On the Telemetric Measurement of Interoception: The Role of Interoception Assessment and Training in Future Digital Therapeutics

David Plans



Submitted for the degree of Doctor of Philosophy in

Psychology, Brasenose College, University of Oxford

September 2025

Abstract

Interoception, the perception of internal bodily signals, is increasingly recognized as a critical factor in mental health and the effectiveness of biometric-based digital therapeutics. This thesis presents a progression of research from developing and evaluating smartphone-based mental health interventions to creating novel measures of interoceptive ability suitable for remote deployment. The work addresses a fundamental gap in our understanding of why individuals with anxiety and depression often struggle to interpret biometric feedback and how to measure this capacity reliably.

The first phase of research (Chapters 2-4) evaluated the BioBase digital therapeutic platform, which combines real-time biometric feedback with psychoeducational content. A randomized controlled trial with 262 university students demonstrated significant reductions in anxiety ($d = 0.81$) and improvements in well-being ($d = 1.16$) sustained at 6-week follow-up. Parallel studies in workplace populations ($n = 55$) and individuals with generalized anxiety disorder ($n = 10$) confirmed therapeutic efficacy across diverse contexts. However, informal user feedback consistently indicated difficulties reconciling objective biometric data with subjective emotional experiences, suggesting that individual differences in interoceptive processing might moderate intervention effectiveness.

This observation motivated the development of novel interoception measures (Chapters 5-7). The CARdiac Elevation Detection (CARED) task employed ecological momentary assessment to capture interoceptive accuracy during daily life, with approximately one-third of participants demonstrating reliable cardiac interoception. The Phase Adjustment Task (PAT) was developed as a smartphone-based measure free from physiological confounds that plague traditional tasks. Validation with 124 participants revealed that 34-48% could reliably perceive their heartbeats, with performance independent of heart rate variability, body mass index, and other physiological factors. A subsequent validation study ($n = 59$) demonstrated convergent validity between the PAT and a newly developed Phase Selection Task ($r = 0.252$-0.349) while confirming discriminant validity through lack of association with exteroceptive control tasks.

The thesis makes three primary contributions: (1) empirical evidence for the effectiveness of biometric-based digital therapeutics for anxiety and depression, (2) identification of interoceptive processing as a potential mechanism underlying individual differences in response to biometric feedback interventions, and (3) development and validation of ecologically valid, smartphone-based measures of interoception suitable for large-scale deployment. These findings have important implications for personalizing digital mental health interventions based on interoceptive abilities and advancing our theoretical understanding of the relationship between bodily awareness and psychological well-being. The work establishes a foundation for interoception-informed digital therapeutics that could match intervention strategies to individual capacities for perceiving and interpreting physiological signals.

Acknowledgements

The work carried out in the seven years it took to run the studies necessary for this thesis was never done by me alone: my family allowed me absences and late nights that most partners and children would have found too difficult. The team I built at BioBeats worked alongside me for eight years. My co-founder, Davide, was a constant source of sanity throughout it. From raising capital to building our own wearable device device in China to dealing with some of the largest healthcare and corporate insurance companies in the world, their indefatigable work allowed me to see my designs and ambitions into tangible realities that touched a good few hundred thousand lives. I couldn't be more grateful to Elise for being an amazing mother to our children whilst putting up with the endless collisions between personal life and the running of a science startup, to which she also contributed. If you eventually moved on, it was no-one else's fault but mine. To my children, I know this work robbed you of a more present and calm father; my hope is that even still, you benefited from seeing someone push through scientific and industrial boundaries to strive for impact, and that this might one day drive you as well: I'm positive you will do it better than I did. To all our investors, partners, academic hosts and Brasenose College, thank you for allowing a vision that most deemed unfeasible to be pursued despite unreasonable demands on everyone's time. To my supervisor, Dr. Geoff Bird, whose undaunted patience and constant belief in even the most unreasonable technical and scientific objectives allowed this work to flourish, I owe an immense debt. To be trusted with building the impossible is often the catalyst to finding a way forward when all seems lost. Without his support through the many, frequent, and often paralysing challenges this project faced, it would have been absolutely impossible to finish it. When I was fishing in the dark for an answer to a question I didn't have the capacity to formulate, you showed me how to answer it, then pointed at a bigger, more profound one. In the end, this thesis is simply a witness to your sustained courage in allowing me to explore it.

Author contributions

Chapter 2: contributed to the study conception and experimental design, contributed to the programming of the BioBase smartphone application and its therapeutic modules, designed and built the BioBeam wearable heart rate monitoring device, conducted statistical analyses of the intervention outcomes, and led manuscript preparation and revision.

Chapter 3: co-designed the workplace adaptation of the study protocol, engineered the BioBeam wearable sensor for extended daily use monitoring, developed the workplace-specific content modules within the BioBase application, performed data processing and linear mixed model analyses, and contributed to the interpretation of feasibility outcomes and manuscript writing.

Chapter 4 : contributed to adapting the study design for clinical populations with generalized anxiety disorder, programmed the ACT-based therapeutic modules and integrated them into the BioBase platform, co-developed the BioBeam firmware for continuous physiological monitoring, analyzed user engagement metrics and clinical outcome data, and led manuscript writing.

Chapter 5 (CARED task development): co-conceived the ecological momentary assessment approach for the CARED task, contributed to the programming of the smartphone application including the heart rate detection algorithm and notification system, designed the data collection infrastructure for continuous monitoring over four weeks, conducted the statistical analyses including signal detection theory calculations, and led the writing of the methodological sections and contributed to other sections.

Chapter 6 (Phase Adjustment Task): primary developer of the Phase Adjustment Task concept and methodology, created the smartphone application implementing the cardiac phase adjustment interface, developed the photoplethysmography-based heartbeat detection system, performed data analyses including consistency scoring algorithms, and was the primary author of the manuscript.

Chapter 7 (PAT and PST validation): co-designed the validation study protocol including the pre-registration, programmed both the PAT and Phase Selection Task applications for cross-validation, developed the exteroceptive control task to assess discriminant validity, conducted the correlation analyses and non-parametric statistical tests, and contributed equally to manuscript preparation with co-authors.

On the Telemetric Measurement of Interoception: The Role of Interoception Assessment and Training in Future Digital Therapeutics 1

Abstract 2

Acknowledgements 4

Author contributions 5

1. Introduction 14

1.1. Definitions and models of interoception 14

1.2. Measurement of interoception 16

1.2.1. Cardiac interoceptive accuracy: current tools and limitations 16

1.2.2. Wider issues with the measurement of interoceptive accuracy 19

1.2.3. Measurement of other interoceptive processes 22

1.2.3.1. Relevance of interoception for well-being and higher-order cognitive processes 22

1.2.3.2. Conditions linked to atypical interoception 23

1.2.4. Interoceptive interventions 28

1.3. Interoception and Biometric Feedback in Digital Therapeutics 31

1.3.1. The Gap Between Biometric Data and Therapeutic Benefit 31

1.3.2. Individual Differences as Moderators of Intervention Effectiveness 32

1.3.3. The Need for Ecologically Valid Interoception Measurement 33

1.3.4. Toward Interoception-Informed Digital Therapeutics 33

1.3.5. Situating Biometric Digital Therapeutics within the MRC Framework for Complex Interventions 34

1.3.6. Rationale for the intervention configurations evaluated in Chapters 2–4 37

1.4. Overview of Studies in this Thesis 39

2. Chapter Two: Digital Biometric Feedback for Student Anxiety: A Randomized Controlled Trial 44

Content adapted from: Ponzo S, Morelli D, Kawadler JM, Hemmings NR, Bird G, Plans D (2020). Efficacy of the Digital Therapeutic Mobile App BioBase to Reduce Stress and Improve Mental Well-Being Among University Students: Randomized Controlled Trial. JMIR Mhealth Uhealth. https://doi.org/10.2196/17767 44

A note on contributions. This study was published with Sonia Ponzo as first author (Ponzo et al., 2020), reflecting her lead on the writing. My role spanned the trial's technical infrastructure — I designed and built the BioBeam wearable and contributed to the programming of the BioBase application and its therapeutic modules — and extended to the statistical analyses and to the drafting and revision of the manuscript. 44

Abstract 44

Introduction 45

Methods 47

Ethics Approval 47

Pre-registration 48

Study Design 48

Recruitment 48

Randomization and Blinding 49

Intervention Group 50

Figure 1. Timeline of the study. Participants were screened for inclusion and exclusion criteria. Of the 805 participants screened, only 262 were eligible to take part. These selected participants were subsequently randomly assigned to one of two groups: intervention or wait-list control. Each group received the questionnaires at baseline, 2, 4

and 6 weeks. The intervention group received access to the program following the first questionnaire completion. 51
Waitlist Control Group 53
Measures and Incentives 53
Primary Outcome 54
Secondary Outcome 54
Additional Measures 54
Statistical Analysis 55
Power 55
Data Exclusion 55
Data Analysis 56
Patient and Public Involvement 57
Results 57
Participant Enrollment and Demographics 57
Figure 2. Flowchart of enrollment and retention rates throughout the study in the intervention and wait-list control groups. 58
Figure 3. STAI scores at baseline, week 2, week 4 and week 6 follow-up from the start of the intervention in both intervention and wait-list control groups. Solid line=median; black dot=mean; whiskers: upper whisker=min(max(x), Q_3 + 1.5 x IQR); lower whisker=max(min(x), Q_1 - 1.5 x IQR). *P<.01. 59
Figure 4. WEMWBS scores at baseline, week 2, week 4 and follow-up from the start of the intervention in both intervention and wait-list control groups. Solid line=median; black dot=mean; whiskers: upper whisker=min(max(x), Q_3 + 1.5 x IQR); lower whisker=max(min(x), Q_1 - 1.5 x IQR). *P<.05. 59
Table 1. Summary of participant characteristics at baseline (T0) 60
Primary Outcomes 60
State-Trait Anxiety Inventory: Baseline to Follow-Up 60
Table 2. Summary of the linear-mixed model on State-Trait Anxiety Inventory scores over the 4 time points in the intervention and waitlist control groups 62
Table 3. Mean, SD, and planned comparisons on State-Trait Anxiety Inventory scores over the duration of the study (T0, T1, T2, and T3) in the intervention and waitlist control groups 62
Secondary Outcomes 63
Warwick-Edinburgh Mental Well-Being Scale: Baseline (T0)—Follow-Up (T3) 63
Table 4. Summary of the linear-mixed model on Warwick-Edinburgh Mental Well-Being Scale scores over the 4 time points in the intervention and waitlist control groups. 64
Table 5. Mean, SD, and planned comparisons on Warwick-Edinburgh Mental Well-Being Scale scores over the duration of the study (T0, T1, T2, and T3) in the intervention and waitlist control groups. 64
Additional Measures 64
Table 6. Summary of the linear mixed model on Patient Health Questionnaire scores over the duration of the intervention (T0, T1, and T2) in the intervention and waitlist control groups. 66
Table 7. Mean, SD, and planned comparisons on Patient Health Questionnaire scores over the duration of the intervention (T0, T1, and T2) in the intervention and waitlist control groups. 66
Figure 5. PHQ-9 scores at baseline and week 2 from the start of the intervention in both intervention and wait-list control groups. Solid line=median; black dot=mean; whiskers: upper whisker=min(max(x), Q_3 + 1.5 x IQR); lower whisker=max(min(x), Q_1 - 1.5 x IQR). *P<.01. 67
Discussion 67
Principal Findings 67

Comparison With Prior Work 68
Limitations 71
Conclusions 74
3. Chapter 3: Workplace Stress Reduction Through Biometric Feedback: A Feasibility Study 76
Content adapted from: Kawadler JM, Hemmings NR, Ponzo S, Morelli D, Bird G, Plans D (2020). Effectiveness of a Smartphone App (BioBase) for Reducing Anxiety and Increasing Mental Well-Being: Pilot Feasibility and Acceptability Study. JMIR Mhealth Uhealth. https://doi.org/10.2196/17767 76
Abstract 78
Introduction 79
Methods 80
Pre-registration 81
BioBase Programme 81
Recruitment 82
Procedure 83
Statistical Analysis 84
Patient and Public Involvement 84
Results 85
Participants 85
Table 1. Sample demographics of participants with baseline self-reported stress on DASS (N=55). 86
Engagement 87
Outcomes 87
Figure 1. After controlling for paid status, differences in (A) 6-item STAI and (B) WEMWBS after 4 weeks of the BioBase program. Association between transformed DASS Stress scores and difference in (C) STAI and (D) WEMWBS, **P<.01, *P<.05. DASS: Depression Anxiety Stress Scale; STAI: State-Trait Anxiety Inventory; WEMWBS: Warwick-Edinburgh Mental Well-being Scale. 88
88
Participant Feedback 88
Table 2. Feedback questionnaire results on intervention engagement, usability, and self-reported behavior change. 89
Participant Feedback: Feasibility and Usability 90
Participant Feedback: Acceptability in the Workplace 90
Participant Feedback: Promoting Behavior Change 91
Discussion 92
Principal Results 93
Comparison With Prior Work 94
Limitations 96
4. Chapter 4: Digital ACT for Generalized Anxiety Disorder: A Person-Based Design Study 101
Content adapted from: Hemmings NR, Kawadler JM, Whatmough R, Ponzo S, Rossi A, Morelli D, Bird G, Plans D (2021). Development and Feasibility of a Digital Acceptance and Commitment Therapy–Based Intervention for Generalized Anxiety Disorder: Pilot Acceptability Study. JMIR Formative Research. https://doi.org/10.2196/21737 101

A note on contributions. This study was published with Nicola Hemmings as first author (Hemmings et al., 2021); Nicola joined the team to lead the Person-Based Approach methodology that frames the study. I contributed to adapting the BioBase design for a clinical-anxiety population, programmed the ACT-based therapeutic modules and integrated them into the platform, co-developed the BioBeam firmware for continuous physiological monitoring, analysed the user-engagement metrics and clinical outcome data, and led much of the manuscript

writing. 101
Abstract 101
Introduction 102
Background 102
Acceptance and Commitment Therapy as a Treatment for GAD 103
Person-Based Approach 106
Aim 107
Methods 107
Pre-registration 107
Phase 1: Development 108
Technical Platform 108
Intervention 109
Stage 1: Development of a GAD Persona 109
Stage 2: Intervention Design and Creation 110
Table 1. The structure of the digital intervention prototype. The outline evidences the 3 core pillars of ACTa alongside the 6 ACT processes and links the intervention design objectives with the intervention key features (module name and exercise). 112
Phase 2: Feasibility Testing 113
Overview 113
Participants 113
Procedure 114
Duty of Care 114
Outcome Measures 115
Statistical Analysis 116
Patient and Public Involvement 116
Results 117
Participants 117
Figure 1. Study flow chart. 118
Table 2. Baseline characteristics of the study sample (N=10). 119
Duty of Care 119
Engagement 119
Figure 2. Completion order of ACT-based modules by participant. Participants onboarded onto the app over 2 days and had access to the app for 14 days. 119
Usability 120
Acceptability 120
Preliminary Efficacy 121
Outcome Measures 121
Preliminary efficacy in participants (N=10) of the digital intervention. The clinical cutoff for GAD-7 and PHQ-9 for moderate anxiety or depression, respectively, is 10. 122
Figure 3. GAD-7 and PHQ-9 results in participants (N=10) of the digital intervention. The clinical cutoff for moderate anxiety or depression, respectively, is 10. 122
Interviews 123
Discussion 124
Principal Findings 124
Acceptability 124
Usability 124
Engagement 125
Analysis of Initial Efficacy 125
Limitations 126

Conclusions 128
5. Chapter 5: CARED Task: Ecological Assessment of Cardiac Interoception 129
Content adapted from: Ponzo S, Morelli D, Suksasilp C, Cairo M, Plans D (2021). Measuring Interoception: The CARdiac Elevation Detection Task. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2021.712896 129
A note on contributions. This study was published with Sonia Ponzo as first author (Ponzo et al., 2021), reflecting her lead on the writing. I co-conceived the ecological momentary assessment approach used in the CARED task, contributed to the programming of the smartphone application — including the heart-rate detection algorithm and the notification system — designed the data-collection infrastructure for continuous monitoring over four weeks, conducted the statistical analyses including the signal-detection-theory calculations, and led the writing of the methodological sections of the manuscript. 129
Abstract 129
Introduction 130
Materials and Methods 132
Pre-registration 132
Participants 133
Materials 133
Design and Procedure 134
Figure 1: Example of a question screen comprising the heart rate judgment and the confidence rating. Following this screen, participants were presented with a free text box and asked to report what they were doing in the 30 min preceding the notification. 135
Data Analysis 137
Table 1: Summary of type 1 responses based on participants' answers to the question "Is your heart beating faster than usual?"(yes/no answer) and whether their heart rate was detected as higher than usual. 138
Patient and Public Involvement 139
Results 139
Figure 2: Empirical cumulative distribution (ECD) of (A) p-values obtained after comparing the Mann–Whitney U statistics against the population of reference and (B) effect sizes. (A) Red line vertical = 0.05, horizontal = 0.05 of the ECD; (B) Red line vertical = 0.2 (small effect size), horizontal = 0.2 of the ECD; Blue line vertical = 0.5 (medium effect size), horizontal = 0.5 of the ECD. Gray lines: upper and lower boundaries of p-values. 140
Figure 3: Empirical cumulative distribution of participants AUROC values obtained by applying the ROC function to correct and incorrect responses predicted by confidence ratings in such answers. Red lines: vertical = 0.5, horizontal = 0.5 of the ECD. Gray lines: upper and lower boundaries of effect sizes. 142
Table 2: Behavioural and questionnaires descriptive statistics and correlations between interoceptive accuracy (effect sizes) and BMI and self-reported measures of interoception and mental health. 142
Discussion 142
Hypotheses Emerging from the AUROC Distributions 143
6. Chapter 6: The Phase Adjustment Task: Smartphone-Based Assessment of Cardiac Interoception 147
Adapted from: Plans, D., Ponzo, S., Morelli, D., Cairo, M., Ring, C., Keating, C. T., Cunningham, A. C., Catmur, C., Murphy, J., & Bird, G. (2021). Measuring interoception: The phase adjustment task. Biological Psychology, 165, 108171. https://doi.org/10.1016/j.biopsycho.2021.108171 148
A note on contributions. I led this study. I was the primary developer of the Phase Adjustment Task concept and methodology, created the smartphone application implementing the cardiac-phase-adjustment interface, developed the photoplethysmography-based heartbeat-detection system, performed the data

analyses including the consistency-scoring algorithms, and was the primary author of the manuscript. 148
Abstract 148
Introduction 149
Study 1 151
Pre-registration 151
Materials and methods 152
Participants 152
Test delivery 152
PAT task: design and procedure 153
Figure 1: Screenshots of the PAT task mobile application. A) Screenshot of one of the 20 trials and B) body map screen. 155
PAT task: scoring 155
Tooling and data analysis 157
Comparison against random responding 157
Classification of participants as interoceptive or non-interoceptive 158
Bodily locations, average delays, and task engagement 159
Physiological variables 159
Task confidence and self-reported measures 160
Patient and Public Involvement 160
Results 161
Comparison against random responding 161
Figure 2: Probability density function of data from simulated participants responding at random (red line) and real participants' data (black line) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.). 162
Figure 3: Probability density function of real participants' data (black line), estimated distribution of non-interoceptive participants (red line) and interoceptive one (blue line) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.). 162
Classification of participants as interoceptive or non-interoceptive 162
Bodily locations, average delays, and task engagement 163
Table 1: Mean and standard deviations of the physiological, task engagement, self-reported interoception, and mental health variables for interoceptive and non-interoceptive participants, when participants are classified as interoceptive and non-interoceptive at Bayes Factor thresholds of 3, 10, and 30 in Study 1. 164
Table 2: T-tests of comparison between interoceptive and non-interoceptive participants on physiological, task engagement, self-report interoception and mental health variables, when participants are classified as interoceptive and non-interoceptive at Bayes Factor thresholds of 3, 10, and 30 in Study 1. 164
Physiological variables 165
Task confidence and self-report interoception measures 165
Empathy and mental health 165
Study 2: consistency of PAT scores after a one week delay 166
Materials and methods 166
Participants and procedure 166
Results 167
Bayes factor classification 167
Table 3: Bayes factor classification of participants successfully passing the screening test for Study 2. 167
Consistency in PAT similarity scores from T1 to T2 167

