	
"yes_newPain"									noPreviousPain	yesNewPain (5)			"Sorry to hear that."	None	
"newPainProfile"						pain_discomfort_location	@human_body_parts	"Can you tell me a bit more"	yesNewPainProfile	newPainProfile (5)	None	Yes			
						pain_discomfort_sensations	@body_sensations	"I'm sorry to hear that. Can"							
						pain_discomfort_familiarity	@familiarity	"Has this \$pain_discom"							
						pain_discomfort_intensity	@sys.number	"Can you pick a number"							
"newPainAffectProfile"						emotion_valence	@emotion_valence	"Would you classify your"	newPainProfile	Empty	None	Yes			End
						emotional_words	@emotiona l_words	"What words would you use"							
						motivation_level	@motivatio n_level	"How motivated are"							
"yes_previous_pain"	User-triggered	Root-level Intent			"Yes"				repeatChat-Pain-	yesPreviousPain (5)			None	"Hey there!"	
"previousPain_still"	User-triggered	Followup intent of			"Yes"	None			yesPreviousPain	yesPreviousPain_			"What's the"		
"previousPain_update"						pain_discomfort_intensity	@sys.number	"What's the current"	yesPreviousPain_still	yesPreviousPain_	"Is your emotio"				
3.2.1.2 "previousPain_affect_update"						emotion_valence	@emotion_valence	"Would you classify your"	yesPreviousPain	Empty	None				
						emotional_words	@emotiona l_words	"What words would you use"							
						motivation_level	@motivatio n_level	"How motivated are"							

"previous Pain_not	User- triggered	Followup intent of			"No"	None			yesPrevious Pain	yesPrevoi ousPain_					
yesNewP ain					"Yes"				yesPrevoio usPain_noE	yesPrevoi ousPain_					
yesNewP ain_profil e						pain_disco mfort_locat	@human_b ody_parts	"Can you tell me a bit more	yesPrevoio usPain_noE xist_newPai n	yesPrevoi ousPain_ noExist_n oNewPai n_affect (5)					
						pain_disco mfort_sens	@body_sen sations	"I'm sorry to hear that. Can							
						pain_disco mfort_famil	@familiarit y	"Has this \$pain_discom							
						pain_disco mfort_inte	@sys.numb er	"Can you pick a number							
yesNewP ain_affec t_profile						emotion_va lence	@emotion_ valence	"Would you classify your	yesPrevoio usPain_noE xist_noNew Pain_affect	empty					End
						emotional_ words	@emotiona l_words	"What words would you use							
						motivation al_level	@motivatio n_level	"How motivated are							
noNewPa in					"No"				yesPrevoio usPain_noE	yesPrevoi ousPain_					
noNewPa in_affect						emotion_va lence	@emotion_ valence	"Would you classify your	yesPrevoio usPain_noE xist_noNew Pain	Empty					End
						emotional_ words	@emotiona l_words	"What words would you use							
						motivation al_level	@motivatio n_level	"How motivated are							
"chatInvit ation- Exertion"	Chatbot's proactive Intent	Root-level Intent	Disabled ML	"&&&&& &"						chatExert ion- followup (2)	"Hi, would you like to have a quick				
"acceptC hatExerti					"Yes"				chatExertio n-followup	Chat_exe rtion (5)					
"rejectCh atExertio					"No"				chatExertio n-followup	chatExert ion-	"Can you				

"rejectChat Reason"					any				chatExertion-noChat-	empty						
"chatExertion"					"three"; "lungs"	exertion_level	@sys.number	"Between 1 to 5, how much	Chat_exertion	Exertion_info (2)		yes				
						exertion_triggers	@sys.any	"There are many factors								
"chatExertionAffect"						emotion_valence	@emotion_valence	"Would you classify your	Exertion_info	Exertion_affect_info (5)						
						emotional_words	@emotional_words	"What words would you use								
						motivation_level	@motivation_level	"How motivated are								
"repeatChatInvitation"	Chatbot's proactive	Root-level Intent	Disabled ML	"2nd_Chat-						2nd_chat Exertion						
"repeatChat-exertion_info"						exertion_level	@sys.number	"Between 1 to 5, how much efforts do you				2nd_chatExert				
						exertion_triggers	@sys.any	"There are many factors that								
"repeatChat-exertion_affect_info"						emotion_valence	@emotion_valence	"Would you classify your emotional state	2nd_chatExertion_affect_info							
						emotional_words	@emotional_words	"What words would you use to describe								
						motivation_level	@motivation_level	"How motivated are you to still achieve								
"start body scanning"	Chatbot's proactive Intent	Root-level Intent	Disabled ML	"body scanning"												