Classifications across measurement points 168
General discussion 168
7. Chapter 7: Validating the PAT: Convergent and Discriminant Validity Evidence. 175
Adapted from Plans, D., Bird, J. M., Morelli, D., Tam, R., Keating, C. T., Palmer, R., Adams, K. L., Piercy, T., Bird, G., & Murphy, J. (in preparation). Assessing the relationship between the Phase Adjustment Task and the Novel Phase Selection Task. 175
A note on contributions. This study was a collaborative effort with co-authors from the Murphy and Bird laboratories. I co-designed the validation study protocol, including the pre-registration, programmed both the PAT and the Phase Selection Task applications for cross-validation, developed the exteroceptive control task used to assess discriminant validity, conducted the correlation analyses and the non-parametric statistical tests, and contributed equally to the preparation of the manuscript with co-authors. 175
Abstract 175
Introduction 176
Methods 179
Participants 180
Materials 181
Phase Adjustment Task (PAT) 181
Task Description 181
PAT Screener Task 184
Task Description 184
Scoring 185
Phase Selection Task (PST) 185
Task Description 185
Figure 1. Example trial structure for the Phase Selection Task. After receiving instructions (see Supplement) participants are presented with a series of tones at one of three different delays (0ms, 250ms or 500ms after the detection of the pulse wave at the finger; presentation of delays is randomised across trials). Each delay is accompanied by a unique tone and each series is accompanied by a random colour-shape combination that changed on each trial. After listening to tones, participants are asked to rank the tones according to their delay from the heartbeat before confirming their response. The Phase Selection Control Task was identical to the Phase Selection Task, except that participants were required to match tones to the flash of the shape, with the same delays used for presentation. 187
Scoring 187
Phase Selection Task – Control Task (PST Control) 188
Task Description 188
Scoring 188
Procedure 189
Analysis Plan 189
Patient and Public Involvement 191
Results 191
Descriptives 191
Table 1. Presents mean and standard deviations of consistency scores on all tasks and the N of participants above chance cut-off. 192
Continuous scores 192
Classification scores 193
Exploratory analyses 194
Table 2. Classification scores across the three tasks 195
Table 3. Task performance as a function of order 195

8. Discussion 202
8.1. Summary of Empirical Chapters 202
8.1.1. Brief Summary 202
8.1.2. Chapter 2 202
8.1.2.1. Summary and interpretation 202
8.1.2.2. Strengths and limitations 204
8.1.3. Chapter 3 204
8.1.3.1. Summary and interpretation 204
8.1.3.2. Strengths and limitations 206
8.1.4. Chapter 4 207
8.1.4.1. Summary and interpretation 207
8.1.4.2. Strengths and limitations 208
8.1.5. Chapter 5 209
8.1.5.1. Summary and interpretation 209
8.1.5.2. Strengths and limitations 210
8.1.6. Chapter 6 211
8.1.6.1. Summary and interpretation 211
8.1.6.2. Strengths and limitations 212
8.1.7. Chapter 7 213
8.1.7.1. Summary and interpretation 213
8.1.7.2. Strengths and limitations 215
8.2. Overall Implications 216
8.2.1. Digital Therapeutics and Biometric Feedback 216
8.2.2. Measurement Innovation in Interoception Science 217
8.2.3. Theoretical Understanding of Interoceptive Difficulties in Mental Health 218
8.3. General Limitations 218
8.3.1. Sample Limitations 219
8.3.2. Methodological Considerations 219
8.3.3. Measurement Validity Questions 220
8.3.4. Data accessibility in industry-academic partnerships 220
Pre-registration and Transparency 221
8.4. Outstanding Questions 222
8.4.1. Mechanisms of Biometric Feedback Effectiveness 222
8.4.2. The Challenge of Binary Classification in Interoception Research 222
8.4.3. Distinguishing Interoceptive Abilities from Task-Specific Demands 224
8.4.4. Clinical Utility of Novel Interoception Measures 224
8.4.5. Longitudinal Development and Training Effects 225
8.5. Strengths of the Work Reported in this Thesis 225
8.5.1. Methodological Innovation in Interoception Measurement 225
8.5.2. Bridging Applied and Basic Science 226
8.5.3. Comprehensive Validation Approach 226
8.5.4. Person-Based Intervention Development 227
8.5.5. Patient and Public Involvement Across the Empirical Programme 228
9. Future Directions 231
9.1. Moving Beyond Binary Classification: Toward Dimensional Approaches 231
9.2. Ecological Validity and Real-World Assessment 233
9.3. Integration with Digital Therapeutics: From Assessment to Intervention 233

9.4. Experience-Based Co-Design for Future Iterations of Interoception-Informed Digital Therapeutics 234
9.5. Mechanistic Understanding Through Multimodal Research 236
9.6. Clinical Translation and Implementation Science 237
9.7. Technological Innovations and Emerging Opportunities 237
9.8. Addressing Fundamental Questions About Interoception 238
9.9. Methodological Innovations for the Phase Adjustment Task and Related Measures 239
9.10. Building an Evidence Base for Clinical Utility 240
9.11. Ethical Considerations and Responsible Innovation 240
10. Conclusions 241
Appendix A 244
Appendix B 245
End of study questions 245
Appendix C 246
Appendix D 250
CONSORT 2025 reporting checklist for the Chapter 2 randomised controlled trial 250
Note on the CONSORT flow diagram 256
Summary of retrospective limitations 256
Reference 257
Intention-to-treat sensitivity analysis for the Chapter 2 randomised controlled trial 258
1. What the published analysis actually estimated 258
Table E1. Trial flow and the participants entering each analytic frame. 258
2. Bias-correction synthesis using the Linardon and Fuller-Tyszkiewicz (2020) meta-analytic benchmark 259
3. Scenario-based sensitivity analysis using published summary statistics 259
Scenario A — Worst case (no-change imputation) 259
Scenario B — Best case (as-completers imputation) 260
Scenario C — Jump-to-reference (MNAR imputation favouring control) 260
Table E2. Summary of ITT-equivalent estimates from all methods. 260
4. The structural limitation: pre-baseline exclusion of 116 randomised participants 261
5. Summary and reported ITT-equivalent estimates 261
References 262
References 263

1. Introduction

1.1. Definitions and models of interoception

It is generally agreed that interoception refers to the processing of bodily signals that arise from the viscera. Introduced as a term by (Sherrington, 1907), the term interoception separated the perception of the internal surface of the body from the external surface of the bodily that directly interacts with the environment ('exteroceptive' signals) and signals relating to the movement and position of the body ('proprioceptive' signals). Since the introduction of this term, various definitions of interoception have been introduced with considerable debate regarding which signals should, and should not, be included under this umbrella term (e.g. (Cameron, 2001; Ceunen et al., 2013; Desmedt et al., 2025; Khalsa et al., 2009; Khalsa & Lapidus, 2016; Suksasilp & Garfinkel, 2022)). Acknowledging the difficulties defining interoception in terms of anatomical location, efferent innervation, or sensory properties (see (Ceunen et al., 2016)), this thesis defines interoception in line with contemporary redefinitions based on physiological pathways. As such, this thesis includes under this umbrella all signals conveyed via cranial nerves (e.g., vagus and glossopharyngeal nerves) and via humoral homeostatic pathways, of which there are multiple, to interoceptive regions in the brain such as the insula and anterior cingulate cortex (Ceunen et al., 2016; Craig, 2002; Critchley & Harrison, 2013), though it is acknowledged that the inclusion of certain signals (e.g., taste) remains a matter of debate (for a recent discussion see (Desmedt et al., 2022).

As well as debate regarding what should be considered an interoceptive signal, there is also debate regarding how best to classify individual differences in interoception. Acknowledging the ambiguity of terms such as 'interoceptive awareness' that had been used to describe performance on a variety of measures (from questionnaires to behavioural tests;

(Ceunen et al., 2013)), one of the first models of interoception made an important distinction between one's engagement by interoceptive signals ('interoceptive sensibility') as assessed via questionnaires, and one's objective performance on behavioural tests of 'interoceptive accuracy' where alignment between some objective metric (e.g., the number of heartbeats that occurred in a given time interval) and the participants' perception (e.g., the number of counted heartbeats) could be established (Garfinkel et al., 2015). The term 'interoceptive awareness' was used to refer to the correspondence between one's interoceptive sensibility and interoceptive accuracy (e.g., their metacognitive insight). Since the introduction and popularisation of this model, various other models of interoception have been introduced, including updates to the original three-dimensional model (e.g., (Desmedt et al., 2025; Khalsa et al., 2018; Murphy, Catmur, et al., 2017; Suksasilp & Garfinkel, 2022)). Whilst not all dimensions included in contemporary models are relevant to this thesis, it is worth acknowledging that newer models generally make a distinction between different processes (e.g., one's accuracy of perception, from one's attention) and acknowledge that one's interpretation of interoceptive signals may be an important individual difference (see Interoceptive interventions).

As well as distinctions across dimensions of interoception (that may not relate to each other; e.g., (Garfinkel et al., 2015; Murphy, Brewer, et al., 2017; Murphy et al., 2020), evidence also suggests that interoceptive ability (specifically accuracy), may dissociate across domains. Whilst some evidence supports associations (e.g., between the perception of cardiac and gastric sensations; e.g., (Herbert et al., 2012; Whitehead et al., 1980), the majority of evidence suggests that interoceptive ability is domain-specific, e.g., (Ferentzi et al., 2018; Garfinkel, Manassei, et al., 2016). Furthermore, evidence also suggests little correspondence between different tasks of interoceptive accuracy within the same domain, suggesting that tasks demands may strongly impact the measurement of interoceptive ability (e.g., (Hickman

et al., 2020)). The same is also true for self-report measures of interoception, that may assess different processes and often show little correspondence with each other (Desmedt et al., 2022; Gabriele et al., 2022; Murphy et al., 2020). In light of this evidence, this thesis is written in line with new suggestions that when reporting results of studies the specific task and domains examined are specified (Desmedt et al., 2025).

1.2. Measurement of interoception

It is widely acknowledged that measuring interoception presents a challenge and in recent years a plethora of novel measures (both behavioural and self-report) have been introduced. Whilst it is beyond the scope of this thesis to review all measures in detail, below a review of tasks of cardiac interoceptive accuracy is provided as it is the domain most relevant to this thesis.

1.2.1. Cardiac interoceptive accuracy: current tools and limitations

Cardiac interoceptive accuracy (the perception of heartbeat sensations) remains the most commonly assessed domain of interoception. Interest in cardiac interoceptive accuracy arose from work examining heart rate control (e.g., voluntary control of heart rate; (Carroll & Whellock, 1980; Dale & Anderson, 1978; Gannon, 1980; McFarland, 1975; Whitehead et al., 1977)) that was hypothesised to relate to individual differences in heart beat perception. Whilst no evidence for such links was observed (e.g., (Whitehead et al., 1977), interest in heart beat perception continued to rise arguably due to both the aforementioned work developing measures of cardiac interoceptive accuracy as well as evidence for links between the perception of heartbeat sensations and emotion (e.g., (Schandry, 1981)). Whilst there are several tasks of cardiac interoceptive accuracy, the two most popular tasks developed in this period that have dominated the field until the present day can be broadly separated into tasks that involve the tracking (or counting) of heartbeat sensations and signal detection measures.

The heartbeat counting task (Dale & Anderson, 1978; Schandry, 1981) requires participants to count their heartbeats over a series of intervals that are then compared to the objective count to determine accuracy. Whilst this task and similar variants (e.g., where participants are asked to tap in response to heartbeats; 'heartbeat tapping tasks'; for a review see (Brener & Ring, 2016)) are easy to implement, in the past 20 years various concerns have been raised regarding the use of this task as a measure of cardiac interoceptive accuracy. These concerns predominately centre on the risk of false positives as participants' beliefs appear to strongly influence performance on this task. For example, in a seminal study from (Windmann et al., 1999), participants with pacemakers were asked to complete the heartbeat counting task where unbeknown to them their pacemakers were set to low, medium and high rates. However, the authors observed no influence of this change on heartbeat counting scores despite comparable accuracy to controls. Since this study, a plethora of studies have been published that provide further evidence for concerns regarding the influence of beliefs on the heartbeat counting task (e.g., (Desmedt et al., 2018; Murphy, Millgate, et al., 2018; Phillips et al., 1999; Ring et al., 2015), as well as papers raising concerns about the psychometric properties of the task and the influence of individual differences in physiology (e.g., underreporting; heart rate (Zamariola et al., 2018). Addressing these concerns, it has been argued that concerns may be mitigated by the use of specific instructions that encourage participants only to report felt heartbeats and not to guess (Desmedt, Corneille, et al., 2020) or adapting the task such that participants report on heart rate change rather than count felt heartbeats (Larsson et al., 2021), or by examining the consistency of participants' responses in terms of the timing of motor responses (where heartbeat tapping tasks are utilised; e.g., (Smith et al., 2021)). Whilst examining consistency, rather than absolute accuracy, of tapping responses may improve precision as discussed below, it is notable that in these tasks participants are generally asked to tap 'in sync' with their heartbeats, rather than tap in

response to perceived heartbeats, meaning that estimation strategies may still play a role. Ultimately, the extent to which the aforementioned strategies fully mitigate the risk of false positives in the heartbeat counting task and variants remains to be established (Desmedt et al., 2023).

Concerns regarding the use of the heartbeat counting task and similar variants that may be influenced by estimation strategies, have prompted increased focus on so-called matching tasks of cardiac interoceptive accuracy (sometimes called signal detection methods) that are not influenced by estimation (Phillips et al., 1999). The most popular of these is known as the two-alternative-force-choice heartbeat detection task (2AFC; (Whitehead et al., 1977)). In this task, participants are presented with an external stimulus (typically a tone or a flash) that is presented in or out of sync with the heartbeat. Generally out of sync intervals are defined between 200-250ms (following the heart's R wave), with out of sync intervals defined as between 500-550ms (for a review see (Hickman et al., 2020)). Participants are required to judge whether the external stimulus is presented in or out of sync with their heartbeat. Whilst not influenced by estimation (reducing the risk of false positives; (Phillips et al., 1999)), the use of pre-defined delays as synchronous and asynchronous means that the 2AFC heartbeat detection task may be susceptible to false negatives (Brener & Ring, 2016; Plans et al., 2020). Indeed, whilst the delays selected for (a)synchronous trials are based on evidence that the majority of individuals perceive their heartbeats as synchronous at approximately 200-250ms following heartbeat contraction, this is by no means universal (Brener et al., 1993; Brener & Kluvitse, 1988; Brener & Ring, 2016; Clemens, 1984; Knoll & Hodapp, 1992; Yates et al., 1985). In fact, there appear to be notable individual differences in individuals' timing of heartbeat perception, so much so that evidence from several studies suggests that some participants present with inverse scores (indicating a preference for the asynchronous interval; (Brener et al., 1993; Wittkamp et al., 2018)). Whilst some analysis techniques can account for

inverse preferences (e.g., by taking an absolute d' value), where an individuals' preferred delay falls between these two intervals (e.g., they have a preference for 350ms where synchronous and asynchronous intervals are 200ms and 500ms, respectively) this individual is likely to respond randomly and may be falsely classified as non-interoceptive when simply they perceive external stimuli as synchronous with their heartbeat at a different delay (Brener & Ring, 2016; Plans et al., 2020).

Evidence (Carroll & Whellock, 1980) for issues with the 2AFC heartbeat detection task generally comes from 'multi-level' psychophysics tasks of interoceptive processes. In these tasks, participants are typically presented with multiple delays following R wave contraction (e.g., 0ms, 100ms, 200ms, 300ms, 400ms and 500ms delays; (Brener et al., 1993; Clemens, 1984; Yates et al., 1985)). In some variants, for example the method of constant stimuli, participants are presented with 5-10 tones for each delay on a trial by trial basis asked to make a synchronicity judgement (e.g., 'was the tone in or out of sync with your heartbeat?'). In others, for example the method of adjustment, participants are permitted to sample all the delays and then select which they believe to be synchronous with their heartbeat. Like newer analysis approaches for the heartbeat tapping task (Smith et al., 2021), the consistency of a participants' preferred delay is taken as a measure of accuracy. Whilst evidence supports the use of multi-level tasks – particularly as evidence suggests only moderate correspondence between multi-level tasks and 2AFC tasks underscoring the aforementioned issues with false positives (Brener et al., 1993) - the demands of these tasks in terms of the length of administration means they are rarely employed.

1.2.2. Wider issues with the measurement of interoceptive accuracy

Beyond cardiac interoceptive accuracy specifically, it should be recognised that the current approach to the measurement of interoceptive accuracy may be limited by factors

beyond those specific to the measurement of cardiac interoception. One such issue may be the conceptualisation of interoceptive accuracy as a stable trait. Indeed, most studies examining the relationship between interoceptive accuracy scores and other processes (e.g., emotion) or examining group differences in interoception typically assess interoceptive accuracy scores on a single occasion. This is despite evidence that both heartbeat counting and 2AFC heartbeat detection scores may be influenced by both state and trait factors (Wittkamp et al., 2018). Indeed, a recent study that examined heartbeat counting scores using an experience sampling approach observed that >50% of the variance could be attributed to within-person variability, questioning the conceptualisation of interoception as a stable trait (Höller et al., 2021).

It is also relevant to note that tasks of interoceptive accuracy typically aim to assess the accuracy of participants' perception without the use of external cues. Whilst a perfectly reasonable method for gauging individual differences in perception, in everyday life participants have access to a plethora of information – both internal and external – for gauging interoceptive states. However, ignoring the contribution of exteroceptive information to the estimation of bodily states may not provide a full picture of an individuals' interoceptive profile as experienced in everyday life or their propensity to use interoceptive signals (Murphy, 2024). This is best illustrated by considering sex differences in interoceptive accuracy. Whilst evidence suggests that males outperform females on lab-based tests of interoceptive accuracy (e.g., blood pressure; blood glucose; as well as cardiac interoceptive accuracy; see (F. Prentice et al., 2022)), no sex differences are observed when these assessments are taken outside of the lab (for an overview see (Pennebaker & Roberts, 1992; F. Prentice et al., 2022). This evidence suggests that men and women rely on different cues for gauging internal states, with males relying on more internal information and females relying more on external information; for example, when estimating hunger, males may rely more on physiological signals with females relying more on external cues, such as the time since they

last ate. This evidence suggests that for a full picture of a person's interoceptive profile, an individual's ability to gauge internal states outside of the lab may be informative for understanding whether an individual has developed (useful) compensatory mechanisms for overcoming any difficulties perceiving interoceptive states. Whilst undoubtedly relevant clinically – as we may be most concerned where an individual has no method for estimating important physiological cues for maintaining homeostasis - it is notable that there currently remains no purpose-made task for assessing cardiac interoceptive ability outside of the lab where external cues are available.

The need for remote testing methods of cardiac interoceptive accuracy is not only relevant for assessing the use of internal vs. external cues, but also for overcoming issues in the field regarding the use of small sample sizes and enabling assessment of interoception on multiple occasions, addressing aforementioned issues inherent with the use of single assessments (e.g., neglecting state effects). Whilst a number of methods have been developed to administer cardiac interoceptive accuracy tasks remotely – for example, using a webcam or polar watch to objectively record heartbeats (e.g., (Di Lernia, Finotti, Tsakiris, Riva, & Naber, 2022; Höller et al., 2021) – it should be noted that the accuracy of these approaches has yet to be established, some require access to wearable technology or computers, and to date only the heartbeat counting task has been administered via remote testing means. It is clear that there is a need for new interoception measures that can be administered remotely and, to ensure accessibility, such measures would ideally make use of smartphone technology given evidence that the proportion of individuals who own a smartphone has now overtaken the number of individuals who own a laptop/desktop in some countries (Perrin, 2021).

Finally, it should be acknowledged that the measurement of interoception – particularly cardiac interoceptive accuracy – is further complicated by the influence of physiological signals. Indeed, it has been shown that certain physiological parameters (e.g., blood pressure,

body mass index, heart rate and heart rate variability) may influence performance on certain tasks of cardiac interoceptive accuracy (O'Brien et al., 1998; Rouse et al., 1988); though this influence may be task-specific; see (Knapp-Kline & Kline, 2005). This means that it is difficult to determine whether individual differences in interoceptive accuracy arise from difficulties in the perception of internal signals specifically, or the size (or signalling) of the signal to be perceived (Desmedt et al., 2022, 2023). Whilst the extent to which this is problematic remains a matter of debate – with some arguing that when examining the clinical relevance of interoceptive accuracy for well-being or higher-order cognitive processes the cause(s) of difficulties is perhaps less important to know than whether they experience difficulties or not (e.g., (Murphy, 2024)) – ideally tasks of interoceptive accuracy would be constructed such that they are not confounded by individual differences in physiological parameters.

1.2.3. Measurement of other interoceptive processes

1.2.3.1. Relevance of interoception for well-being and higher-order cognitive processes

The importance of developing new measures for assessing interoception is best illustrated by considering the relevance of interoception for well-being and higher-order cognitive processes. Whilst interoception has been linked to a plethora of individual differences and clinical conditions (Brewer et al., 2021), much evidence for such links comes from the tasks described above that have various limitations. As it is beyond the scope of this thesis to review all conditions and processes in which interoception has been implicated, the following section provides an overview of the conditions and higher-order cognitive processes most relevant to this thesis – including, anxiety, stress, depression, alexithymia and empathy. Further underscoring clinical relevance, in the final section an overview of interventions focusing on aspects of interoception is provided.

1.2.3.2. Conditions linked to atypical interoception

Anxiety is often considered to be the prototypical disorder of interoception, particularly anxiety disorders such as panic that are characterised by physiological symptoms (e.g., breathlessness, racing heart rate) and catastrophizing of physical sensations (Khalsa & Lapidus, 2016). Despite good reason to expect interoceptive differences in anxiety, evidence regarding the relationship between anxiety and interoception is fairly mixed (Brewer et al., 2021). In terms of interoceptive accuracy, most research has focused on cardiac interoceptive accuracy. Whilst both positive and negative associations have been reported in the literature, suggesting individuals with anxiety might have either heightened or dampened sensitivity to cardiac sensations respectively, a recent comprehensive meta-analysis of the relationship between anxiety and cardiac interoceptive accuracy (as measured by the heartbeat counting task and heartbeat detection task) failed to find an association between anxiety and cardiac interoceptive accuracy (Adams et al., 2022). Considering that the null relationship between anxiety and cardiac interoceptive accuracy (as measured by the heartbeat counting task and heartbeat detection task) was not moderated by factors such as the age, gender and clinical diagnoses of participants, the anxiety measures used or the study design, such results cast doubt on theoretical proposals implicating atypical accuracy in the aetiology or progression of anxiety disorders." The extent to which this is due to the measures of cardiac interoceptive accuracy employed that are subject to limitations remains to be established. Indeed, evidence that cardiac interoceptive accuracy training may reduce anxiety (Quadt et al., 2021) is in conflict with evidence suggesting no simple relationship between these factors. One possibility is that training targets aspects beyond interoceptive accuracy. Indeed, despite ambiguity regarding the relationship between cardiac interoceptive accuracy and anxiety, there is evidence to suggest that anxiety may be associated with increased attention to, and negative interpretations of, interoceptive sensations (e.g., (Anderson & Hope, 2009; Palser et

al., 2018; Paulus, 2013; Taylor et al., 1991), as well as discrepancies in metacognitive components of interoception (Garfinkel, Tiley, et al., 2016; Palser et al., 2018). Indeed, there is evidence linking anxiety to poorer awareness of interoceptive sensations (e.g., confidence-accuracy discrepancies; reduced metacognitive insight) as well as some suggestion that anxiety may relate to increased engagement by interoceptive sensations relative to accuracy. Although promising, it is worth noting that the measures used to assess self-reported aspects of interoceptive processing have been subject to various criticisms and show little correspondence with each other making comparisons across studies difficult (Desmedt et al., 2022; Gabriele et al., 2022). The measurement of metacognitive components are also impacted by limited trial numbers where confidence-accuracy relationships are assessed, and concerns regarding self-reported components also shed doubt on metacognitive indices calculated using measures where participants may interpret questionnaires differently (e.g., the Body Perception Questionnaire; (Gabriele et al., 2022)). Overall, with respect to anxiety, atypical or negative interpretations of interoceptive sensations remains the most well established correlate of anxiety, though further research is required to establish whether the absence of a relationship between anxiety and cardiac interoceptive accuracy is due to the measures used, or is truly absent in contrast to theoretical proposals.

In contrast to anxiety, where positive relationships are predicted by theory, *Depression* is often associated with poorer cardiac interoceptive accuracy and weaker engagement by interoceptive signals (Eggart et al., 2019; Harshaw, 2015; Khalsa & Lapidus, 2016; Paulus & Stein, 2010). However, the relationship between depression and cardiac interoceptive accuracy is complex for several reasons. First, depression frequently co-occurs with anxiety (with correlations between depression and anxiety scales often exceeding $r = 0.50$; Clark & Watson, 1991), making it difficult to isolate depression-specific effects on interoception. Second, the relationship may be non-linear: Dunn et al. (2007) found that mild depression was

associated with reduced interoceptive accuracy, while severe depression showed no such relationship, suggesting a potential inverted-U shaped function. Pollatos et al. (2009) similarly reported that the association between depression and interoceptive accuracy was moderated by anxiety levels, with effects only evident in those with low concurrent anxiety. As with anxiety research, these mixed findings may reflect limitations of the heartbeat counting task rather than true absence of interoceptive differences in depression.. Indeed, a meta-analysis examining the relationship between heartbeat counting and depression failed to find an association between the two (Hickman et al., 2020) and very few studies have examined the relationship between depression and cardiac interoceptive accuracy using other measures (for reviews see (Brewer et al., 2021; Eggart et al., 2019). In terms of other factors, again limitations of self-report measures and other measurement issues mean that it is difficult to make strong conclusions regarding the relationship between depression and other facets of interoception, highlighting that further work is required.

Beyond common mental health conditions, interoceptive accuracy has also been linked to stress, both in terms of simple relationships between measures of stress (e.g., heart rate variability, questionnaires, cortisol) as well as when stress is induced experimentally (for a review see (Schulz & Vögele, 2015). Where stress has been induced experimentally, it is notable that the effects differ across interoceptive tasks with stress reducing accuracy on the heartbeat detection task and increasing accuracy on the heartbeat counting task (Schulz et al., 2013). As these tasks index different processes (with only ~4% of the variance shared between the two tasks as demonstrated by meta-analytic work examining this association; (Hickman et al., 2020)), one possibility is that stress enhances attention to interoceptive processes (e.g., increased vigilance; resulting in better counting scores), but negatively impacts exteroceptive processes and/or the integration of interoceptive and exteroceptive signals (e.g., impacts multisensory integration on the heartbeat detection task). In terms of

simple relationships, whilst results are mixed, it is notable that early life stress does not impact cardiac interoceptive accuracy (heartbeat counting) under resting conditions, with associations (e.g., reduced accuracy) only observed under conditions of stress (Schaan et al., 2019). Overall, whilst results are consistent with a relationship between interoceptive accuracy and stress, to date associations have only been explored using classical tasks of cardiac interoceptive accuracy.

Whilst the above conditions have been considered as diagnostic categories, it is worth noting that atypicalities in aspects of interoception have been hypothesised to be a transdiagnostic risk factor for poor mental health (Barrett & Simmons, 2015; Brewer et al., 2021; Murphy, Brewer, et al., 2017). Although interoception has been linked to a variety of fundamental cognitive domains that may increase psychiatric risk, one area of focus has been that of emotion. Indeed, almost every theory of emotion ascribes a fundamental role to the perception of bodily sensations in emotional experience (Critchley & Nagai, 2012; Damasio, 2008; Gendron & Barrett, 2009; James, 1894; Schachter & Singer, 1962; Seth, 2013). What follows from these theoretical predictions is that individuals who have difficulties perceiving interoceptive states (e.g., have poor accuracy) should also present with emotional difficulties. One line of enquiry for testing these predicted relationships has been work examining the relationship between alexithymia – characterised by difficulties identifying and describing one's own emotions - and interoceptive accuracy (Brewer et al., 2016; Herbert et al., 2011; Mul et al., 2018; Murphy, Brewer, et al., 2018; Murphy, Catmur, et al., 2017; NEMIAH & JC, 1976; Shah et al., 2016; Trevisan et al., 2019). However, like other conditions, mixed results have been reported regarding the relationship between alexithymia and interoceptive accuracy that may depend on the measures employed, the domains examined and the distribution of alexithymia scores present in any sample (for reviews see (Brewer et al., 2021; Trevisan et al., 2019). Whilst results are broadly consistent with respect to self-perceived interoceptive

accuracy, with alexithymic individuals' routinely reporting difficulties perceiving interoceptive states on questionnaires (e.g., the Interoceptive Accuracy Questionnaire; (Brewer et al., 2016; Murphy, Brewer, et al., 2019)), results are mixed regarding the relationship between alexithymia and performance on objective tasks of interoceptive accuracy, both within the domain of cardiac interoceptive accuracy and beyond.

However, it is not only the processing of one's own emotions that has been theoretically linked to interoceptive accuracy. Indeed, as evidence suggests one's ability to understand one's own emotions relates to one's ability to recognise emotions in other people (Bird & Cook, 2013; Cook et al., 2013; Heaton et al., 2012), theoretically we would expect to find relationships between one's interoceptive accuracy and sensitivity to others' emotions. Whilst evidence is consistent with this possibility – suggesting associations between the recognition of emotional states in others' and cardiac interoceptive accuracy as assessed using common tasks (e.g., (Terasawa et al., 2014)) – evidence is particularly mixed with respect to empathy (Ainley et al., 2015; Ernst et al., 2013; Fukushima et al., 2011; Mul et al., 2018). Whilst empathy involves aspects such as affect sharing and the ability to focus on the emotions of others (e.g. is separable from personal distress; (Coll et al., 2017)), as the first stage of empathy involves recognising the emotional state of the other, it would be predicted that individuals who have difficulties recognising interoceptive states in themselves would have difficulties recognising emotions in others which in turn should impact on their ability to display empathetic concern. Whilst mixed results may be due to poor measurement of cardiac interoceptive accuracy, it is worth noting that the measurement of empathy may be problematic, conflating empathy with emotion recognition and personal distress contributing to mixed results. As such, further work is required to test these theoretical proposals using refined measures of interoceptive accuracy and reliable measures of empathy.

1.2.4. Interoceptive interventions

Given links between interoception and well-being (e.g., mental health conditions, stress), there has been much focus on whether training targeting aspects of interoception (accuracy, attention, interpretation) can improve interoceptive ability, having positive consequences for well-being. Whilst very recent evidence has focused on training of cardiac interoceptive accuracy, suggesting that such feedback training alongside educational resources may improve anxiety (Quadt et al., 2021), notably much of the work examining interoceptive training has focused on respiratory signals given the relevance of these signals for asthma (e.g., for health monitoring) and panic (e.g., interoceptive exposure therapy), as well as their use in various forms of relaxation therapies (e.g., mindfulness, meditation). As interventions reported in this thesis follow this tradition of utilising respiratory signals, below a brief overview of interoceptive interventions focusing on respiratory processing is provided.

Accurate perception of respiratory symptoms is critical for self-initiated asthma control methods, with both poor and heightened sensitivity to bronchoconstriction associated with negative outcomes (e.g., lack of appropriate use of mitigation strategies leading to life threatening episodes and over use of mediations such as rescue inhalers that have negative side effects, respectively; e.g., (Kifle et al., 1997)). As such, much research has focused on whether training can improve perception of respiratory sensations in asthma patients. One early example of this is respiratory load training, where patients are asked to breathe through devices of varying resistance and are provided with feedback regarding the accuracy of their perception. Whilst variants of this protocol has been shown to improve accuracy of perception in asthma patients (Harver, 1994) and in later work was shown to improve quality of life, it is notable that feedback training may have negative consequences, increasing the amount of self-reported asthma triggers, which is associated with a lower quality of life (Janssens & Harver, 2015). Such work suggests that whilst feedback training may improve accuracy of

perception (e.g., detection of relationships between triggers and symptoms), it may also have negative consequences for well-being.

One possible reason for this is that external respiratory loads may not neatly map to sensations of bronchoconstriction that are internally generated (Mass et al., 1992). Indeed, there is some evidence that respiratory load training does not result in improvements in detection of actual lung function (Fritz et al., 2007). As such, an alternative training method involves patients tracking actual lung function and perceived lung function and noting discrepancies in a daily diary, with improvements observed in the relationship between perceived and actual estimates over the course of the recording period resulting from continued feedback (Schandry et al., 1996; Silverman et al., 1990). However, like respiratory load training, this type of training has also been associated with negative consequences as a result of heightened perception (e.g., overuse of medication; (J. M. Feldman et al., 2012)). Overall, evidence suggests both positive and negative consequences of interoceptive training targeted solely on interoceptive accuracy.

A further example of interoceptive training comes in the form of interoceptive exposure techniques, typically CO2 inhalation, lactate infusions, breathing through a straw or exercise that induce physiological sensations such as increased breathing rate (Barlow, 2017; Boettcher et al., 2016; Griez & van den Hout, 1983). Such treatments, which have been applied beyond the treatment of panic attacks to other forms of anxiety (e.g., agoraphobia) using a variety of specific techniques to induce sensations, have a long history of improving panic symptoms in patients (Arntz, 2002; Beck et al., 1997). Notably, in contrast to interoceptive training in asthma (that aims to improve interoceptive accuracy), interoceptive exposure is predominately based on classical conditioning techniques where the aim of interoceptive exposure is to promote learning of a safe contingency between physiological sensations and panic that competes with the learnt fear association (Craske et al., 2014;

Sherman, 1972). In this way, interoceptive exposure therapy may be best thought of as a technique that aims to reduce sensitivity to interoceptive sensations (possibility via both accuracy and attentional mechanisms) and reduce associations between the presence of physical sensations and negative emotional responses. Notably these techniques are often implemented alongside treatments that aim to target maladaptive interpretations of interoceptive sensations (e.g., Cognitive Behavioural Therapy; (Craske et al., 1994; Telch et al., 1993)), underscoring that it is not improving interoceptive accuracy *per se* that is most beneficial for all therapeutic endeavours, and that targeting maladaptive attention and interpretation may be more useful for certain conditions.

A focus on training aspects of interoception beyond interoceptive accuracy to improve well-being can also be well-illustrated by the benefits of meditation, mindfulness and yoga practices that promote passive observation of bodily sensations (e.g., breathing). Mindfulness practice generally involves promoting moment-to-moment non-judgemental attention to bodily processes such as breathing to cultivate sustained attention and has been shown to promote well-being and reduce stress (Kabat-Zinn, 1982; Vaitl, 2004). However, the interoceptive mechanisms by which mindfulness training improves well-being remains unclear. Whilst it has been argued that mindfulness training does not result in improvements in interoceptive accuracy, such assertions are based on evidence that training does not improve cardiac interoceptive accuracy (Khalsa et al., 2020). Indeed, as interoceptive accuracy dissociates across domains (Ferentzi et al., 2018; Garfinkel, Manassei, et al., 2016), such results may not be surprising. Whilst only examined by one study, evidence suggests that respiratory interoceptive accuracy may be improved via mindfulness training (Daubenmier et al., 2013) and further evidence suggests some preliminary evidence that yoga practice may improve cardiac interoception (as assessed via heartbeat counting; (Schillings et al., 2021)). Evidence is also consistent with the observation that such practices may improve confidence

in one’s interoceptive ability (Khalsa et al., 2008). As such, whilst the mechanisms by which such training result in improvements remain unclear, the above evidence converges to suggest that training techniques that target interoceptive processes, particularly when combined with non-judgemental observation or therapies that target maladaptive interpretations of signals (e.g., cognitive behavioural therapy), may be particularly useful for enhancing well-being.

1.3. Interoception and Biometric Feedback in Digital Therapeutics

Given the evidence reviewed above regarding the role of atypical interoceptive processing across multiple mental health conditions, a critical question emerges: how might individual differences in interoceptive abilities influence the effectiveness of digital therapeutic interventions that rely on biometric feedback? This question has particular relevance given the rapid proliferation of smartphone applications and wearable devices that provide users with real-time physiological data as a means of supporting mental health and well-being.

1.3.1. The Gap Between Biometric Data and Therapeutic Benefit

Digital therapeutics incorporating biometric feedback operate on the implicit assumption that providing users with objective physiological information will enhance self-awareness and facilitate improved self-regulation. However, this assumption overlooks a fundamental distinction between passive biometric monitoring and the active interoceptive processing required to derive therapeutic benefit from such information. While wearable devices can continuously track parameters such as heart rate variability, respiratory patterns, and sleep metrics, translating these data into meaningful insights about one's internal state requires the very interoceptive abilities that may be compromised in individuals seeking mental health support.

This presents a potential therapeutic paradox: the populations most likely to benefit from digital mental health interventions — those experiencing anxiety, depression, or other

conditions for which disrupted interoceptive processing has been hypothesised as a candidate contributing mechanism (Brewer et al., 2021; Khalsa et al., 2018), though the empirical evidence remains mixed (Adams et al., 2022) — may be less well placed to effectively utilise biometric feedback approaches. The informal feedback received during the studies reported in this thesis, where participants indicated they "could not fully reconcile the biometric data provided by the applications with their emotional states," illustrates this fundamental challenge in translating objective physiological measurements into subjective therapeutic understanding.

1.3.2. Individual Differences as Moderators of Intervention Effectiveness

The substantial individual differences in interoceptive abilities documented in the literature suggest that biometric feedback interventions are unlikely to be uniformly effective across all users. Individuals with accurate interoceptive processing may readily integrate external biometric data with their ongoing awareness of internal states, using such information to enhance self-regulation and emotional awareness. Conversely, those with disrupted interoceptive processing may find biometric feedback confusing, overwhelming, or simply irrelevant to their subjective experience.

This variability has important implications for intervention design and personalization. Current biometric feedback applications typically adopt a "one-size-fits-all" approach, providing standardized physiological feedback without considering users' underlying interoceptive abilities. A more effective approach might involve: (1) preliminary assessment of users' interoceptive profiles to inform intervention selection, (2) incorporation of interoceptive training components to enhance users' capacity to benefit from biometric feedback, or (3) development of alternative therapeutic approaches for individuals with significantly compromised interoceptive processing.

1.3.3. The Need for Ecologically Valid Interoception Measurement

Understanding the relationship between interoceptive abilities and biometric feedback effectiveness requires reliable, accessible measures of interoceptive processing that can be deployed in real-world settings. Traditional laboratory-based interoceptive tasks, while valuable for research purposes, may not adequately capture the interoceptive demands of engaging with biometric feedback during daily life. The development of smartphone-based interoception measures therefore represents a crucial step toward understanding how interoceptive abilities influence digital therapeutic outcomes.

Such measures could enable researchers to investigate whether specific dimensions of interoceptive processing—accuracy, attention, or evaluation—are differentially important for biometric feedback effectiveness. They could also inform the development of more personalized digital therapeutics that adapt their approach based on users' interoceptive profiles, potentially improving therapeutic outcomes while reducing the frustration and disengagement currently observed in users who struggle to connect biometric data with their subjective experiences.

1.3.4. Toward Interoception-Informed Digital Therapeutics

The progression of research reported in this thesis—from observing user difficulties with biometric feedback to developing novel measures of interoceptive processing—reflects a broader need in the digital therapeutics field to ground intervention design in established psychological and neuroscientific principles. Rather than assuming that more physiological data necessarily leads to better therapeutic outcomes, future digital therapeutics should consider how such data can be meaningfully integrated with users' existing interoceptive abilities and limitations.

This approach could inform the development of adaptive interventions that modify their biometric feedback strategies based on users' demonstrated interoceptive capacities,

potentially offering simplified physiological information for those with compromised interoceptive processing while providing more detailed biometric feedback for individuals capable of effectively utilizing such information. Understanding these relationships represents a critical step toward developing more effective, personalized digital mental health interventions that account for the fundamental role of interoceptive processing in translating objective physiological data into subjective therapeutic benefit.

1.3.5. Situating Biometric Digital Therapeutics within the MRC Framework for Complex Interventions

A useful integrative lens for the therapeutic studies of Chapters 2–4 is the updated Medical Research Council framework for developing and evaluating complex interventions (Skivington et al., 2021). The framework distinguishes four core elements of complex-intervention research — *development or identification* of the intervention, *feasibility* of delivering and evaluating it, *evaluation* of its effects, and *implementation* into routine practice — and identifies six cross-cutting considerations applicable across all phases: considering context, developing programme theory, engaging stakeholders, identifying key uncertainties, refining the intervention, and weighing economic considerations. Reframing the empirical chapters of this thesis through this lens helps clarify what kind of evidence each chapter contributes (and what it does not), and provides a structured vocabulary for the calibration of conclusions that runs through Chapters 2–4 and the matching summaries in the General Discussion (s. 8.1.2–8.1.4 and s. 8.3.2).

Mapping the empirical chapters to the framework. Chapter 4 (Hemmings et al., 2021) sits clearly within the *development* phase of the MRC framework. The study deployed the person-based approach (Yardley et al., 2015) to ground the ACT module content in qualitative work with adults with diagnosed GAD, and used the resulting prototype in a small (n = 10) pilot to evaluate acceptability and usability — exactly the activities the updated MRC guidance identifies as development-phase work. Chapter 3 (Kawadler et al., 2020) is best characterised as a *feasibility* study: a single-arm evaluation of an existing intervention in a real-world workplace context, addressing questions about delivery in a target population, engagement under real conditions, and acceptability among completers (with the engagement bias documented in the new "Engagement" subsection of Chapter 3 being itself a feasibility finding). Chapter 2 (Ponzo et al., 2020) is best characterised as a *preliminary evaluation*: a randomised controlled trial with a waitlist comparator, designed to provide initial evidence of effects on self-reported anxiety and wellbeing in university students. The decisions taken in Chapter 2 — describing the analysis as modified intention-to-treat (s. 2.4.4.2 and Appendix E), reframing outcomes as preliminary indicative evidence (s. 2.8), and repositioning BioBase as a candidate treatment adjunct (s. 2.8 and s. 8.1.2.1) — are all consistent with the cautious early-evaluation framing that the MRC framework recommends ahead of the *implementation* phase.