11.4.3 A Selection of Example Phrases for Chatbot Training.

Table 11-3: A Selection of Example Phrases for Chatbot Training

Bodily Locations	Feelings & Emotion
feet, left foot, right foot	Accepting
hands, left hand, right hand	Calm
mouth	Centered
legs, left leg, right leg	Content
foot	Fulfilled
abdomen	Patient
Adam's apple	Peaceful
adenoids	Present
adrenal	Relaxed
gland	Serene
anatomy	Trusting
ankle	Aliveness
anus	Joy
appendix	Amazed
arch	Awe
artery	Bliss
backball of the foot	Delighted
belly	Eager
belly button	Ecstatic
big toe, left big toe, right big toe	Enchanted
bladder	Energized
blood	Engaged
blood vessels	Enthusiastic
bone	Excited
brain, left brain, right brain	Free
breast	Happy
buttocks	Inspired
calf	Invigorated
capillary	Lively
carpa	Passionate
cartilage	Playful
cell	Contempt
cervical vertebrae	Disgruntled
cheek chest chin	Disturbed
circulatory system	Edgy

clavicle	Frustrated, annoyed, irritated, impatient, mad
coccyx	Hurt
collar bone	Insecure
diaphragm	Jealous
digestive system	Lonely
ear, left ear, right ear	Overwhelmed
ear lobe	Aggravated
elbow, left elbow, right elbow	Bitter

11.5 Appendix F-Chapter 8

11.5.1 Proposal for Further Chatbot Implementation

As demonstrated in Chapter 7, the chatbot's dialogue scripts were developed and defined through an iterative research and design process. These dialogue scripts, in a way, acted as prescribed interventions for facilitating the participants' self-reporting behaviours. To make the dialogue meet the research goal of sampling SARE self-reports, considerable effort was put into designing the overall conversation topics, logic, and language the chatbot should follow. The multiple versions of utterances and scripts design for the Chabot did provide user a dynamic conversation experience. However, this was not very easy as it required the researcher to have a good grasp of linguistics and language. Therefore, the chapter suggests that current Large Language Models (LLM), e.g., GPT-4, Bert, can be used to generate such human-like scripts with reduced effort for researchers based on simple user requests.

Although LLMs are powerful in generating human-like scripts, they are essentially mathematical prediction models, and the output of such generative content is still not reliable and sometimes unpredictable due to the nature of AI. Consequently, without the guidance or regulation from the researcher, conversations with an LLM-based chatbot might still be randomised in a naturalistic conversation context.

Neither the LLM or the current chatbot platform could proactively and systematically control the conversation interaction on its own. Therefore, researchers are still required to manipulate and design the topics, structure, and logic of the chatbot's conversation, especially when there is a specific research goal involved (e.g., the conversation needs to be designed to sample rich SARE self-reports).

Therefore, this chapter proposes a new conceptual architecture for **a three-user (Researcher, Chatbot, and User), semi-controlled, and collaborative conversation flow powered by LLM.** In this context, the researcher acts as a prescriber or supervisor, setting topics and conversation logic according to their research interests. The LLM is responsible for generating human-like scripts based on those requirements, and the voice-based chatbot serves as the facilitator, proactively using these prescribed, AI-generated, human-like scripts to interact with users.