Cross-cutting considerations. Several of the six cross-cutting considerations in Skivington et al. (2021) bear directly on this thesis. *Context* was an explicit determinant of intervention configuration across the empirical chapters (university students, working adults under

elevated stress, adults with diagnosed GAD) and shaped both the choice of content scaffold (e.g., the Job Demands–Resources model in the workplace study; Bakker & Demerouti, 2007) and the recruitment context (the industry-academic AXA partnership reflected on in s. 3.7.3). *Programme theory* — how the intervention is expected to produce its effects — was articulated post-hoc rather than pre-specified for Chapters 2 and 3, and is positioned in s. 2.8 (Chapter 2 Conclusions) as a candidate mechanism rather than an established active ingredient. *Stakeholder engagement* in the form of the person-based approach was used in Chapter 4 but not in Chapters 2–3; the General Discussion subsection on user-involvement methodology (s. 8.5.4) examines the implications of this asymmetry and considers experience-based co-design as an alternative for future work. *Key uncertainties* — about which components drive observed effects, about how BioBase performs against active comparators, and about whether interoceptive processing moderates response to biometric feedback — are catalogued in s. 8.4 and motivate the measurement work of Chapters 5–7.

An honest note on chronological mismatch. The empirical work was conducted in the order Chapter 2 (2020) → Chapter 3 (2020) → Chapter 4 (2021), with the preliminary evaluation completed before the structured development work. In MRC terms this is an inverted sequence: the framework recommends development first, then feasibility, then evaluation. The inversion reflects the practical conditions in which the work was conducted — I joined ongoing intervention-development streams within the BioBeats partnership, where the BioBase platform had already been deployed at the point at which the formal empirical thesis work began (see s. 3.7.3 for the chapter-level reflection on this collaboration) — and is itself part of the thesis's learning trajectory. The introduction of the person-based approach in Chapter 4, rather than earlier, is a direct consequence of recognising during the early empirical work that structured user-centred development was a gap in the platform's existing design pipeline. Reframing this trajectory through the MRC lens does not retrospectively impose a coherence it did not have at the time; it makes explicit the developmental learning that produced the measurement studies of Chapters 5–7.

1.3.6. Rationale for the intervention configurations evaluated in Chapters 2–4

The three therapeutic studies in this thesis (Chapters 2, 3 and 4) employed two specific intervention configurations: (i) the BioBase programme, combining brief CBT- and mindfulness-based modules with passive biometric feedback from a wrist-worn wearable (heart rate variability, activity, sleep) and ecological momentary mood declarations, evaluated in a student population (Chapter 2, randomised controlled trial) and a workplace population (Chapter 3, single-arm feasibility); and (ii) a six-module Acceptance and Commitment Therapy adaptation of the BioBase platform, evaluated in adults with diagnosed generalised anxiety disorder (Chapter 4, pilot acceptability). The reasons underlying these selections were a combination of theoretical, empirical and pragmatic considerations; honest acknowledgement of the pragmatic component matters here because the intervention configurations were not the outcome of systematic head-to-head comparison.

Theoretical and empirical grounds. CBT- and mindfulness-based modules were chosen for the BioBase content used in Chapters 2 and 3 for three reasons. First, the evidence base for both approaches in non-clinical and subclinical anxiety, stress and student/workplace populations is

substantial, and meta-analytic estimates for digitally delivered variants fall in the small-to-moderate range (e.g., g ≈ 0.80, 95% CI [0.68, 0.93] for digital interventions for anxiety; Pauley et al., 2023). Second, both modalities are structurally well-suited to brief modular digital delivery: short standalone modules with clear behavioural targets — cognitive restructuring, attentional re-orientation, breath-based relaxation, body scan — can be completed in five-minute blocks without requiring concurrent therapist contact. Third, the integration of these therapeutic modules with continuous passive biometric feedback was the central novel contribution of the BioBase platform and aligned directly with this thesis's interest in whether physiological signals can be meaningfully integrated into mental-health interventions (s. 1.3.1–1.3.4). For the workplace context of Chapter 3 specifically, the therapeutic content was further structured around the Job Demands–Resources model (Bakker & Demerouti, 2007), an established framework for understanding occupational stress that provided population-appropriate scaffolding for tailoring module content to workplace stressors.

The choice of ACT (rather than CBT, mindfulness-based stress reduction, behavioural activation, or metacognitive therapy) for the clinical Chapter 4 study is described in detail in s. 4.4.2. In summary, ACT was selected because its emphasis on present-moment bodily awareness aligns directly with this thesis's interoception focus; because its transdiagnostic structure is well-suited to GAD given its high comorbidity with depression; because its metaphor-based, brief-format modules are particularly amenable to digital adaptation; and because published evidence supports digital ACT efficacy in anxiety populations (Kelson et al., 2019).

Pragmatic and contextual factors. The empirical work of this thesis was conducted within an industry-academic collaboration with BioBeats Ltd (subsequently acquired by Huma Therapeutics Ltd; the consequences of this for post-publication data accessibility are discussed in s. 8.3.4). Two pragmatic factors followed. First, the BioBase platform was the digital delivery infrastructure available to the research team; building an alternative biometric-feedback platform from scratch was not feasible within the timeframes of the empirical studies. Second, I joined ongoing intervention-development streams within the partnership: the studies of Chapters 2 and 3 were led by Sonia Ponzo and Jamie Kawadler respectively as first authors, with me as co-author and platform lead, and the Chapter 4 ACT pilot was developed under the leadership of Nicola Hemmings using the person-based approach (s. 8.5.4). Each chapter's intervention configuration was therefore shaped in part by team composition and prior infrastructure decisions, rather than being initiated from a clean slate (see the author-contribution paragraphs at the start of each empirical chapter for my specific role in each study).

Honest limitations of the selection process. The intervention configurations evaluated here were not the outcome of a systematic head-to-head comparison of candidate approaches. CBT-, mindfulness- and ACT-based digital modules, the Job Demands–Resources framework, and biometric feedback were selected because they fit the theoretical, empirical and pragmatic constraints described above; competing approaches — for example behavioural activation, metacognitive therapy, structured psychoeducation without biofeedback, or biofeedback alone without therapeutic content — were not formally evaluated as alternatives. A consequence is that the present studies cannot answer the question of which combination of

components would produce the largest effects in the relevant populations. The multi-component non-specific-effects issue documented across Chapters 2–4 (s. 8.1.2–8.1.4 and s. 8.3.2) is therefore both a methodological limitation of these specific trials and a signal that the field's next priority should be dismantling-design studies isolating individual active ingredients.

1.4. Overview of Studies in this Thesis

The aim of this thesis is to develop novel measures of interoception that can be deployed remotely and examine the efficacy of remote interoceptive training methods to promote well-being. This research journey evolved from initial attempts to use biometric data for mental health interventions toward the recognition that reliable interoceptive measurement was a critical missing component in understanding the relationship between bodily awareness and psychological well-being.

The thesis comprises six empirical studies (Chapters 2-7) that document this progression from biometric-driven therapeutic interventions to the development and validation of novel interoception measures. The early studies (Chapters 2-4) investigated smartphone applications that utilized biometric feedback to support mental health, while the later studies (Chapters 5-7) focused specifically on developing reliable, remotely deployable measures of interoception.

Chapter 2 presents a randomized controlled trial examining the efficacy of the BioBase digital therapeutic platform among university students with elevated anxiety and stress (Ponzo et al., 2020). This study utilized a smartphone application paired with a wearable device (BioBeam) that provided real-time biometric feedback including heart rate variability (HRV) and movement data. The intervention combined psychoeducational content based on cognitive behavioral therapy and mindfulness with personalized biometric insights delivered over four weeks.

Chapter 3 reports a pilot feasibility and acceptability study of the same BioBase intervention in working adults experiencing workplace stress (Kawadler et al., 2020). This study employed a single-arm design with 55 participants over four weeks, focusing on the real-world implementation of biometric feedback for stress reduction.

Chapter 4 describes the development and pilot testing of a digital acceptance and commitment therapy (ACT) intervention for generalized anxiety disorder (GAD) using the person-based approach (Hemmings et al., 2021). This intervention was delivered through the BioBase platform and included modules targeting the core ACT processes of awareness, openness, and active engagement with values-based behavior. The study employed a two-week intervention period with 10 participants diagnosed with GAD.

Throughout these early therapeutic studies, users frequently reached out via support emails and messages indicating they could not fully reconcile the biometric data provided by the applications with their emotional states, which tried to illustrate autonomous nervous system activity using Heart Rate Variability and circadian rhythm data gathered from the wearable sensor, and to correlate it to potential stress events in their daily lives. This informal feedback suggested that while the interventions were effective, participants struggled to meaningfully interpret physiological signals in relation to their subjective experiences. This pattern became particularly apparent during the ACT study, where present-moment bodily awareness is a core therapeutic component, leading to my hypothesis that measuring interoceptive abilities might be essential for understanding individual differences in therapeutic responsiveness and for developing more targeted interventions.

The recognition that interoception might be a crucial factor in therapeutic effectiveness prompted me to shift toward developing reliable measures of interoceptive ability that could be deployed remotely in clinical populations in subsequent studies.

Chapter 5 presents the development and validation of the CARdiac Elevation Detection (CARED) task, the first novel measure of interoception designed for remote, ecological assessment (Ponzo et al., 2021). The CARED task utilized continuous heart rate monitoring via wearable devices combined with ecological momentary assessment delivered through smartphone notifications over four weeks. Participants received notifications asking whether their heart rate was elevated above normal levels, providing a naturalistic measure of cardiac interoceptive accuracy in real-world contexts.

Chapter 6 reports the development and validation of the Phase Adjustment Task (PAT), a smartphone-based measure designed to overcome the methodological limitations of existing interoception tasks (Plans et al., 2021). The PAT required participants to synchronize audio tones with their perceived heartbeat by adjusting the phase relationship between tones and heartbeats using a simple dial interface. This approach eliminated potential confounds related to heart rate knowledge, counting strategies, and physiological variables that plagued earlier measures.

Chapter 7 presents a pre-registered validation study examining the relationship between the Phase Adjustment Task and a newly developed Phase Selection Task (PST) to establish construct validity. In the PST, 59 participants were presented with tones at three different delays following objective heartbeat detection (0ms, 250ms, and 500ms) and asked to rank the tones according to their perceived synchrony with heartbeats.

Together, these studies document a research pathway that began with the practical challenge of developing effective digital mental health interventions and evolved into fundamental measurement development in interoception. The early therapeutic studies demonstrated that biometric feedback interventions could effectively reduce anxiety and improve well-being, but informal user feedback suggested that effectiveness was potentially limited by

participants' difficulties in connecting biometric data to their subjective emotional experiences. This observation led to the hypothesis that reliable measurement of interoceptive abilities was essential for understanding individual differences in therapeutic responsiveness. The subsequent measurement development studies (Chapters 5-7) addressed this need by creating the first validated, remotely deployable measures of cardiac interoceptive accuracy, providing tools that could potentially be integrated into future digital therapeutic interventions to enable more rigorous investigation of the role of interoception in mental health and well-being.

A note on the central premise. A central interest of this thesis is the proposition that atypical interoceptive processing may contribute to mental health difficulties, and this interest motivates both the digital therapeutic work of Chapters 2–4 and the measurement work of Chapters 5–7. It is important to be explicit about the evidential status of this premise. The empirical evidence linking interoception to anxiety and depression is mixed: a recent meta-analysis (Adams et al., 2022) found no association between anxiety and cardiac interoceptive accuracy as measured by the heartbeat counting and heartbeat detection tasks ($r = .02$), and that null relationship was robust to moderation by age, gender, clinical diagnosis, anxiety measure or study design. Where positive associations have been observed for depression and for other conditions, results are heterogeneous and frequently confounded with comorbid anxiety, methodological variability across interoceptive tasks, and individual differences in body composition and arousal. The thesis therefore treats atypical interoception as a candidate transdiagnostic mechanism worthy of continued investigation, rather than as an established characteristic of anxiety, depression or other mental health conditions. The measurement work in Chapters 5–7 was designed in part to make this empirical question more tractable, by developing ecologically deployable tasks that are less subject to some of the methodological limitations of the heartbeat counting and detection tasks that drove the Adams et al. (2022)

null. References to "disrupted" or "atypical" interoceptive processing elsewhere in this thesis should be read against this calibration.

2. Chapter Two: Digital Biometric Feedback for Student Anxiety: A Randomized Controlled Trial

Content adapted from: Ponzo S, Morelli D, Kawadler JM, Hemmings NR, Bird G, ***Plans D (****2020). Efficacy of the Digital Therapeutic Mobile App BioBase to Reduce Stress and Improve Mental Well-Being Among University Students: Randomized Controlled Trial. JMIR Mhealth Uhealth. https://doi.org/10.2196/17767*

A note on contributions. This study was published with Sonia Ponzo as first author (Ponzo et al., 2020), reflecting her lead on the writing. My role spanned the trial's technical infrastructure — I designed and built the BioBeam wearable and contributed to the programming of the BioBase application and its therapeutic modules — and extended to the statistical analyses and to the drafting and revision of the manuscript.

Abstract

University students in the United Kingdom experience increasing levels of anxiety, yet many do not access traditional mental health services due to barriers including stigma, cost, and availability. This randomized controlled trial evaluated the efficacy of BioBase, a 4-week digital therapeutic intervention combining biometric feedback with psychoeducational content, for reducing anxiety and improving well-being in university students with elevated stress and anxiety. Two hundred and sixty-two students (aged 18-25) with elevated anxiety (DASS-21 anxiety >7) or stress (DASS-21 stress >14) scores were randomized to either the BioBase intervention (n=130) or waitlist control (n=132). The intervention comprised a smartphone application paired with a wearable device providing real-time biometric feedback including heart rate variability and movement data, combined with daily psychoeducational modules based on cognitive behavioral therapy and mindfulness principles. Linear mixed models revealed significant group × time interactions for the primary outcome of anxiety (STAI-S-6) at 4 weeks (estimate = 7.09, 95% CI 2.94-11.24, p = .001, d = 0.67) and at 6-week follow-up (estimate = 4.56, 95% CI 0.49-8.63, p = .03, d = 0.81), with the intervention group showing significant reductions while the control group remained unchanged. Secondary

outcomes of well-being (WEMWBS) and depression (PHQ-9) also improved significantly in the intervention group. These findings demonstrate that digital interventions combining biometric feedback with therapeutic content can effectively reduce anxiety and improve well-being in university students, with effects sustained at 2-week follow-up, supporting the potential of such approaches for addressing the mental health crisis in higher education.

Introduction

Building on the theoretical framework establishing the role of interoception in anxiety (Section 1.2.3.2) and the potential of biometric feedback interventions (Section 1.3), this study examines whether the BioBase digital therapeutic can effectively reduce anxiety in university students. While this population experiences high anxiety prevalence with only 25-36% seeking treatment (Blanco et al., 2008; Eisenberg et al., 2009, 2011), digital interventions offer a promising approach to overcome barriers including stigma (Cardwell et al., 2013) and preference for self-management (Czyz et al., 2013). As discussed in Section 1.3, digital therapeutics incorporating biometric feedback represent a novel approach to mental health intervention. The effectiveness of such interventions may depend on users' ability to perceive and interpret the bodily signals that underlie emotional experiences—a capacity central to the concept of interoception outlined in this thesis. Studies calling for more timely and preventative therapeutic interventions have highlighted the need for digital interventions (Lungu & Sun, 2016; J. L. Prentice & Dobson, 2014; Randall & Bewick, 2016), such as internet-based self-help resources and mobile apps, which have been on the rise in the past decade due to their increased accessibility, availability, and anonymity (Donker, Batterham, et al., 2013; Huberty et al., n.d.; Kählke et al., 2019; Mak et al., 2015; Mowbray et al., 2006), as well as their cost-effectiveness (Herrman et al., 2004). Owing to the widespread use of mobile phones, mobile apps could constitute effective therapeutic support for periods

when students are away from the university, as well as increasing the capacity of on-site counseling services (P. Brown, 2016; Price et al., 2014).

Mobile apps, paired with biosensors and wearable devices, are also effective in gathering passive data (eg, physical activity (He & Li, 2013)) and self-report measures (eg, mood journaling). Accordingly, apps are increasingly used as a real-time monitoring tool, with personalized feedback, insights, and therapeutic content offered to users within the context of mental health interventions (Donker, Petrie, et al., 2013) and illness prevention (Ebert et al., 2017). A number of these digital interventions have proven effective in treating a variety of mental health disorders, ranging from anxiety and depression, to substance use disorder (Waltz, 2018). For example, an intervention lasting 2 weeks comprising brief, daily conversations and mood tracking with a Cognitive Behavioral Therapy (CBT)-oriented conversational agent (Woebot) found that, in comparison with an information-based digital control group, those in the Woebot group significantly reduced their symptoms of depression, while participants in both groups showed significantly reduced levels of anxiety (Fitzpatrick et al., 2017). Furthermore, an 8-week intervention in US university students with the mobile-app Calm was found to produce a significantly greater degree of stress reduction than that seen in a waitlist control group (Huberty et al., 2019). Despite these promising results, studies investigating the efficacy of a combined intervention, including both passive data collection and active therapeutic content, are still lacking. The app BioBase (BioBeats, Ltd) aims at increasing individuals' well-being by combining elements of mindfulness, biofeedback interventions (such as diaphragmatic breathing exercises), CBT, and behavioral activation theory [33-35]. Specifically, its psychoeducational content is based on the job demands-resources model, which has been shown to be associated with students' well-being and stress management [36]. Alongside therapeutic content, data on physical activity, sleep quality, and heart rate are collected via a wrist-worn wearable device (BioBeam) and made available to individuals using the app to foster an increased awareness of users' current well-being. Furthermore, available in-app tools include an ecological momentary assessment tool based on the Circumplex Model of Affect [37], allowing individuals to log their mood in the moment, and reflect back on their entries at a later date to gain insights into longer-term patterns of emotion. The app also includes diaphragmatic breathing exercises and relaxation techniques for in-the-moment stress reduction. The purpose of this study was to test the efficacy of a 4-week intervention delivered via a mobile app and wearable device (ie, the BioBase program) in comparison with a waitlist control group on anxiety and general mental well-being in university students with elevated anxiety or stress. The study also examined sustained effects (at 6 weeks from baseline) of the intervention on anxiety and well-being. Finally, in the current study, measures of depression were collected to investigate the impact of the BioBase program on depressive symptoms.

We hypothesized that university students in the intervention group, but not in the waitlist control, would have significant improvements in anxiety and well-being following a 4-week intervention with BioBase. We also predicted that anxiety and well-being would have sustained effects in the intervention group, but not in the waitlist control, at 2 weeks following the end of the intervention. Furthermore, it was hypothesized that being enrolled in the BioBase program would reduce depressive symptoms after 4 weeks of usage.

Methods

Ethics Approval

This study was approved by an Institutional Ethics Committee at the University of Exeter (UEBS Research Ethics Committee, ethics application number: eUEBS002252). All participants provided informed, electronic consent prior to their enrollment in the study.

Pre-registration

The trial was pre-registered on the Open Science Framework (OSF; registration ID 2zd45; https://osf.io/2zd45/) prior to participant enrolment. The pre-registration sets out the study design, randomisation and minimisation procedures, primary and secondary outcome measures, and analysis plan, and the pre-registration materials are publicly accessible at the OSF link. Trial registration details are also recorded in the corresponding peer-reviewed publication (Ponzo et al., 2020). The analyses reported in this chapter follow the pre-registered analysis plan; where analyses deviate from the pre-registration, this is noted in the Methods or Results.

The reporting of this trial has been audited against the CONSORT 2025 statement (Hopewell et al., 2025); the completed checklist is presented in Appendix D, indicating the location of each reporting item within this chapter, the items addressed in the underlying peer-reviewed publication (Ponzo et al., 2020), and the items for which retrospective alignment is limited or which were not addressed in the original protocol.

Study Design

The current study was a randomized, waitlist control trial with assessments conducted at baseline, 2 weeks, postintervention (4 weeks), and follow-up (6 weeks). Participants randomly assigned to the intervention group took part in a 4-week well-being intervention (the BioBase program). Those assigned to the waitlist control group received the intervention after 8 weeks.

Recruitment

Participants were recruited using institutional participant pools at different UK universities as well as via social media, mailing lists, and flyers and through university staff. Recruitment took place between October and November 2019, and potential participants were screened for eligibility via a Qualtrics survey. Inclusion criteria comprised being a full-time university student attending a university in the United Kingdom and (1) being aged between 18 and 25 years, (2) having scored >14 points on the Depression, Anxiety and Stress Scale-21 items (DASS-21 (S. H. Lovibond & Lovibond, 1996)) stress subscale or >7 points on the DASS-21 anxiety subscale, (3) owning an iPhone 6 or above, (4) not having any previous psychiatric or neurological conditions, (5) not being pregnant at the time of testing, and (6) being able to read and understand English. Participants were also excluded if they were currently in therapy or were using counseling services. Individuals taking part in the initial screening survey were entered into a lottery to win a UK £50 (US $63.93) Amazon Voucher.

Randomization and Blinding

The original design was devised as a single-blind study; however, due to logistical reasons (ie, clarity of communications between the research team and participants) it was decided to unblind the design. Eligible participants (n=262) were sent a reminder email prompting them to confirm their willingness to take part in the study. A total of 130 participants were randomly assigned to the intervention group and 132 participants were randomly assigned to the waitlist control group based on minimization factors: gender (2 categories: male and female), age (7 categories: 18, 19, 20, 21, 22, 23, 24, and 25 years), DASS-21 anxiety subscale (5 categories: normal, mild, moderate, severe, extremely severe), and DASS-21 stress subscale (5 categories: normal, mild, moderate, severe, extremely severe). DASS-21 categories were used for inclusion (ie, participants scoring within the normal range at screening were excluded from the study and those scoring normal at baseline were excluded from the analysis) and minimization purposes only. The first participant was allocated at

random. Each subsequent participant's group membership was allocated such that their addition to that group would lead to a closer match between the groups according to the minimization factors at screening. The random number list used to create the 2 groups was generated using the R Minirand package. Following randomization, the intervention group received their BioBeams (which are not functional until paired with a registered BioBase account) via post at their selected address.

Intervention Group

After randomization, participants in the intervention group were emailed the first set of questionnaires to complete (Figure 1). At the end of the questionnaires, they were given details on how to download and register on the BioBase app.

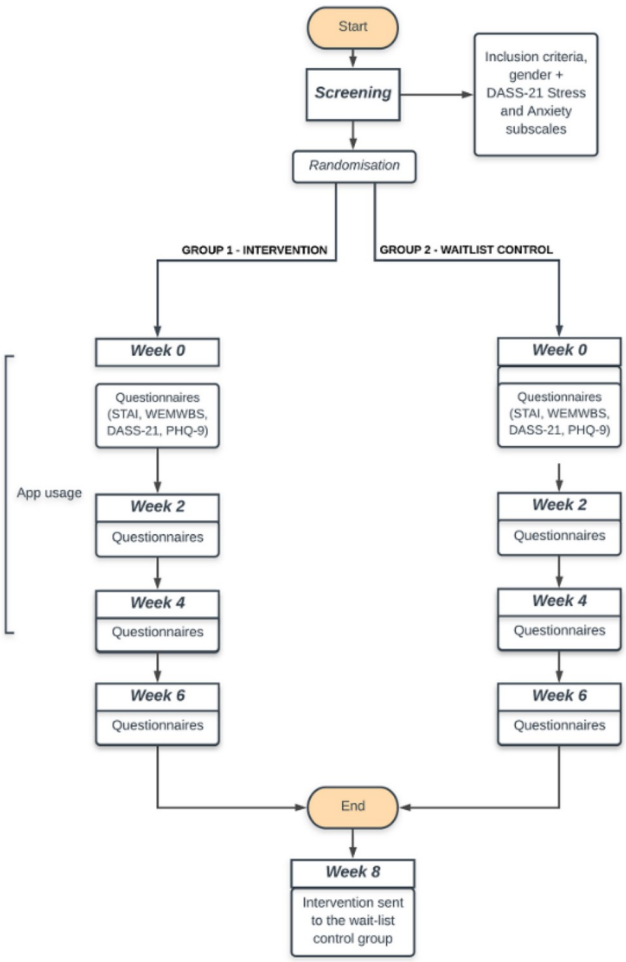


Figure 1. Timeline of the study. Participants were screened for inclusion and exclusion criteria. Of the 805 participants screened, only 262 were eligible to take part. These selected participants were subsequently randomly assigned to one of two groups: intervention or wait-list control. Each group received the questionnaires at baseline, 2, 4 and 6 weeks. The intervention group received access to the program following the first questionnaire completion.

The BioBase program is a multidimensional mobile app comprising psychoeducational content on mental health and well-being, mood tracking (via an ecological momentary assessment, EMA (Shiffman et al., 2008)), and in-the-moment exercises (eg, deep breathing and relaxation techniques). Furthermore, passive data on sleep, heart rate, and physical activity are collected via a wearable device (BioBeam) and presented to the users via a dashboard view.

The psychoeducational content is delivered via 42 five-min long modules, each tackling different aspects of psychological and emotional distress (see Appendix A) for a detailed description of the modules). The content is organized in three different courses, based on the job demands-resources model (A. B. Bakker & Demerouti, 2007). Each course relates to a different aspect of the model (ie, demands, control, and support) and it comprises 14 modules. Demands and control are widely recognized as relevant workplace stressors (Häusser et al., 2010; Karasek, 1979), while social support has been shown to positively impact perceived well-being (Haslam et al., 2018). Embedded in these modules are elements of CBT and self-compassion. Digitally delivered CBT interventions have been proven efficacious in reducing the levels of anxiety and depression (Proudfoot et al., 2004) and similarly, self-compassion has been shown to predict symptom severity in anxious and depressed individuals (Van Dam et al., 2011). By incorporating these therapeutic elements, the courses aim to foster an individual's recognition of internal physiological and emotional processes as a trigger for stress and identify effective coping strategies (eg, setting achievable goals aligned with the individual's personal values).

The EMA tool allows an individual to report their mood in the moment by choosing a mood from a list of options, each with different valence (positive or negative) and arousal (high or low). Furthermore, individuals can specify any ecological component surrounding the moment they chose to declare their mood (ie, where they were, whether they were alone or with somebody, and what activities they were engaged with). EMAs are a valuable mood-tracking tool in the context of digital interventions specifically aimed at reducing levels of anxiety and depression (see (Schueller et al., 2017) for a review).

The deep-breathing tool is designed as a quick intervention aimed at reducing stress and increasing relaxation and consists of 10 guided deep diaphragmatic breaths. Respiration biofeedback has been shown to lead to a reduction of symptoms of depression and anxiety (see (Schoenberg & David, 2014) for a review on results of biofeedback interventions). Similarly, the body scan has been devised as a standalone quick relaxation intervention due to its effectiveness in reducing anxiety and depression (for a detailed review, see (Blanck et al., 2018)).

Finally, passive data collection on physical activity (ie, number of steps performed every 20 seconds), sleep duration and quality (via a triaxial accelerometer with a sample rate of 100 Hz), and heart rate (via a photoplethysmography sensors) was obtained via the BioBeam wearable. This information was made available to participants via an in-app dashboard. Increased sleep awareness and implementation of sleep hygiene techniques have been recognized as a mediating factor in anxiety (Chou et al., 2015), thus supporting the notion that insights into individuals' sleeping patterns may prove beneficial in stress reduction. Furthermore, as physical inactivity is associated with greater levels of anxiety (Stubbs et al., 2017), awareness of and insight into one's own activity patterns may foster improvements in individuals' well-being.

Participants were not prompted to use the app in any specific fashion and were left to freely engage with it for the whole intervention (ie, 4 weeks). Participants were, however, encouraged to continuously wear the BioBeam and engage with the therapeutic content (modules and tools) on a daily basis for at least 5 min. App usage was discontinued after the 4-week intervention ended.

Waitlist Control Group

The waitlist control group received the baseline questionnaire at the same time as the intervention group, followed by an email stating that they would be provided with the app and the wearable device in 8 weeks. Throughout the 8 weeks during which the intervention group used BioBase, the waitlist control participants received the 2-, 4-, and 6-week questionnaires, preceded by a reminder email to complete them. After 8 weeks, participants received a BioBeam at their selected address as well as an email with instructions on how to download and register the app.

Measures and Incentives

Both groups completed four surveys via an online platform (Qualtrics). The surveys consisted of the following questionnaires: the State-Trait Anxiety Inventory (STAI-S-6 (Marteau & Bekker, 1992)), the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS (Tennant et al., 2007)), the DASS-21, and the Patient Health Questionnaire (PHQ-9 (Kroenke et al., 2001)). The DASS-21 Stress and Anxiety Subscales were used as a screening tool for participants' inclusion in the study, whereas the depression subscale, together with the PHQ-9, was used as an outcome measure for depression. Demographic characteristics of the sample were collected at baseline. At the end of the study, each participant received a monetary incentive of £40 (£10 per each completed set of questionnaires at T0, T1, T2, and T3) plus an additional £5 if they decided to send back the wearable device received as part of the intervention.

Primary Outcome

The primary outcomes of the study were responses on the STAI [52]. The STAI-S-6 is a short version of the 10-item state subscale of the STAI. It is a 6-item scale, measuring state anxiety, with responses ranging from 1 (Not at all) to 4 (Very much). Scaled scores are obtained by multiplying the summed responses to each item by 20 and subsequently dividing the score by 6 (range 20-80).

Secondary Outcome

The secondary outcome of the current study was the WEMWBS (Tennant et al., 2007), measuring perceived well-being. WEMWBS is a 14-item scale assessing subjective well-being and psychological functioning. Scoring is obtained by summing each response, ranging from 1 (None of the time) to 5 (All of the time) (range 14-70). WEMWBS has been validated for use in the United Kingdom with those aged 16 and above (Tennant 2007).

Additional Measures

Anxiety and stress were further measured via the DASS-21 subscales to ensure participants were still reporting elevated levels of stress or anxiety at baseline as well as during the screening procedure. Moreover, depression levels were investigated via the DASS-21 Depression subscale and the PHQ-9 questionnaire, a widely employed clinical tool. Although DASS-21 focuses on 1-week periods, PHQ-9 instructs individuals to report changes in the previous 2 weeks. Given that the focus on longer periods may mitigate the effects of random fluctuations in mood, both measures were collected.

Statistical Analysis

Power

The a priori power analysis was based on the within-subjects effect size observed in our group's prior workplace feasibility pilot (Kawadler et al., 2020; reported in Chapter 3 of this thesis), which found a pre-post reduction in STAI anxiety of d = 0.96 over four

weeks of BioBase use. To achieve .95 power to detect this effect at α = .05, the target final sample was 55 participants per group, requiring at least 100 randomised per group (200 in total) on the assumption that approximately 45% of randomised participants would drop out before the four-week endpoint. Two retrospective caveats apply. First, d = 0.96 was derived from a single-arm pre-post comparison, which tends to inflate effect estimates relative to between-groups controlled designs; the comparable controlled meta-analytic estimate for digital interventions in anxiety disorders is g = 0.80 [95% CI 0.68–0.93] (Pauley et al., 2023), with effects in subclinical and student populations typically smaller. A more conservative effect size assumption — for example d ≈ 0.50 — would have indicated a substantially larger required sample. Second, although the 262 randomised participants exceeded the 200-participant target, observed attrition was higher than assumed: 123 of 262 randomised participants completed the T3 follow-up at six weeks, giving an observed dropout rate of approximately 53%.

Data Exclusion

The analytic strategy described below departs from a strict intention-to-treat (ITT) analysis in two respects, both addressed in detail in s. 2.7 and in Appendix E. First, the analytic sample was restricted to participants who provided a baseline (T0) measurement: 146 of the 262 randomised participants met this criterion, with 116 either failing to attend the T0 assessment or scoring below the baseline DASS-21 thresholds. Second, the linear mixed model included the 123 of these baseline-measured participants who provided at least one post-baseline observation. The published analysis is therefore best characterised as a modified intention-to-treat (mITT) analysis on the baseline-measured sample, using maximum-likelihood handling

of post-baseline missingness. Implications and a sensitivity analysis for ITT-equivalent estimates are presented in s. 2.7 and Appendix E.

Given that the inclusion criteria for the current study comprised indication of anxiety or stress (as indexed by DASS-21 Anxiety and Stress subscale scores), 15 participants from the intervention group and 16 from the waitlist control group who initially scored above the normal range at screening but who scored in the normal range at baseline (T0) were excluded from statistical analysis. Participants were further excluded from final analyses if they did not download or open the app during the 4-week intervention as well as if they did not complete all questionnaires.

Data Analysis

The current study employed a mixed design with a between-subjects variable (group) with 2 levels: intervention versus waitlist control and a within-subjects variable (time) with 4 different levels: baseline, 2 weeks, 4 weeks, and 6 weeks. Given the advantages of linear mixed models (LMMs) in dealing with lack of homogeneity of variance and incomplete data sets across time points (Hesselmann, 2018), LMMs were used to analyze our primary and secondary outcomes. Specifically, group and time were the fixed effects and time/subjects were the nested random effects. Planned comparisons (paired-samples t tests) were conducted to explore the direction of significant interactions between group and time. Effect sizes for planned comparisons were calculated using Cohen d (pooled SD) to allow maximum comparability with previous research (Field, 2024). The P values reported later have not been corrected for multiple comparisons but remain significant if corrected. Data were analyzed and plotted using the tydiverse, ggplot2 (Wickham, 2009), lmer4 (Bates et al., 2014), and lmerTest (Kuznetsova et al., 2017) packages for the R statistical language.

Patient and Public Involvement

No formal patient and public involvement (PPI) was undertaken for this study, reported here against the GRIPP-2 Short Form checklist (Staniszewska et al., 2017). Intervention design was led by the BioBeats Ltd research team with academic supervision; the BioBase platform had been developed prior to study commencement and was not iteratively adapted in response to public consultation specific to this trial. Study protocol, outcome measures, and randomisation procedures were determined by the research team without an advisory group of people with lived experience of anxiety or stress. The absence of formal PPI is recognised as a limitation, addressed in the thesis-level reflection in s. 8.5.5 and connected to the experience-based co-design approach proposed for future iterations in s. 9.4.

Results

Participant Enrollment and Demographics

Figure 2 illustrates the flow of participants through the study and reasons for exclusion. Of 805 participants that were screened via an online questionnaire for inclusion and exclusion criteria, 262 participants were deemed eligible and were randomized into either the intervention (n=130) or waitlist control (n=132) groups. Of those, 59 participants from the intervention group and 64 from the control group completed the final questionnaire at T3 and were included in the analysis. Engagement data from the participants in the intervention group showed participants engaged with the app 21.9 of 29 days on average (median 26 days, IQR 13 days, range: 2-29 days). On average, participants engaged with the app 5.33 (SD=5.03) minutes per day (range: 2.13-28.68) over the 29 days of the intervention (see Figures 1 and 2).

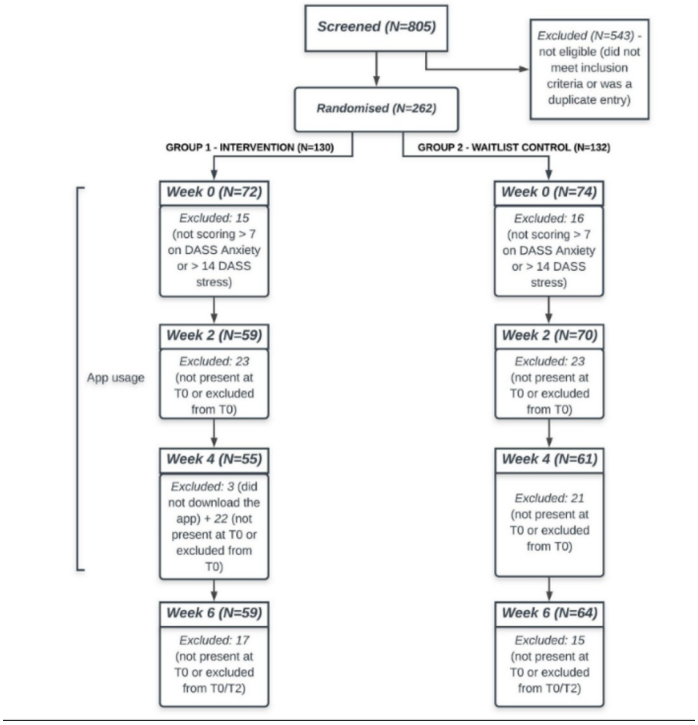


Figure 2. Flowchart of enrollment and retention rates throughout the study in the intervention and wait-list control groups.

However, no correlation was found between the total amount of engagement with the app and differences, from baseline to T2 (4-weeks), in the main outcome measures (see Figure 3 and Figure 4).

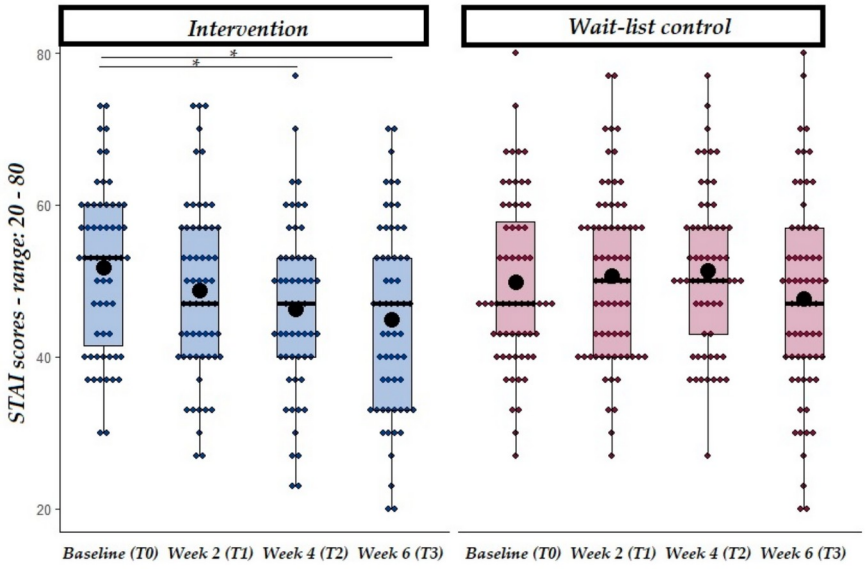


*Figure 3. STAI scores at baseline, week 2, week 4 and week 6 follow-up from the start of the intervention in both intervention and wait-list control groups. Solid line=median; black dot=mean; whiskers: upper whisker=min(max(x), Q_3 + 1.5 x IQR); lower whisker=max(min(x), Q_1 - 1.5 x IQR). *P<.01.*

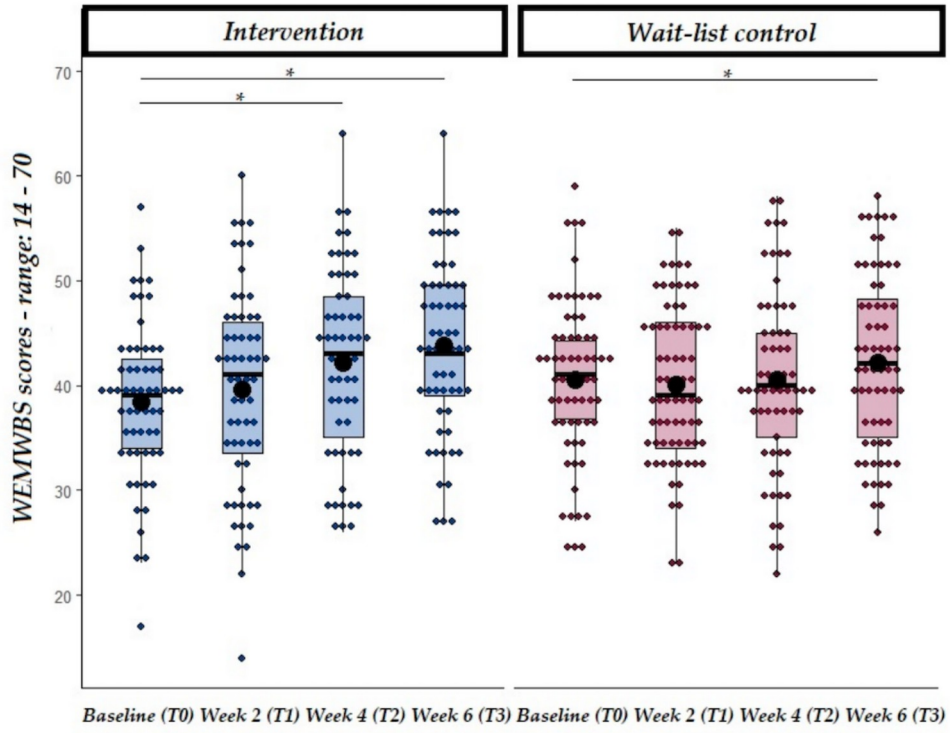


*Figure 4. WEMWBS scores at baseline, week 2, week 4 and follow-up from the start of the intervention in both intervention and wait-list control groups. Solid line=median; black dot=mean; whiskers: upper whisker=min(max(x), Q_3 + 1.5 x IQR); lower whisker=max(min(x), Q_1 - 1.5 x IQR). *P<.05.*

Participants in the two groups did not differ significantly with respect to age, gender, nor their levels of stress and anxiety at baseline (see Table 1). A total of 59 participants from the intervention group and 70 from the control group partook in the second questionnaire (intervention group: 38 females, age range: 18-25 years, mean 19.9 years, SD 1.9; waitlist control: 48 females, age range: 18-25 years, mean 19.9 years, SD 1.89; Figure 2), 55 and 61 (respectively) in the third (intervention group: 36 females, age range: 18-25 years, mean 19.9, SD 1.82; waitlist control: 43 females, age range: 18-25 years, mean 19.93 years, SD 1.95); and finally 59 and 64 participants completed the follow-up questionnaire (intervention group:

38 females, age range: 18-25 years, mean 19.92 years, SD 1.86; waitlist control: 43 females, age range: 18-25 years, mean 20, SD 1.90).