This conceptual architecture can be implemented on top of the Nested Macro-Micro Conversation Pathway presented in Chapter 7. In the Macro-CP layer, an interface can be built for researchers to easily prescribe essential conversation elements. The prescription can include the components and elements identified in DAM Framework. Such information can be stored in a list variable referred to as the Researcher's Prescription (RP), with each component as RP_# (e.g., RP_1) and each sub-element as RP_#_# (e.g., RP_1_1).

Within the Macro-Conversation Pathway, RP_1 will be used as parameters to feed the LLM to generate human-like Researcher's Prescribed Dialogue Scripts (RP_DS_1). The voice-based chatbot then uses RP_DS_1 as the Chatbot's Dialogue Script 1 (C_DS_1) to proactively send to users, who then respond with User Response 1 (UR_1), activating the inner Micro-Conversation Pathway for conversing RP_#_#.

Within the Micro-Conversation Pathway, C_DS_1 and UR_1 are then fed into the LLM to generate a Generative human-like Response 1_1 (GR_1_1). Due to the unpredictable nature of the LLM, GR_1_1 might contain something unexpected or

not relevant to the research goal. Therefore, the Micro-CP accesses the RP_1_1 to feed the LLM to generate Researcher's Prescribed Dialogue Scripts 1_2 (RP_DS_1_2). GR1_1 and RP_DS_1_2 are together used as the Chatbot's Dialogue Script 1_2 (C_DS_1_2) to respond to users, who then respond with User Response 1_2 (UR_1_2). C_DS_1_2 and UR_1_2 can be fed into LLM models for GR1_2. RP_1_3 will be used to obtain RP_DS_1_3. GR_1_2 and RP_DS_1_3 are used as the Chatbot Dialogue Script 1_3 (C_DS_1_3). This process iterates until all sub-element (RP_#_#) are discussed.

Then the Macro-Conversation Pathway start conversation about RP_2, RP_DS_2, C_DS_2, UR_2, followed with the Micro-Conversation Pathway about GR_2_1, RP_2_1, RP_DS_2_1, C_DS_2_1, UR_2_2, RP_2_3, RP_DS_2_3, C_DS_2_3. Such process iterates until all components and elements in Researcher's Prescription (RP) are discussed.

11.6 Appendix G-Chapter 9

11.6.1 General Chatbot Conversation Flowchart

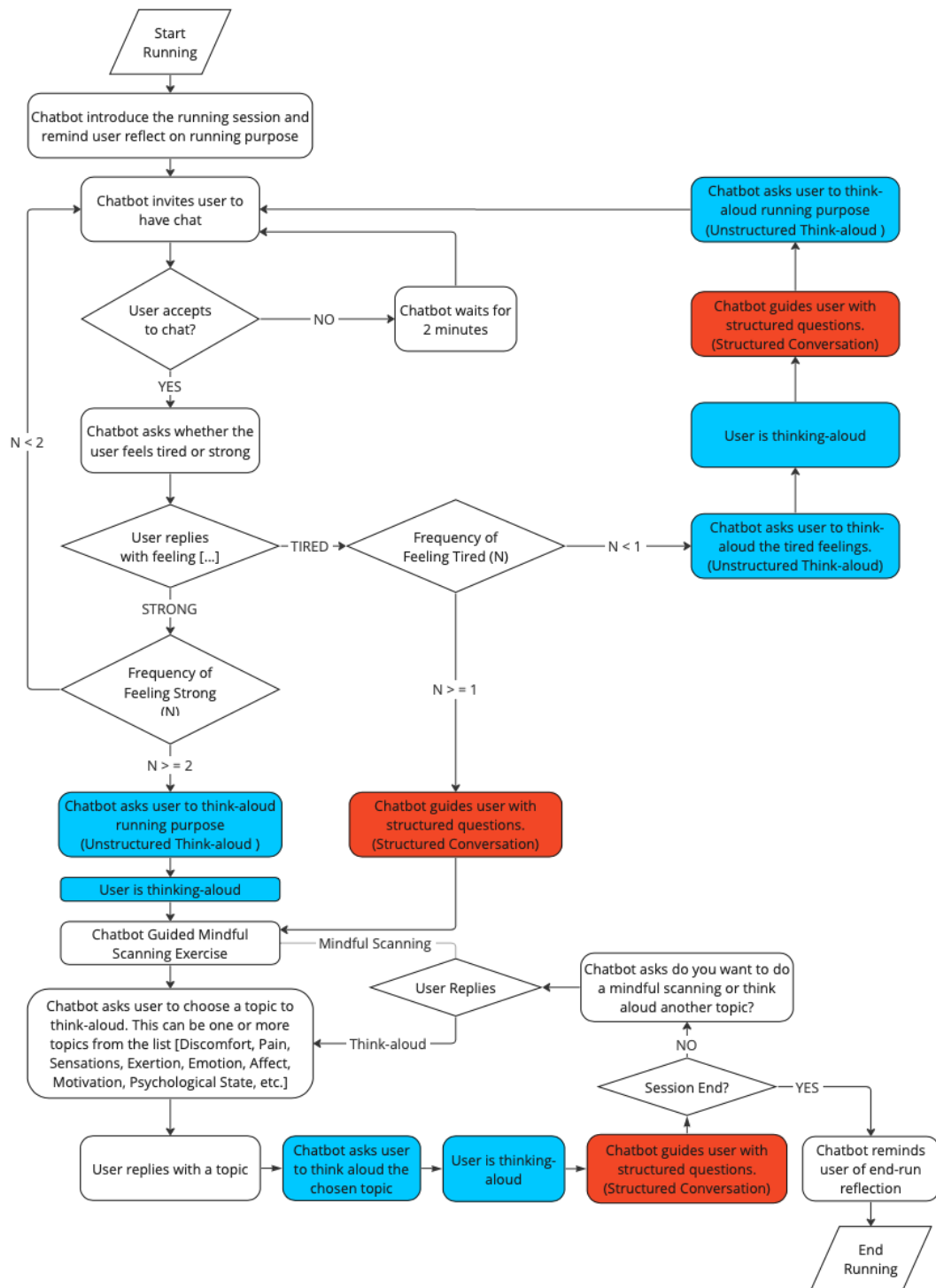


Figure 11-8: General Chatbot Conversation Flowchart

11.6.2 A taxonomy of Noise in Voice Self-reports while Running

In Chapter 9, the analysis of runners' voice recordings yielded a preliminary taxonomy of noise observed in voice recordings from runners during verbal self-report, suggesting ways for improved noise categorization, and extraction of contextual information from non-verbal aural signals.

This section provides a reflection on the collection of voice-based self-reports from runners using the microphone on the eSense smart earbud device. Although the R-ESM offers potential for collecting continuous voice self-reports from runners, they also capture a significant amount of noise signals from various sources, which presents challenges in data processing and analysis. Thus, it is useful to gain an in-depth understanding of the types of noise present in such voice self-reports, in order to provide insight into how technology could be designed to address this. Based on a qualitative analysis of voice recordings from eleven runners across fourteen in-the-wild running sessions (as described in the section 5.2), an initial taxonomy of noise in runners' voice self-reports data via eSense has been achieved. The taxonomy details and characterizes the observed noise, the challenges in achieving good-quality self-reports, and opportunities for extracting useful contextual information. Furthermore, a noise-categorization API for the eSense or other similar platforms is suggested, not only for the purpose of noise-cancellation, but also for incorporating the mining of contextual information for more meaningful event recognition.

11.6.2.1 Related Work

Traditional experience sampling methods (ESM) that use mobile phone-based applications require users to physically interact with a phone multiple times [46], which is not practical when collecting self-reports from runners while running, as the use of hands and arms is necessary for balance and the maintenance of movement flow. Smart earables offer great potential for experience sampling and the collection of user self-reports [47] in ubiquitous contexts, as they require less physical or

biomechanical interference with the runner's body movement [48], hence provide a hands-free experience for runners to self-report via voice recordings [21]. Additionally, earbuds are lightweight and widely accepted by many runners for listening to music while running. Despite the potential of earbuds to provide a more ubiquitous and less intrusive ESM, it is known that audio-based ESM is subject to potential recognition errors due to noise [33]. However, the noise type in voice self-reports in the context of running has yet to be specifically investigated.