Table 1. Summary of participant characteristics at baseline (T0)

Characteristics	Intervention group (n=72)	Waitlist control (n=74)	*P* value
Gender (n)			
Females	45	47	N/A[a]
Males	27	27	N/A
Age (years), mean (SD)	19.9 (1.83)	19.84 (1.76)	.83
DASS-21[b] anxiety, mean (SD)	15.39 (6.68)	14.46 (7.23)	.42
DASS-21 stress, mean (SD)	21.08 (7.02)	19.86 (7.66)	.32

[a]N/A: not applicable.
[b]DASS-21: Depression, Anxiety, and Stress Scale-21 items.

Primary Outcomes

State-Trait Anxiety Inventory: Baseline to Follow-Up

The primary hypothesis was that, in comparison with the waitlist control group, the intervention group treated with BioBase would show a significant reduction in anxiety levels (measured via STAI-S-6) at the end of the intervention (ie, 4 weeks following baseline measures). Furthermore, it was hypothesized that such effects would be sustained at follow-up (ie, 2 weeks after the end of the intervention). An LMM with STAI-S-6 as the dependent variable and group and time (as well as their interaction) as independent variables was carried out. This analysis revealed a significant main effect of time (at both week 4 and week 6), with scores being lower in comparison with baseline. Furthermore, a significant interaction between group and time (at both week 4 and week 6) on perceived anxiety levels (see Table 2 for a summary of the LMMs) was observed.

To further explore the significant interaction between time and group, planned comparisons were conducted separately in the intervention and waitlist control groups comparing STAI-S-

6 values at baseline with week 4 and follow-up (6 weeks), respectively. Findings revealed that STAI-S-6 at week 4 was significantly lower in the intervention group but not in the control group (see Table 3 for a summary of descriptive statistics and planned comparisons and Figure 3) and that such a reduction was still present at follow-up in the intervention group only.

One of our secondary hypotheses was that, in line with the results from Fitzpatrick and colleagues (Fitzpatrick et al., 2017), the BioBase intervention would show efficacy in decreasing self-reported levels of anxiety after 2 weeks of treatment compared with the control group. However, no interaction between time and group was found at week 2, suggesting that changes in anxiety did not occur within the first 2 weeks of the intervention (Table 2).

Table 2. Summary of the linear-mixed model on State-Trait Anxiety Inventory scores over the 4 time points in the intervention and waitlist control groups

Predictors	STAI-S-6[a]		
	Estimates	95% CI	*P* value
Intercept	53.61	47.07-60.15	<.001
Group	−1.90	−5.99 to 2.19	.36
T1—2 weeks	−6.35	−12.87 to 0.17	.06
T2—4 weeks	−12.25	−18.90 to −5.59	*<.001* [b]
T3—6 weeks	−11.32	−17.84 to −4.81	*.001*
Group: Time T1	3.44	−0.65 to 7.52	.10
Group: Time T2	7.09	2.94 to 11.24	*.001*
Group: Time T3	4.56	0.49 to 8.63	*.03*
Random effects			
σ^2	16.82	N/A[c]	N/A
τ00 Time: Participants ID	49.44	N/A	N/A
τ00 Participants ID	67.20	N/A	N/A
ICC[d]	0.87	N/A	N/A
N time	4	N/A	N/A
N ID	123	N/A	N/A
Observations	491	N/A	N/A
Marginal R^2/conditional R^2	0.037/0.879	N/A	N/A

[a]STAI: State-Trait Anxiety Inventory.
[b]Italicized values are significant.
[c]N/A: not applicable.
[d]ICC: intraclass correlation coefficient.

Table 3. Mean, SD, and planned comparisons on State-Trait Anxiety Inventory scores over the duration of the study (T0, T1, T2, and T3) in the intervention and waitlist control groups

Time point	STAI-S-6[a]: planned comparisons							
	Intervention group				Waitlist control			
	Mean (SD)	*t* test (*df*)	*P* value	Effect size, *d*	Mean (SD)	*t* test (*df*)	*P* value	Effect size, *d*
T0—baseline	51.71 (10.78)	N/A[b]	N/A	N/A	49.81 (10.96)	N/A	N/A	N/A
T1—week 2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
T2—week 4	46.31 (11.32)	3.507 (54)	<.001	0.67	51.33 (10.35)	−1.449 (60)	.15	0.26
T3—week 6	44.95 (12.52)	4.35 (58)	<.001	0.81	47.61 (13.29)	1.542 (63)	.13	0.27

[a]STAI-S-6: State-Trait Anxiety Inventory.
[b]N/A: not applicable.

Secondary Outcomes

Warwick-Edinburgh Mental Well-Being Scale: Baseline (T0)—Follow-Up (T3)

One of the secondary hypotheses was that participants in the intervention group only would report higher levels of well-being (as measured by WEMWBS) at both the end of the intervention and at follow-up. An LMM with WEMWBS as the dependent variable and group and time (as well as their interaction) as the independent variables revealed a significant main effect of time (at both week 4 and week 6), suggesting that perceived well-being increased over time regardless of groups. Furthermore, a significant interaction between group and time (week 4 and week 6) was found (see Table 4), which was further analyzed with planned comparisons. t tests were conducted separately in the intervention and waitlist control group comparing WEMWBS values at baseline (T0) and following the 4-week intervention as well as at follow-up (week 6). Results showed that in the intervention group only, WEMWBS values significantly increased between baseline and week 4, suggesting a higher perceived well-being in the intervention group (see Table 5 and Figure 4). WEMWBS values significantly increased between baseline and follow-up in both groups, but with higher values on average in the intervention group, suggesting an increase in perceived well-being.

Table 4. Summary of the linear-mixed model on Warwick-Edinburgh Mental Well-Being Scale scores over the 4 time points in the intervention and waitlist control groups.

Predictors	WEMWBS[a]		
	Estimates	95% CI	*P* value
Intercept	36.40	31.65 to 41.16	*<.001* [b]
Group	2.07	−0.90 to 5.04	.17
T1—2 weeks	2.69	−1.25 to 6.64	.18
T2—4 weeks	7.41	3.38 to 11.44	*<.001*
T3—6 weeks	8.94	4.99 to 12.88	*<.001*
Group: Time T1	−1.57	−4.05 to 0.90	.21
Group: Time T2	−3.96	−6.47 to −1.45	*.002*
Group: Time T3	−3.65	−6.11 to −1.18	*.004*
Random effects			
σ^2	5.00	N/A[c]	N/A
τ00 Time: Participants ID	19.25	N/A	N/A
τ00 Participants ID	46.28	N/A	N/A
ICC[d]	0.93	N/A	N/A
N time	4	N/A	N/A
N ID	123	N/A	N/A
Observations	533	N/A	N/A
Marginal R^2/conditional R^2	0.029/0.794	N/A	N/A

[a]WEMWBS: Warwick-Edinburgh Mental Well-Being Scale.
[b]Values in italics are significant.
[c]N/A: not applicable.
[d]ICC: intraclass correlation coefficient.

Table 5. Mean, SD, and planned comparisons on Warwick-Edinburgh Mental Well-Being Scale scores over the duration of the study (T0, T1, T2, and T3) in the intervention and waitlist control groups.

Time point	WEMWBS[a]—planned comparisons							
	Intervention group				Waitlist control			
	Mean (SD)	*t* test (*df*)	*P* value	Effect size, *d*	Mean (SD)	*t* test (*df*)	*P* value	Effect size, *d*
T0—Baseline	38.47 (7.54)	N/A[b]	N/A	N/A	40.55 (7.76)	N/A	N/A	N/A
T1—week 2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
T2—week 4	42.15 (9.02)	−3.385 (54)	.001	0.65	40.51 (8.64)	0.814 (62)	.42	0.15
T3—week 6	47.76 (8.31)	−6.260 (58)	<.001	1.16	42.19 (8.37)	−2.127 (63)	.04	0.38

[a]WEMWBS: Warwick-Edinburgh Mental Well-Being Scale.
[b]N/A: not applicable.

Additional Measures

To explore the potential of the BioBase program to reduce depression over the 4-week period of use, depression was measured via the PHQ-9 questionnaire and a linear mixed model with depression scores as the dependent variable and group and time (as well as their interaction)

as the independent variables was carried out. This analysis revealed that depressive symptoms decreased at 4-weeks from the start of the intervention, regardless of groups, but that in the intervention group this effect was more pronounced (as suggested by the significant interaction between group and time at week 4). This significant interaction (Table 6) was further analyzed via planned comparisons on depression scores in the intervention and waitlist control group at baseline and following the 4-weeks intervention. Findings revealed that in the intervention group only, PHQ-9 values significantly decreased between baseline and week 4, suggesting a lower perceived level of depression (see Table 7 and Figure 5). Changes in the Depression subscale of the DASS-21 were also explored. This analysis revealed a main effect of Time at both week 2 and week 4, with depression levels reducing over time irrespective of groups. Although the same pattern highlighted by the PHQ-9 scores was observed (intervention group: Baseline: mean 18.58, SD 10.87; week 4: mean 12.76, SD 8.77; waitlist control: baseline: mean 16.44, SD 9.67; week 4: mean 12.16, SD 8.90), there was no significant interaction between Group and Time. Such finding could be due to the intrinsic characteristics of the scales (ie, DASS-21 focuses on 1-week periods, while PHQ-9 asks individuals to report changes in the previous 2-weeks).

Table 6. Summary of the linear mixed model on Patient Health Questionnaire scores over the duration of the intervention (T0, T1, and T2) in the intervention and waitlist control groups.

Predictors	PHQ-9[a]		
	Estimates	95% CI	*P* value
Intercept	12.65	9.73-15.58	*<.001* [b]
Group	−0.87	−2.70 to 0.95	.35
T 1—2 weeks	−1.34	−3.53 to 0.84	.23
T 2—4 weeks	−4.91	−7.15 to −2.67	*<.001*
Group: Time T1	0.46	−0.90 to 1.83	.51
Group: Time T2	2.07	0.67 to 3.46	*.004*
Random effects			
σ^2	4.25	N/A[c]	N/A
τ00 Time:Participants ID	3.22	N/A	N/A
τ00 Participants ID	19.22	N/A	N/A
ICC[d]	0.84	N/A	N/A
N time	3	N/A	N/A
N ID	123	N/A	N/A
Observations	368	N/A	N/A
Marginal R^2/conditional R^2	0.026/0.845	N/A	N/A

[a]PHQ-9: Patient Health Questionnaire.
[b]Value in italics are significant.
[c]N/A: not applicable.
[d]ICC: intraclass correlation coefficient.

Table 7. Mean, SD, and planned comparisons on Patient Health Questionnaire scores over the duration of the intervention (T0, T1, and T2) in the intervention and waitlist control groups.

Time point	PHQ-9[a]—planned comparisons							
	Intervention group				Waitlist control			
	Mean (SD)	*t* test (*df*)	*P* value	Effect size, *d*	Mean (SD)	*t* test (*df*)	*P* value	Effect size, *d*
T0—baseline	11.78 (5.2)	N/A[b]	N/A	N/A	10.91 (4.93)	N/A	N/A	N/A
T1—week 2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
T2—week 4	8.71 (4.45)	5.139 (54)	<.001	0.99	9.85 (5.38)	1.392 (60)	.17	0.25

[a]PHQ-9: Patient Health Questionnaire.
[b]N/A: not applicable.

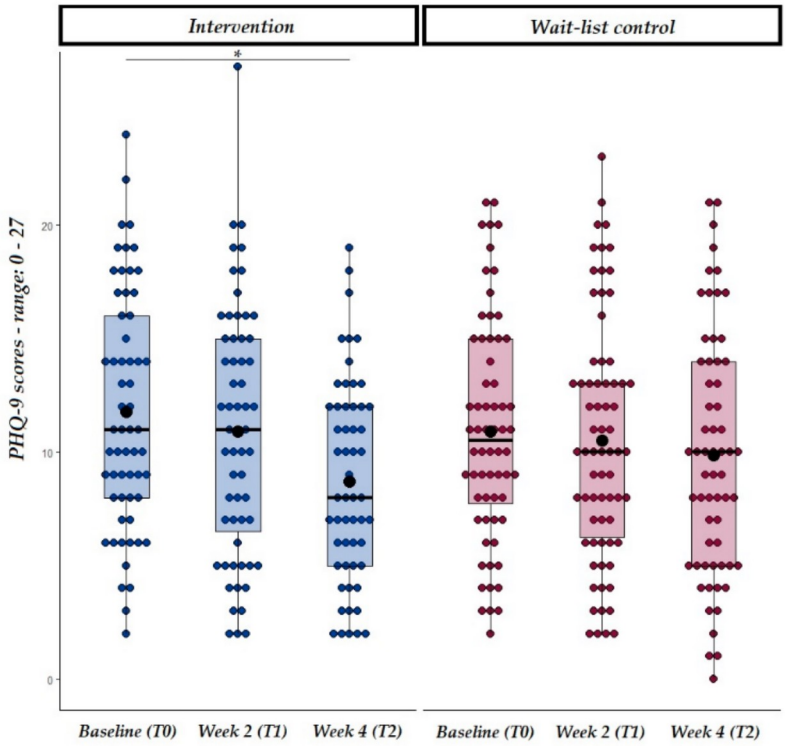


*Figure 5. PHQ-9 scores at baseline and week 2 from the start of the intervention in both intervention and wait-list control groups. Solid line=median; black dot=mean; whiskers: upper whisker=min(max(x), Q_3 + 1.5 x IQR); lower whisker=max(min(x), Q_1 - 1.5 x IQR). *P<.01.*

Discussion

Principal Findings

The aim of the current study was to investigate the efficacy of BioBase, a 4-week app-based intervention, in reducing anxiety and increasing well-being in university students with high self-reported levels of stress or anxiety. Results revealed that using the BioBase program for 4 weeks led to reduced self-reported anxiety and increased self-reported well-being. Such results were sustained at follow-up, with participants in the intervention group maintaining lower levels of self-reported anxiety and higher levels of well-being at 6 weeks from the study start date. Effect sizes ranged from moderate to large throughout the different outcomes.

Comparison With Prior Work

The primary hypotheses of the current study were that in the intervention group only, levels of anxiety would decrease following enrollment in the BioBase program and that this reduction would be sustained after 2 weeks from the end of the intervention. In line with our first primary hypothesis, we found that self-reported levels of anxiety were significantly reduced in the intervention group after 4 weeks of app usage. This finding is in line with results from previous studies using digital interventions in both student (Huberty et al., 2019) and nonstudent (Fitzpatrick et al., 2017)populations. As mentioned in the Introduction section, Huberty et al (2019) found that the mobile app Calm, consisting of a guided mindfulness meditation program, was effective in reducing stress levels among university students. In contrast to the BioBase 4-week program of 5 min a day; however, the Calm intervention was an 8-week program, requiring participants to first complete a 1-week course and then actively engage with the therapeutic content for at least 10 min a day. The efficacy of the BioBase program despite the reduced dosage may be related to the nature of the BioBase program: the therapeutic content is only one aspect of the hypothesized factors at play in anxiety reduction. Interactions with the app dashboard (showing participants their levels of activity, sleep quality, mood declarations over time, and heart rate), as well as usage of the tools, are hypothesized to be causally efficacious in addition to the traditional therapeutic content. Future studies using BioBase could shed light on the individual contribution of each of these aspects in reducing anxiety levels.

These results are also in line with previous findings (Kawadler et al., 2020), suggesting a significant reduction in anxiety following a 4-week intervention with the BioBase program in a sample of full-time employees. However, in this previous study, the effect of the intervention was not assessed beyond the end of the program. In the current study, we showed

that the effect of the intervention persisted for 2 weeks following the end of the program. This result, in line with previous research (Huberty 2019), highlights the efficacy of mobile apps to reduce stress and anxiety over time, and their potential to supplement existing therapeutic support. Future studies should investigate the extent to which these effects persist over longer timeframes, with the aim of identifying optimal guidelines for engagement to maximize outcomes.

A secondary hypothesis was that reduction in anxiety would be present following 2 weeks of enrollment in the BioBase program in the intervention group (but not in the waitlist control group). However, we did not find evidence of efficacy at 2 weeks. This finding is in contrast with a previous study conducted in the young adult population (Fitzpatrick et al., 2017), using a CBT-based intervention to reduce anxiety and depression, which found significant results following 2-week long interactions with a Web-based conversational agent. Nevertheless, the current study significantly differed in both methods of delivery (app vs Web-based) as well as type of intervention. Although Fitzpatrick and colleagues employed a daily intervention, comprising specific time windows of interaction with the therapeutic content, the current study had a more ecological approach, with the BioBase program being available to participants at all times yet not being a daily commitment. Thus, the reason behind the lack of efficacy following a 2-week enrollment in the program may be due to differences in perceived benefit from the participants' perspective, that is, it may be easier to recognize the impact of a daily conversational intervention versus a natural, progressive engagement with a multidimensional program. Further research, comparing different kinds of interventions, would be needed to shed light on these findings.

In terms of secondary outcomes, it was hypothesized that perceived well-being would increase following a 4-week intervention with BioBase and that this effect would be sustained at follow-up (6 weeks). As predicted, we found that participants in the intervention group

reported higher levels of perceived well-being after 4 weeks, which were still significant at 2 weeks from the end of the intervention. Nevertheless, we also found a main effect of time, with levels of perceived well-being being higher at T2 and T3, regardless of the grouping. Further studies with single- or double-blind designs could investigate the impact of being enrolled in a study on perceived well-being.

Finally, additional measures of depression were obtained via the PHQ-9 questionnaire and DASS-21 Depression subscale to assess the feasibility of the BioBase program in reducing depressive symptoms. Results showed that participants taking part in the current study reported lower depression levels after 4 weeks of BioBase usage and sustained effects at follow-up (as measured via the PHQ-9). Nevertheless, despite showing the same pattern of reduction, the same results were not significant for the DASS-21 Depression subscale. Such a discrepancy may be due to differences in sensitivity of the 2 measures, given the focus on periods of different length, and further research is needed to shed light on these findings. Furthermore, given that the trial was conducted in November 2019 through December 2019, it is possible that the reduction in DASS Depression scores observed in the waitlist control group could be due to changes in university work demands, such as coursework deadlines and exams, over this period.

This result is nonetheless particularly relevant when assessing the lack of engagement of individuals at risk of suicide with established pathways of support. Specifically, the possibility to access a digital mental health intervention, which could be efficacious in reducing depressive symptomatology could, represent a novel approach in students at risk of suicide (Czyz et al., 2013; Thorley, 2017). Future studies should specifically investigate the efficacy of such intervention in a student population with individuals suffering from self-reported depressive symptoms.

Limitations

A limitation of the current study is the lack of a blinding procedure. As mentioned in the Methods section, the current study was an unblinded, randomized controlled trial, with participants in the control group being aware of the fact that they were not currently partaking in the intervention. This was a consequence of the type of control group employed. However, both groups received the same kind of communications and were prompted to respond to the questionnaires in the same way. A targeted standardized email was sent every week, with the timeline of the study and key dates as a reminder to participants. Although these measures reduced the possibility that unblinding could influence the results of the current study, future studies should investigate the extent to which being enrolled in an intervention leads to improvements in anxiety and well-being by employing a single-blind design, with an information-based control group.

Moreover, owing to lack of data on ethnicity, or information on the characteristics of the students underusing mental health services, it was not possible to assess the generalizability of our sample. Further studies should further investigate this, by replicating the current study while controlling for these variables.

Furthermore, in the current study, it was not possible to differentiate the effect of the different components of the BioBase program. Although this is a characteristic of digital interventions (Whitton et al., 2015), future studies should explore what components of the BioBase program are most efficacious for which individuals.

In addition, the current study targeted subclinical levels of anxiety; therefore, participants with a psychiatric diagnosis of anxiety were excluded. This decision was made to explore symptom reduction and well-being increase without the confounding factors of being currently in treatment for anxiety. It could be the case, however, that effect sizes were underestimated if BioBase is more efficacious in participants with higher anxiety levels.

Further research is needed to better understand the potential effects of BioBase in individuals with a clinical diagnosis of anxiety or stress.

In terms of the follow-up measure, the current study employed a 6-weeks follow-up, aimed at investigating the sustained effects of the intervention. However, it should be noted that further research is needed to explore long-lasting effects of the intervention (eg, 8 weeks).

Finally, in the current study, no specific criterion was used with regard to app usage. Given that we wanted to observe how participants would naturally engage and interact with the program, there was no strict indication nor control on participants' way to use the app. Nevertheless, the majority of the sample engaged with the intervention, with only 3 people not downloading or installing the app. Future research could explore whether a more controlled intervention, with specific engagement criteria, could lead to more efficacious results while still maintaining ecological validity.