11.6.2.2 Methods

Overall, 11 runners (five males, and six females) had a total of 14 running sessions. Nine of the runners completed only one session, one runner completed two sessions, and another runner completed three sessions. The duration of running sessions (and eSense data recording) averaged 34.7 ± 15.1 minutes. The researcher guided participants during the 5-min trail session. The study did not control the running environment, runner ability, or runner performance. Participants were given with flexibility to run anywhere at any self-selected pace. More details on the data collections can be seen in earlier chapters.

During the process of transcribing the voice-recording data in Nvivo [51], the researcher followed a semantic coding approach to identify sound that could potentially be regarded as noise in the runners' self-reports. The noise coding process was mainly based on the researchers' listening, subjective transcription and interpretation of audio. A name was given to each labelled sound or noise, together with a description of its characteristics. Other contextual information was also noted, based on the researchers' recall and notes taken during the data collection stage, such as the observation of the runners' clothes and accessories during the study sessions when participants ran indoors or at a reachable distance outdoors. The following sections present an initial taxonomy of the noises observed in the runners' voice self-reports.

11.6.2.3 Results: Taxonomy

As shown in Table 5.3.1 (see next page), this taxonomy consists of 14 noise categories: Car noise, Traffic noise, Sound of foot strike, Outdoor terrain noise, Earbuds rubbing noise, Breathing sound, Clothes rubbing noise, Wind noise, Personal item vibration noise, Treadmill machine noise, Foot strike and treadmill impact sound, Animal sound, Passenger talking, and Shop's speaker noise.

Each noise category is associated with its relevant running context (e.g., road run, gym treadmill run). For each category, the taxonomy also includes the characteristics and factors that influence the impact of the noise (e.g., pitch, volume, wind direction). It also classifies whether the noise is synchronized with the run, i.e. whether or not the noise is concomitant with the run. In addition, each category is evaluated with respect to its impact on the quality of voice self-reports. For example, a noise category is labelled as high impact if it severely renders the voice recording difficult to be understood and hence transcribed.

The noise category itself as well as the characteristics of the noise can be useful contextual info for understanding the noise experience, e.g., outdoor surface vs treadmill surface enables detection of running context, animal sounds might explain references to animals in self-report, frequency of foot strike can be useful in capturing fatigue or pacing, etc.

Table 11-4: Taxonomy of noise in runners' voice self-reports

Noise type	Characteristics & Factors	Synchronized with running?	Running context	Impact on speech recognition
Car noise	Scrapping or chirping sound when a car passes; Increasing pitch when the car is approaching; Decreasing pitch when the car is driving away; Honking sound;	No	Road run	High
Traffic noise	Motorbike engine noise; Car engine noise; Police car or emergency car alarm sound	No	Road run Park run	High
Sound of foot strike	High-rhythm sound during a fast run; Low-rhythm sound during a slow run; High volume sound on heavy landing; Low volume sound on light landing; Heel-to-front transition sound	Yes	Road run Gym run	High
Outdoor terrain noise	Road; Running Track; Grass; Trail; Treadmill	Yes	All run	High
Earbuds rubbing noise	sSense earbuds are unstable in intense physical activities. The rubbing noise has a very high frequency.	Yes	All run	High
Breathing sound	Inhale and exhale sound different.	Yes	All run	High
Clothes rubbing nose	Same frequency & rhythm as foot strike. Chuffing noise from arm, hands, legs	Yes	All run	High

Table 11-5: Taxonomy of noise in runners' voice self-reports (cont.)

Noise type	Characteristics & Factors	Synchronized with running?	Running context	Impact on speech recognition
Wind noise	Headwind; downwind; crosswind Wind direction causes different noise levels. Wind direction interacts with voice volume.	No	Outdoor run	High
Personal item vibration noise	Keys; Earrings; Necklaces	Yes	All run	High
Treadmill machine noise	Treadmill belt noise: squeaking, screeching and whining noises; Treadmill motor noise; High frequency & High volume; Other gym machine noise	Yes	Gym run	High
Foot strike and treadmill impact sound	Landing sound volume; Soft or heavy landing; High speed creates high pitched noise from treadmill machine;	Yes	Gym run	High
Animal sound	Birds sound, e.g., seagulls	No	Outdoor run	Low
Passenger talking	Group talk vs individual talk lower frequency voice vs higher frequency voice	No	Road run Gym run	High
Shop's noise	Music; Ads	No	Road run	High

11.6.2.4 Results: Demonstrations of audio and noise signals in a runner's voice self-reports

The taxonomy above shows that the noise observed had different characteristics and were typically high impact. Here, we further provide visual representations of the raw audio highlighting the noise signals, using the software Audacity [25], based on the data captured from one of the running sessions introduced in Section 2 as a case study. These visual representations aim to 1) illustrate not only how different noise categories and their characteristics impact the raw audio signals, and 2) illustrate how raw audio and noise signals could infer meaningful information.

Figure 11-9 shows the full voice recording for the running session. In this recording, all sounds (voice and noise signals) overlapped, which makes it difficult to recognize a runner's speech. Figure 11-10 shows an excerpt (approximately 3 minutes) of the full voice recording in Figure 11-9. In Figure 11-10, we distinguished between noise signals and speech. Voice signals when the runner was talking (segment C in the figure) occur throughout the excerpt. Meanwhile, there are aural occlusions at several points, e.g., from noise of one foot (segments A in the figure) or other (segments B in the figure) landing. The feet strike noise (segments A and B) overlaps with the voice signal (segment C). Further, the noise (segments A and B) are of higher volume than the voice (segment C), which makes it challenging to extract the voice content from the overlapped area. Beyond the negative implications of the noise, we deduced potentially valuable information from them. For instance, as can be seen in Figure 11-10, segment A always has higher volume than segment B, which suggests that this runner may have a higher landing impact in one foot than in the other.