A further consideration concerns the power assumptions underlying the trial. The d = 0.96 effect size used in the a priori power calculation (s. 2.4.4.1) was derived from a single-arm pre-post feasibility pilot (Kawadler et al., 2020; Chapter 3), which is known to inflate effect estimates relative to controlled comparisons; the pooled meta-analytic estimate for digital interventions in anxiety disorders, based on controlled trials, is g = 0.80 (Pauley et al., 2023), and effects in subclinical and student populations tend to be smaller still. The trial was therefore powered for an optimistic effect. Observed attrition (approximately 53%) also exceeded the assumed dropout rate (approximately 45%), although the higher-than-target number randomised (n = 262 against a target of 200) partly compensated for this in absolute terms. Any future replication should be powered against a more conservative effect size estimate and a higher anticipated attrition rate.

Per-protocol versus intention-to-treat analysis. The analyses reported above departed from a strict intention-to-treat (ITT) analysis as now emphasised in CONSORT 2025 (Hopewell et al., 2025) in two respects. First, the analytic sample was restricted to the 146 of 262 randomised participants who provided a baseline measurement; 116 randomised participants (44%) did not reach the T0 assessment. Second, within the baseline-measured sample, the published linear mixed model included the 123 participants with at least one post-baseline observation, treating the remaining 23 as missing. The published analysis is therefore best characterised as a modified intention-to-treat (mITT) analysis using maximum-likelihood handling of post-baseline missingness, rather than a per-protocol analysis in the strict sense. A formal ITT re-analysis on the full randomised sample (n = 262) was originally planned for the post-viva revisions but could not be performed: the trial dataset is held by Huma Therapeutics Ltd (which acquired BioBeats Ltd in 2021), and is no longer accessible following decommissioning of the company's research infrastructure (s. 8.3). Appendix E presents a sensitivity analysis constructed from the published summary statistics and from the meta-analytic per-protocol-to-ITT bias benchmark of Linardon and Fuller-Tyszkiewicz (2020, $\Delta d = 0.18$ across 70 smartphone mental-health RCTs). Three converging methods — a Linardon-derived bias correction, a worst-case no-change imputation, and a jump-to-reference imputation following Carpenter, Roger, and Kenward (2013) — yield an ITT-equivalent range for the post-intervention effect (T2) of $d \approx 0.47–0.65$ (5.14–7.09 STAI-S-6 points), with the originally published estimate at the upper bound. The corresponding range for the 2-week follow-up effect (T3) is $d \approx 0.17–0.35$ (2.24–4.56 STAI-S-6 points), with the Linardon-corrected

lower bound (d ≈ 0.17) falling below the conventional small-effect threshold. The post-intervention effect at T2 is therefore robust across imputation strategies; the follow-up effect at T3 is less robust and should be interpreted as preliminary indicative evidence of sustained effect, a recalibration argued further under panel point C2-4. The 44% pre-baseline attrition rate is a structural limitation that cannot be addressed by any imputation method without screening-stage data, and is acknowledged in s. 8.3.

Conclusions

This randomised waitlist-controlled trial provides preliminary indicative evidence that a four-week multi-component digital programme combining biometric feedback, brief CBT- and mindfulness-based modules, and ecological momentary mood declarations was associated with reductions in self-reported state anxiety and improvements in self-reported wellbeing in a student population with elevated baseline stress and anxiety. The post-intervention reduction in STAI-S-6 (four weeks) was robust across the intention-to-treat sensitivity analyses reported in Appendix E (ITT-equivalent d ≈ 0.47–0.65); the 6-week follow-up reduction was less robust, with the lower bound of the sensitivity range falling below the conventional threshold for a small effect (d ≈ 0.17), and so the durability of the intervention's effect beyond the intervention period should be interpreted cautiously and treated as a question requiring further investigation rather than a confirmed claim.

Several features of the trial limit the strength of the conclusions that can be drawn. The intervention combined therapeutic content (CBT and mindfulness modules), biometric feedback, and EMA-based mood declarations in a single multi-component package; the present design cannot isolate the active ingredient, and the observed effects may reflect non-specific components common to many supportive digital tools (engagement with structured content, self-monitoring, perceived clinical attention) rather than a specific contribution of biometric feedback (Whitton et al., 2015). The waitlist control did not include an active or sham comparator, and the trial cannot rule out attention or expectation effects attributable to being enrolled in a programme. Pre-baseline attrition was substantial (44% of randomised participants; see Appendix E), and the trial's reporting departs from CONSORT 2025 in respects detailed in Appendix D. The broader evidence base linking interoception, biometric feedback and mental-health outcomes is mixed (Adams et al., 2022; see s. 1.4 and s. 8.2.3 for calibration).

Within these limitations, the findings support continued investigation of multi-component digital programmes of this kind as a candidate adjunct to existing pathways of support — particularly for student populations who under-use on-site mental-health services — rather than as a standalone efficacious treatment. The measurement-development work in Chapters

5–7 of this thesis provides instruments that could, in future work, address one of the most important questions raised by the trial: whether individual differences in interoceptive capacity moderate who benefits from biometric-feedback interventions of this kind, and whether the apparent reduction of effect at follow-up reflects a fade-out characteristic of a particular interoceptive phenotype.

3. Chapter 3: Workplace Stress Reduction Through Biometric Feedback: A Feasibility Study

Content adapted from: Kawadler JM, Hemmings NR, Ponzo S, Morelli D, Bird G, ***Plans D*** *(2020). Effectiveness of a Smartphone App (BioBase) for Reducing Anxiety and Increasing Mental Well-Being: Pilot Feasibility and Acceptability Study. JMIR Mhealth Uhealth. https://doi.org/10.2196/17767*

Chapter overview and framing

This chapter is structured as a feasibility and acceptability study consistent with the framing of the Medical Research Council framework for complex interventions described in s. 1.3.5 (Skivington et al., 2021). Although the published version of this work reports an analysis of pre-post symptom change in anxiety and wellbeing alongside its acceptability and engagement findings, the design constraints — single-arm, no control group, modest sample (n = 55), 4-week intervention — are those of a feasibility-phase study, and the chapter is read here primarily for what it contributes to questions about whether a multi-component biometric digital therapeutic can be delivered, engaged with, and accepted by working adults under elevated stress, rather than as a test of intervention efficacy.

Three research questions structure the chapter:

Feasibility of delivery. Can a 4-week biometric digital therapeutic (BioBase) be delivered to working adults under elevated stress, with adequate retention to support outcome interpretation? Operationalised through recruitment, retention, completion rates, and technical performance.

Acceptability and engagement. Do participants find the intervention acceptable, and to what extent and pattern do they engage with it? Operationalised through user feedback,

recommendation rates, and engagement metrics, with explicit caveats around the paid-versus-unpaid recruitment asymmetry.

Preliminary indicative signal of effect. What pattern, if any, do self-reported anxiety and wellbeing scores show across the intervention period? Given the single-arm design, this question can be addressed only descriptively; the symptom-change findings are positioned as a tertiary preliminary indicative signal that may justify subsequent controlled evaluation, not as efficacy evidence in their own right.

The three research questions correspond to the three core elements that the MRC framework's feasibility phase recommends evaluating prior to formal evaluation: whether the intervention can be delivered, whether it is acceptable, and whether the indicative signal is strong enough to justify continued development (Skivington et al., 2021). The interpretation throughout the chapter follows this priority order: feasibility and acceptability findings are read as substantive evidence about whether this intervention configuration is worth developing further, and the symptom-change findings are read as descriptive signal whose interpretation must be qualified by what the feasibility/acceptability data also show.

Several thesis-level revisions are integrated into this chapter to support this framing. The engagement bias arising from the paid-versus-unpaid recruitment design is documented in a new Discussion subsection ("Engagement and the limits of the acceptability findings"; see C3-2) and connects to broader smartphone-mental-health attrition benchmarks (Linardon & Fuller-Tyszkiewicz, 2020) and to qualitative research on workplace-digital-mental-health engagement barriers (Carolan & de Visser, 2018). A candid reflection on the industry-academic collaboration with AXA that shaped the recruitment context appears in a separate new Discussion subsection ("Reflections on industry-academic collaboration"; see C3-3). The chapter-level summary in the General Discussion (s. 8.1.3.1) reads the present study,

alongside the Chapter 2 student RCT, as feasibility/acceptability evidence supporting continued investigation of digitally delivered workplace-adapted multi-component programmes rather than as confirmation of intervention effectiveness.

A note on contributions. This study was published with Jamie Kawadler as first author (Kawadler et al., 2020), reflecting her lead on the writing. I co-designed the workplace adaptation of the BioBase protocol, engineered the BioBeam wearable for the extended daily-use monitoring this study required, developed the workplace-specific content modules within the application, performed the data processing and linear mixed-model analyses, and contributed to the interpretation of feasibility outcomes and to the drafting and revision of the manuscript.

Abstract

Workplace-related stress and anxiety result in significant physical and mental illness, yet many employees lack access to effective interventions (Joyce et al., 2016). This pilot feasibility study evaluated the BioBase program, a smartphone application with wearable biometric monitoring, for reducing anxiety and improving mental well-being in 55 healthy working adults screened for elevated stress (DASS Stress subscale). The intervention utilized the same BioBase platform described in Chapter 2, adapted with workplace-specific psychoeducational content based on the job demands-resources model (Bakker & Demerouti, 2007). After 4 weeks of program use, participants showed significant reductions in anxiety measured by the State-Trait Anxiety Inventory (baseline M = 45.52, SD = 13.2; 4-week M = 39.82, SD = 11.2; $t_{54} = -3.51$, $p < .001$, Cohen's d = 0.96) and improvements in mental well-being measured by WEMWBS (baseline M = 48.12, SD = 6.4; 4-week M = 50.4, SD = 6.9; $t_{53} = 2.41$, $p = .019$, Cohen's d = 0.66). Participants with higher baseline stress showed significantly greater improvements in both anxiety reduction ($R^2 = 0.180$) and well-being enhancement ($R^2 = 0.101$). Seventy percent of participants would recommend the program to

a friend, and the majority found the content relevant to workplace stressors. These findings demonstrate the feasibility and acceptability of biometric feedback interventions in workplace settings, replicating the pattern observed in university students while highlighting the potential for population-specific adaptations of digital mental health interventions.

Introduction

The workplace can be a major cause of anxiety that leads to depression and burnout (Bhui et al., 2012; Joyce et al., 2016) as well as stress-related illnesses (Chandola et al., 2006; Hammer & Sauter, 2013). In the UK, 12.8 million working days were lost in 2018-2019 due to mental ill health, equating to 21.1 days lost per year per working individual (The Stationery Office, 2019).

Interventions in the workplace may be able to reduce the effects of anxiety or facilitate the recovery of employees with anxiety (Joyce et al. 2016). Structured and targeted psychological interventions have been found to be more effective than general stress management training (Cuijpers et al., 2012), and those employees experiencing higher levels of anxiety have been shown to benefit most (Ebert et al., 2016). Smartphone apps offer a practical, scalable, accessible, and cost-effective solution to promoting employee mental health (Carolan & de Visser, 2018; Price et al., 2014). As smartphone use becomes more ubiquitous, a mobile solution to mental health is becoming increasingly more common (Deady et al., 2018). As discussed in Section 1.3 of this thesis, digital interventions that integrate biometric feedback with evidence-based therapeutic content may address barriers to traditional mental health support in occupational settings.

Many therapist-assisted or guided workplace digital interventions have been shown to be effective in reducing work-related stress (Aikens et al., 2014; Heber et al., 2013, 2016; Ly et al., 2014; Ruwaard et al., 2007); however, these suffer from the limited availability and high cost of more traditional forms of support, and therefore are rarely available to the entirety of

the workforce (Ebert et al., 2016). Self-guided digital interventions, which overcome some of these barriers for individuals seeking support, have also shown promising results in reducing stress and anxiety (Proudfoot et al., 2013).

Engaging, innovative, and effective interventions that address specific workplace-related mental health difficulties are still needed. Integrated technology in self-guided interventions offers several advantages that increase engagement, behavior change, and positive psychological outcomes (Howarth et al., 2018), such as personalized mental health/well-being summaries and insights, real-time monitoring of health data from linked biosensors (Donker, Petrie, et al., 2013), and a secure and private means to access confidential support (Anton & Jones, 2017; D. Bakker et al., 2016). Combined interventions that utilize advanced technology for real-time data feedback plus active therapeutic content may yield a synergistic effect to decrease anxious symptoms, especially in those with higher baseline levels of stress and anxiety (Weisel et al., 2018).

Building on the findings from Chapter 2, this study aimed to evaluate the usability, acceptability, feasibility, and preliminary efficacy of BioBase adapted for individuals in full-time employment with varied levels of stress on perceived anxiety and mental well-being. We hypothesized that after 4 weeks of using the programme (1) there would be a significant decrease in anxiety and (2) a significant increase in mental well-being compared to baseline. Further, we hypothesized that (3) higher baseline levels of self-reported stress would be associated with larger decreases in anxiety and increases in well-being.

Methods

Framing note on Methods (thesis-level). The Methods reported below describe a 4-week single-arm pre-post evaluation of the BioBase intervention in n = 55 working adults under elevated stress, using a workplace-adapted version of the platform described in Chapter 2. Within the thesis framing of this chapter (s. 3.1), the outcomes described in the Methods are

read in the following priority order: (i) feasibility outcomes — recruitment yield, retention through the 4-week intervention period, completion rates for the planned assessments, and technical performance of the platform and BioBeam wearable; (ii) acceptability outcomes — user feedback responses, recommendation rates, qualitative free-text responses, and engagement metrics including the paid-versus-unpaid asymmetry now documented as a feasibility finding in the new "Engagement and the limits of the acceptability findings" Discussion subsection; (iii) preliminary indicative signal — pre-post change scores on the State-Trait Anxiety Inventory (STAI), Warwick-Edinburgh Mental Well-Being Scale (WEMWBS), and Patient Health Questionnaire (PHQ-9), interpreted as descriptive signal given the absence of a control group rather than as effect estimates. The reordering is presentational only — the underlying measures and analyses are as published in Kawadler et al. (2020) — but it foregrounds the feasibility and acceptability framing that the design supports.

Pre-registration

This feasibility study was not pre-registered. At the time the work was conducted (recruitment 2018–2019), pre-registration of mHealth feasibility studies was not yet routine practice in the field, and the study was conducted within a commercial research environment in which pre-registration infrastructure was not yet standard. The research questions, outcome measures and analytic approach reported in this chapter correspond to those documented in the published version of the work (Kawadler et al., 2020). The absence of pre-registration is acknowledged as a transparency limitation, with cross-thesis implications discussed in s. 8.3.

BioBase Programme

The BioBase intervention utilized the same platform described in Chapter 2, comprising a smartphone application paired with the BioBeam wearable device. For this workplace population, the therapeutic content was adapted to address occupational stressors based on the

job demands-resources model (A. B. Bakker & Demerouti, 2007) and HSE work stressor dimensions, while maintaining the core features of biometric feedback, psychoeducational modules, and self-monitoring tools. As with the student population study, data on physical activity, sleep quality, and heart rate are collected via its paired wearable device, BioBeam, a device that was designed and built specifically to allow BioBeats to collect high-quality long-term cardiovascular and proprioceptive measurements, and fed back to the individual via a dashboard view in order to increase awareness of the individual's current well-being state and to help them learn how physiological patterns are related to their well-being in parallel with the content of the modules. For this population, the app was augmented to also contain on-demand tools such as an in-built ecological momentary assessment based on the Circumplex Model of Affect (Posner et al., 2005), allowing individuals to log their mood in the moment, and reflect back on their entries at a later date to gain insights into patterns and themes

Recruitment

Paid participants were individually recruited from a recruitment agency specializing in user research and usability testing (People for Research UK) alongside unpaid participants recruited through convenience sampling through a local Facebook campaign. Recruitment of small businesses was also conducted by a local small business network.

Participants were eligible to take part if they were aged between 18 and 65, in full-time employment, had access to an Apple iPhone, and were able to read and understand English. Exclusion criteria were a history of clinical diagnosis of a mental health or neurological disorder, pregnancy, or current engagement in counselling or therapy. Ethical approval was granted by the University of Exeter and informed consent was obtained electronically from all participants.

Procedure

The study lasted 4 weeks. Participants were screened for inclusion and exclusion criteria using an online questionnaire. At baseline, online questionnaires were used to collect demographic information on age, gender, education, and employment categories, as well as self-reported stress from the Stress subscale of the 21-item Depression Anxiety Stress Scale (DASS-21) (S. H. Lovibond & Lovibond, 1996). The stress subscale consists of 7 questions with range 0-21 and relates to the experience of stress symptoms in the previous week. The internal consistency of the scale is good (Cronbach α=.85) and construct validity measured in previous studies is also good (Henry & Crawford, 2005).

At baseline and at 4 weeks, questionnaires included the 6-item State-Trait Anxiety Inventory (STAI; (Marteau & Bekker, 1992)) and the Warwick-Edinburgh Mental Well-being Scale (WEMWBS; (Tennant et al., 2007)). The 6-item STAI measures state anxiety, with responses ranging from 1 (Not at all) to 4 (Very much). Scaled scores are obtained by multiplying the summed responses to each item by 20 and subsequently dividing the score by 6 (range 20-80). The STAI is widely used and has a good internal consistency (α=.92) (Marteau & Bekker, 1992).

The WEMWBS is a 14-item scale assessing subjective mental well-being. Scoring is obtained by summing each response, ranging from 1 (None of the time) to 5 (All of the time) (range 14-70). WEMWBS has been validated for use in the UK with those aged 16 and above with good internal consistency (α=.91) (Tennant et al., 2007).

Participants were asked to engage naturally with the app, but encouraged to continuously wear the BioBeam, use the BioBase app for around 5 minutes per day for 4 weeks, and complete at least 14 daily 3-5-minute-long modules. Measures of app engagement consisted of number of days in which the app was opened (ie, “active” days), total duration of use, number of modules completed, and total tools completed over the 4 weeks.

An additional feedback questionnaire with multiple-choice questions and free-text answers was distributed at the end of the study along with an optional semistructured interview with the researcher to expand on their responses in the feedback form. Quotes from the free text and interviews were given to illustrate insights into the feasibility, usability, and acceptability of the BioBase programme.

Statistical Analysis

Analysis was performed using R version 3.6.0. To test the first hypothesis of change in STAI and WEMWBS scores between baseline and 4 weeks, variables were first tested for a normal distribution using Shapiro–Wilk tests and linear mixed-effects models (R package: “lmer”) with participants as random effects were fitted to the data.

To test the third hypothesis that baseline DASS stress scores would be associated with greater decreases in STAI and greater increases in WEMWBS, baseline DASS stress scores were first transformed using ordered quantiles to follow a normal distribution (method: OPQ, R package: “bestNormalize”). Linear regression models were used to determine association between baseline DASS stress scores and change in STAI and WEMWBS scores. All analyses used paid status as a covariate. P-values were considered statistically significant at $<.05$.

Patient and Public Involvement

No formal patient and public involvement (PPI) was undertaken for this study, reported here against the GRIPP-2 Short Form checklist (Staniszewska et al., 2017). Study design was led by the BioBeats Ltd research team; recruitment was conducted through the BioBase–AXA workplace wellness partnership (see s. 3.7.3 for the full reflection on this industry-academic collaboration), but neither AXA employees nor BioBase users were involved in a formal advisory capacity for study design, outcome selection, or intervention adaptation. The workplace-specific tailoring of the BioBase content drew on the Job Demands–Resources

framework (Bakker & Demerouti, 2007) rather than on direct consultation with workplace stakeholders. The absence of structured workplace-stakeholder consultation is acknowledged; the engagement bias arising from the paid-versus-unpaid recruitment design documented in the C3-2 revisions is itself partly a consequence of this absence. See s. 8.5.5 for the thesis-level reflection.

Results

Framing note on Results (thesis-level). The Results below are presented in the order used in the published paper, leading with baseline characteristics and primary symptom outcomes. The thesis framing of this chapter (s. 3.1) reads the same data in a different priority order: the feasibility findings (retention through the 4-week intervention period; technical performance with no platform-level failures reported; engagement frequency data) and acceptability findings (70% would-recommend rate among 30 of 55 feedback respondents; usability ratings; qualitative themes from open-text responses, including the paid-versus-unpaid engagement asymmetry of $p < .001$) constitute the chapter's primary contribution. The symptom-change findings on STAI and WEMWBS — significant pre-post reductions in anxiety and increases in wellbeing — are reported below for completeness and as the preliminary indicative signal described in s. 3.1, with their interpretation deferred to the Discussion subsections that explicitly address the single-arm design limitation.

Participants

A total of 70 participants were initially recruited; 4 were excluded due to current engagement in counselling/therapy and 1 was excluded because they were employed part-time. The final sample consisted of 55 participants (25 paid and 30 unpaid) as 10 participants did not complete the exit questionnaires. While paid participants were significantly older, there were no significant differences between paid and unpaid participants in terms of gender

distribution; education or employment category; or baseline stress, anxiety, or well-being scores (Table 1).