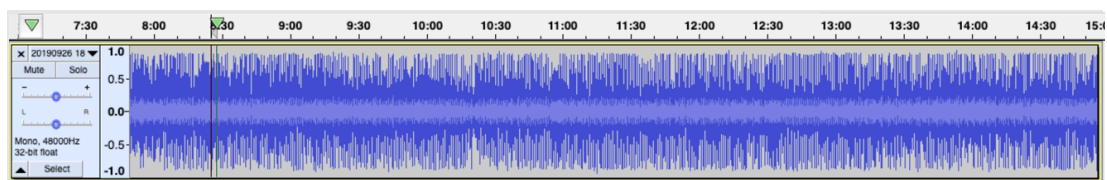


Figure 11-9: An overview of a runner's voice self-report while running

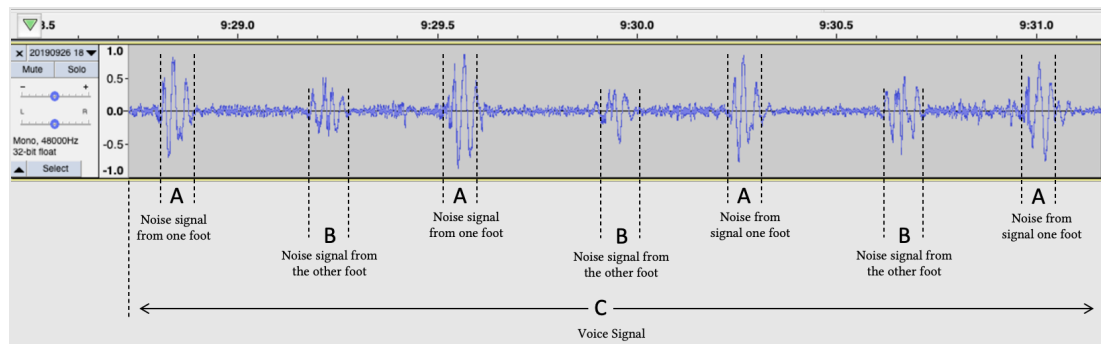


Figure 11-10: An excerpt of a runner's voice self-report

11.6.2.5 Discussions: How noise can contribute to useful information to SARE ground truth

As shown in Table 11-4, noise in runner voice self-report is complex but rich in information. Although we applied simple noise reduction techniques offered by Audacity [26], the results were not satisfactory enough to produce a good-quality voice recording. This is due to the irregular pattern of some of the noise such as wind noise, traffic noise, and noise from road surfaces with variations. In addition, the volumes of step and strike noise were much higher than the runner's voice. The noise from the earbuds' rubbing was even stronger as it was the closest to the microphone. Such factors and characteristics make it challenging to apply existing noise reduction techniques. While our analysis and discussion is limited in that we did not review how existing noise cancellation technologies may address the challenges that the observed noise and their characteristics pose, our findings outline critical issues that arise in running scenarios, some of these issues may be unique to ESM in running and may not be observed in other use cases (e.g. listening to music during outdoor walking, private phone call in a public space) that noise cancellation technologies will typically target.

The proposed taxonomy sheds a light on what type of noise existed in the running context, and the summary of features of each noise category can be useful to guide

development of noise reduction API (or real-time noise cancellation software) that address the specific scenario of running ESM or ESM in the wild in general. For instance, as highlighted in Table 11-4, headwind, downwind, or crosswind, not only due to the natural wind direction but also the running direction of a runner themselves, degrades the quality of voice signal as wind vibrates air particles that the voice vibrates [245]. Further, while crosswind direction could additionally alter the direction of voice transmission to one side; headwind could collide with the voice, more than tailwind. Noise cancellation and voice augmentation techniques could better accommodate such factors, to help researchers generate a better-quality voice recording via eSense or other similar platforms.

However, noise can have value, particularly in providing contextual information and personal affective experience. For instance, wind noise can provide information useful for contextualizing the runner's experience self-report. Headwind noise signals can be different from downwind noise, whereas headwind can contribute to making a run extremely difficult. In the transcribed voice self-reports, a runner for example said *"fight the headwind"* which could indicate that they were struggling or putting much higher effort into the moment. On the contrary, a runner *"feels a bit easier now, with the downwind"*. Here is another example, a runner said, *"I am cold, I'm going downhill so the wind catches me, I'm sure it will get better shortly"*. Such wind direction can be a factor that makes a runner perceive the run as difficult. In such a situation, runners might benefit from some sort of digital cheering that can be delivered via earbuds. On another hand, headwind can also be a feel-good factor in the run. For instance, a runner referred to wind as *"nice to run into a breeze on a hot day"*. When the temperature is high, but wind is gentle and cool, this could be a moment in which a runner mentally enjoys the run. Therefore, wind noise can not only serve as a contextual measurement but also a measurement for personal affective experience. Similarly, car noise can infer traffic and be used as a safety measurement. Digital reminders could be sent to runners to be aware of traffic if they are running on the road. Foot strike noise even has more information that can be

used to detect a runner's running performance such as cadence and stride type (heel stride, middle stride, front stride). It could also be used to infer a runner's mental state again. For example, fatigue could make a runner switch to more heel strikes, which is associated with a higher impact on the ground [162].

11.6.2.6 Conclusion

This part of results reports a preliminary taxonomy of noise observed in voice recordings from runners during verbal self-report, based on a qualitative analysis of the audio recordings. The taxonomy includes the noise types, their characteristics and factors, whether the noise type is synchronized with running or not, and how bad the noise affects the speech recognition quality. The findings highlight need for noise investigation before audio data collection in the wild. This taxonomy further highlights that noise can be both negative and positive factors. While noise clearly undermines extraction of the desired self-report of experience, it has the potential to provide rich contextual information for deeper understanding of the experience. eSense or other similar smart earbuds platforms will be more valuable when equipped with modules (e.g., API) capable of noise reduction, noise categorization, and extraction of contextual information from non-verbal aural signals.