Table 1. Sample demographics of participants with baseline self-reported stress on DASS (N=55).

Participant demographics	Paid	Unpaid	*t* test (*df*) or chi-square (*df*)	*P* value
Total participants	25 (45)	30 (55)		
Age, mean (SD), years	38.6 (13.3)	31.4 (6.7)	2.45 (54)	.02
Gender			0 (54)[a]	>.99
Female, n (%)	14 (56)	16 (53)		
Male, n (%)	11 (44)	14 (47)		
Education				.35[b]
School (age 16), n (%)	3 (12)	0 (0)		
Sixth form/college (age 18), n (%)	3 (12)	8 (27)		
Some university, n (%)	1 (4)	1 (3)		
2-year degree, n (%)	4 (16)	4 (13)		
4-year degree, n (%)	10 (40)	14 (47)		
More than 4-year degree, n (%)	4 (16)	3 (10)		
Job category[c]				.43[b]
Administrative, n (%)	5 (20)	2 (7)		
Service, n (%)	2 (8)	1 (3)		
Technical, n (%)	1 (4)	4 (14)		
Sales, n (%)	2 (8)	6 (21)		
Professional, n (%)	10 (40)	10 (34)		
Executive, n (%)	5 (20)	6 (21)		
Baseline DASS[d] stress, mean (SD)	13.7 (8.0)	16.7 (9.9)	1.23 (53)	.22
Baseline STAI[e], mean (SD)	43.9 (11.0)	46.9 (14.8)	0.87 (54)	.39
Baseline WEMWBS[f], mean (SD)	48.1 (5.6)	48.2 (7.2)	0.05 (53)	.96

[a]Presented as chi-square (*df*).
[b]Fisher exact test.
[c]N=29 for unpaid.
[d]DASS: Depression Anxiety Stress Scale.
[e]STAI: State-Trait Anxiety Inventory.
[f]WEMWBS: Warwick-Edinburgh Mental Well-being Scale.

Engagement

Participants used the app on average 20.9 days (median 22; IQR 7.5) of 4 weeks, completing an average of 28.4 tools (median 25) and completing an average of 8.9 modules (median 7). The mean total duration of use was 164.3 minutes (median 157.0 minutes).

Paid participants used the app on more days (paid: mean days 23.9; unpaid: mean days 18.4; t54=3.39; P=.014; Cohen d=0.89), completed more tools (paid: average tools 36.8; unpaid:

average tools 21.3; t54=3.12; P=.003; Cohen d=0.86), completed more modules (paid: average modules 13.7; unpaid: average modules 4.9, t54=4.81; P<.001; Cohen d=1.30), and had a longer total duration of use (paid: 210.2 minutes; unpaid: 125.9 minutes; t54=3.20; P=.003; Cohen d=0.88) than unpaid participants.

Outcomes

The first 2 hypotheses were supported. After controlling for paid status, linear mixed models revealed that anxiety as measured by STAI significantly decreased (baseline mean 45.52 [SD 13.2]; 4-week mean 39.82 [SD 11.2]; t54=–3.51; P<.001; CI –8.88 to –2.52; Cohen d=0.96) and mental well-being significantly increased (baseline mean 48.12 [SD 6.4]; 4-week mean 50.4 [SD 6.9]; t53=2.41; P=.019; CI 0.44-4.23; Cohen d=0.66) after 4 weeks of using the BioBase programme. There were nonsignificant effects of paid status on STAI (t54=1.97; P=.054) and WEMWBS (t53=0.28; P=.78; Figure 1).

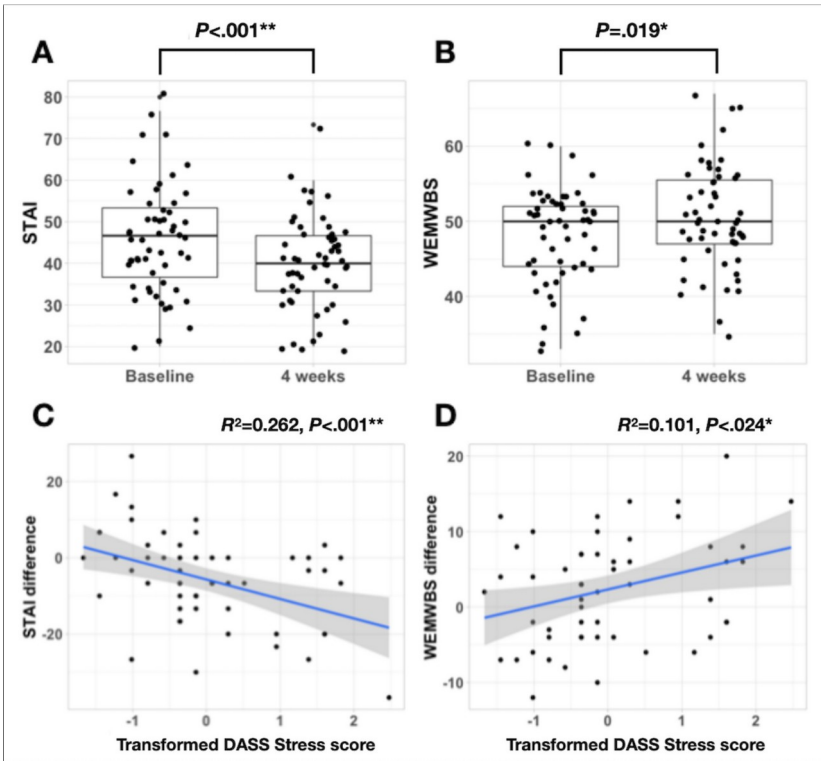


*Figure 1. After controlling for paid status, differences in (A) 6-item STAI and (B) WEMWBS after 4 weeks of the BioBase program. Association between transformed DASS Stress scores and difference in (C) STAI and (D) WEMWBS, **P<.01, *P<.05. DASS: Depression Anxiety Stress Scale; STAI: State-Trait Anxiety Inventory; WEMWBS: Warwick-Edinburgh Mental Well-being Scale.*

After controlling for paid status, linear regression models revealed that a higher baseline DASS stress significantly predicted a greater decrease in STAI (t53=–3.93; P<.001; CI –8.62 to –2.80; R2=0.262) and significantly predicted a greater increase in WEMWBS (t52=2.32; P=.024; CI 0.30-4.12; R2=0.101; Figure 1).

Participant Feedback

As much as 32 of 55 participants completed a feedback questionnaire with both multiple-choice questions and free-text answers and were given the opportunity to expand their responses in an interview with an experienced researcher (NH). Responses from 2 participants were withdrawn because the participant did not want their responses shared. Table 2 describes responses from 30 participants (18 paid and 12 unpaid) of the multiple-choice questions.

Table 2. Feedback questionnaire results on intervention engagement, usability, and self-reported behavior change.

Theme	Paid (N=18, 60%)	Unpaid (N=12, 40%)	*P* value
Feasibility and usability			<.001[a]
How many times did you usually log in?			
Never, n (%)	0 (0)	0 (0)	
Once a week, n (%)	0 (0)	1 (83)	
Several times a week, n (%)	0 (0)	7 (58)	
Once a day, n (%)	10 (56)	2 (17)	
Several times a day, n (%)	8 (44)	2 (17)	
What time of day did you usually log in?			.50[a]
Morning before work, n (%)	3 (17)	2 (17)	
Morning at work, n (%)	2 (11)	0 (0)	
Lunch time, n (%)	0 (0)	2 (17)	
Afternoon at work, n (%)	1 (6)	1 (8)	
Evening at home, n (%)	12 (67)	7 (58)	
Where did you usually log in?			>.99[a]
Home, n (%)	14 (78)	10 (83)	
Work, n (%)	4 (22)	2 (17)	
Commute, n (%)	0 (0)	0 (0)	
Acceptability			
Would you recommend the BioBase programme to a friend?			.13[a]
Yes, n (%)	15 (83)	6 (50)	
Maybe, n (%)	3 (17)	5 (42)	
No, n (%)	0 (0)	1 (8)	
Would you like to continue using the BioBase programme?			.26[a]
Yes, n (%)	11 (61)	4 (33)	
No, n (%)	7 (39)	8 (67)	
Promoting behavior change			
Have there been any changes to your physical health since starting the BioBase programme?			>.99[a]
Yes, n (%)	6 (33)	3 (25)	
Maybe, n (%)	2 (11)	2 (17)	
No, n (%)	10 (56)	7 (58)	
Have there been any changes to your mental health since starting the BioBase programme?			.06[a]
Yes, n (%)	5 (28)	3 (25)	
Maybe, n (%)	3 (17)	7 (58)	
No, n (%)	9 (50)	2 (17)	
Have there been any changes to your engagement at work since starting the BioBase programme?			.69[a]
Yes, n (%)	6 (33)	3 (25)	
Maybe, n (%)	2 (11)	3 (25)	
No, n (%)	10 (56)	6 (50)	

[a]Fisher exact test.

Participant Feedback: Feasibility and Usability

Participants stated that the app was easy to use and navigate (“really good experience” [unpaid participant]). The wearable was also considered usable (“easy to wear and keeps you

informed" [paid participant]), with participants paying particular attention to the battery life ("the battery life of the device is exceptional" [paid participant]).

The BioBase programme was convenient and feasible to use as 73% (22/30) of participants reported checking the app at least once per day ("I think we all checked it when we first got in 'cause we were talking about our sleep and our rubbish sleep, and I guess at lunchtime I'd have a go and then when I'd go home" [unpaid participant]). There was a significant difference between paid and unpaid participants in how often they logged in, with unpaid participants logging in less often (P<.001; Table 2).

There were no differences between paid and unpaid participants in the time of day or location in which they logged in; 80% (24/30) of participants logged in at home and 70% (21/30) in the evening or in their lunch break in the workplace ("Yeah normally at home, I do it on my lunch break as well, when I was at work at times, when was stuck in the queue or something" [paid participant]).

Participant Feedback: Acceptability in the Workplace

The majority of participants responded they would recommend the programme to a friend (70%, 21/30) and 50% (15/30) would have liked to continue with the programme after 4 weeks. There were no differences between paid and unpaid participants in this respect (Table 2).

BioBase was seen as an acceptable solution in the workplace to support employee health ("I think it's definitely needed. I think it's a really great way to track your company satisfaction and well-being and this changes the way people build their HR policies, 100%. I think there are a lot of companies that say they are doing really amazing things, but might not be, and this is a good way to prove it, because you're not like messing around with the data" [unpaid participant]). Participants appreciated the accessible and personal approach ("on demand,

private-ish, lots of great education" [unpaid participant]) and the concept of increasing self-awareness ("it provides great exposure to education and techniques for managing and discussing mental health maintenance," "The programme has made me more self-aware...I think the BioBase is excellent for workplace use" [unpaid participant]). Many participants had support and encouragement from their colleagues ("I just used the app mostly at work because I was sort of reminded by everyone here" [unpaid participant], "we would do the breathing things, if one of us was stressed" [unpaid participant]).

Participant Feedback: Promoting Behavior Change

There were no differences between paid and unpaid participants with regard to perceived changes in physical health, mental health, or engagement at work (Table 2). Key themes in the feedback revealed that real-time data helped participants to understand their current state ("BioBase has helped me become more aware of my body and how external factors were affecting me" [paid participant]) and the impact of environmental triggers ("being aware of what triggers stress and being taught and successful ways of managing this. Has been a game changer" [paid participant]).

Learned proactive and preventative mechanisms from the modules and tools helped achieve positive outcomes ("I approach some tasks, particularly procrastination differently sometimes. I am also even more mindful of taking time to unwind." [unpaid participant]; "[I learned] how to breathe properly, coping mechanisms for stress, how to get things into proportion where stress concerned and how to cope with stress by developing interests outside work and relaxing after work" [paid participant]). Specifically, the breathing tool was found to be very useful to achieve on the spot stress reduction ("I like the breathing exercises; this is something I will continue with after the trial. I find them very useful for winding down in the evenings before sleep" [paid participant]).

Changes in sleep habits was seen as a consistent behavior change with multiple participants reporting various new health behaviors (“[I] prepare myself for sleep more deliberately than previously” [unpaid participant], “I'm more conscious of going to bed at a reasonable time” [unpaid participant], “I make the room completely dark 30 mins before going to sleep” [paid participant], “Ensuring I go to bed at a particular time even on the weekend” [paid participant]).

There were also reports of increased social connection (“the tribes concept stuck with me and I intend to invest more in building social groups” [unpaid participant], “I consciously try and arrange social activity to increase my activity [levels, like] playing golf, badminton, swimming” [paid participant]).

Discussion

Framing note on Discussion (thesis-level). The Discussion below presents the published paper's interpretation, organised around symptom-change findings, comparison with prior work, and limitations. The thesis framing of this chapter (s. 3.1) leads with a different priority order, summarised here before the published Discussion begins. The chapter's primary contributions are: (i) preliminary evidence that the BioBase platform can be delivered to working adults under elevated stress with adequate retention (feasibility); (ii) evidence that respondents who completed the feedback questionnaire found the intervention acceptable, with important caveats around the paid-versus-unpaid engagement asymmetry that limits how this finding generalises beyond paid workplace cohorts (acceptability, qualified — see the new "Engagement and the limits of the acceptability findings" subsection below); (iii) a preliminary indicative pre-post symptom-change signal on STAI and WEMWBS that, given the single-arm design and the multi-component non-specific effects discussed in s. 8.3.2 and Whitton et al. (2015), is consistent with — but does not establish — a real intervention effect, and that justifies subsequent controlled evaluation (preliminary indicative signal, not

efficacy). The published Discussion below should be read with these framings in mind; the new "Engagement and the limits of the acceptability findings" subsection (from C3-2), the new "Reflections on industry-academic collaboration" subsection (from C3-3), and the Chapter conclusion (s. 3.8) elaborate these qualifications.

Principal Results

This study explores the efficacy of BioBase, which to our knowledge is the first self-guided intervention that combines measurement and biofeedback of passively collected physiological data and linked content on physical and mental health. We found significant decreases in anxiety and increases in mental well-being after 4 weeks of using BioBase in a workplace setting. Importantly, this study found that higher baseline stress levels were associated with greater decreases in anxiety and increases in well-being. These results fit into a wider body of literature showing that digital health interventions may improve symptoms of depression, stress, and anxiety in a workplace environment (Bostock et al., 2019; Tan et al., 2014), with more pronounced benefit in those with elevated stress levels (Heber et al., 2013; Weisel et al., 2018).

Engagement and the limits of the acceptability findings

The acceptability outcomes reported here should be read with two engagement-related caveats. First, the published feedback dataset (Kawadler et al., 2020, Table 2) represents 30 of the 55 enrolled participants who completed the post-intervention feedback questionnaire; the views of the 25 enrolled participants who did not complete this questionnaire are not represented. Second, the published engagement data show a clear paid-versus-unpaid asymmetry in log-in frequency ($p < .001$; Kawadler et al., 2020, Table 2): all 18 paid participants logged in at least once a day (10 once daily, 8 several times daily), whereas only 4 of 12 unpaid participants logged in daily or more frequently, with the remaining 8 logging in once a week or several times a week. The headline acceptability statistic — that 70% (21 of 30) of respondents would recommend BioBase to a friend — should therefore be understood as a 70% endorsement among the subset of participants who engaged sufficiently to complete the feedback questionnaire, predominantly drawn from the paid arm. The 25 non-respondents and the differential engagement patterns by payment status mean that the acceptability findings are likely to over-represent positive views of the programme.

A formal descriptive comparison of engagers and non-engagers on baseline characteristics — DASS-21 stress, age, gender, baseline anxiety and wellbeing — was requested at viva and was originally planned for these revisions. It could not be performed because the underlying

trial dataset is held by Huma Therapeutics Ltd (the acquirer of BioBeats Ltd) and is no longer accessible following decommissioning of the company's research infrastructure; this data-availability constraint is discussed as a thesis-wide transparency limitation in s. 8.3.4. The broader smartphone-mental-health attrition literature provides a relevant external benchmark: average study attrition in 70 RCTs of smartphone mental-health interventions is 24.1% (95% CI [19.3, 29.6]) at short-term follow-up and 35.5% (95% CI [26.7, 45.3]) at longer-term follow-up (Linardon & Fuller-Tyszkiewicz, 2020). The 25 of 55 (45%) non-respondents to the feedback questionnaire in the present study is consistent with the upper end of that distribution and underscores that the present feasibility evidence is partial.

Comparison With Prior Work

Self-guided digital health interventions have been shown to have positive effects on mental health and well-being (Weber et al., 2019) and can be as effective as face-to-face treatments for anxiety (Cuijpers et al., 2010; Ebert et al., 2016). The results from this study show BioBase is also potentially effective. The effect sizes found were d=0.96 for reduction in anxiety and d=0.66 for increase in well-being, which is higher than previous studies using self-guided interventions (d=0.44 for work-related worry (Ebert et al., 2016); d=0.47 for anxiety (Proudfoot et al., 2013)).

The BioBase programme was found to be feasible for use in the work environment, due to ease of access on a personal or work phone (Howarth et al., 2018). Participants found it relevant, acceptable, and that it worked within their time constraints, as it focused on work-related stressors with simple, short tasks that are easily implemented in the working day (Ebert et al., 2016).

From an employer’s perspective, self-guided interventions do not require the vast resources needed for either traditional organizational support or therapist- or coach-assisted interventions. In addition, they provide more accessibility than in-house training and reach a larger population with comparable clinical outcomes (Andersson, 2009). Our findings suggest that participants with higher levels of stress had significantly greater reductions in anxiety and increases in well-being, which may help employers deliver a solution for those employees most at risk.

Self-guided interventions have the advantages of allowing the user to go at their own pace and access content that is right for them; however, they rely more on an individual's willingness to take part and often have lower engagement rates or failed adoption (Muuraiskangas et al., 2016). Participants in this study used the app on average 72% of days (20.9/29 days), which is higher than (Ebert et al., 2016; Weber et al., 2019) or similar to (Proudfoot et al., 2013) other self-guided digital interventions. It is possible that engagement was higher because the BioBase programme is a unique intervention that provides targeted and relevant content along with integrated technology (Howarth et al., 2018). The feedback suggests that people appreciated different aspects of the programme based on what content and insights were tailored (ie, feedback on sleep quantity and quality and ability to customize recommended modules). These findings replicate and extend those from Chapter 2, demonstrating that the pattern of greater improvement among participants with higher baseline stress is consistent across different populations and contexts.

Limitations

This study was conducted with participants recruited from a workplace setting across a variety of workplaces and job roles; however, it is beyond the scope of this study to separate participants by type of role or organizational conditions. Future studies with larger samples should be carried out to explore these effects.The current feasibility study did not utilize a control group and therefore was not able to rule out a regression to the mean or a placebo effect of using a digital intervention. Future studies should conduct a randomized controlled trial to add evidence that the anxiolytic effects were more specific than a placebo effect.

Due to the personalized nature of the intervention and the multiple therapeutic touch points in the BioBase programme, it is difficult to conclude which elements of the BioBase programme contributed most to the observed reductions in anxiety and increases in well-being. Further practical investigation is needed into breaking into the "black box" of digital health

interventions to understand the main therapeutic effects (McKay et al., 2018; Whitton et al., 2015).

Future directions: actively recruiting non-engagers

A methodological priority for future feasibility work on workplace digital mental-health interventions is to actively recruit a subsample of non-engagers for follow-up — for example through brief structured exit interviews with participants who downloaded the app but did not progress, or with those who declined to enrol — to characterise the barriers to engagement they experienced. Qualitative work of this kind has been conducted in adjacent workplace digital mental-health programmes (Carolan & de Visser, 2018, identified time pressures, perceived irrelevance to immediate stressors, and privacy concerns as recurrent barriers among workplace digital mental-health non-engagers) and should be considered standard practice for feasibility-phase research within the MRC framework for complex-intervention development (Skivington et al., 2021). Without such non-engager data, acceptability findings derived only from respondents are systematically biased toward favourable views.

3.7.3 Reflections on industry-academic collaboration

The study reported in this chapter was conducted within a specific industry-academic collaboration that warrants explicit reflection. During the Chapter 3 study I held the position of Chief Executive Officer of BioBeats Ltd, the company that developed and provided the BioBase platform; my co-authors Jamie Kawadler, Nicola Hemmings, and Sonia Ponzo were BioBeats employees, and Geoffrey Bird was a freelance consultant compensated by BioBeats for his work on the study. BioBeats Ltd provided the participant recruitment fees and the wearable devices used in the study. These affiliations were declared in the published Conflicts of Interest statement of Kawadler et al. (2020). A subset of the participants — the eighteen paid participants who showed substantially higher app log-in frequency than the twelve unpaid participants ($p < .001$; see the Engagement subsection above) — were recruited through a workplace wellness partnership between BioBeats and AXA. AXA's role in the study was therefore structural: as a workplace partner providing recruitment access, AXA was also a commercial customer of BioBeats during the period in which the research was being conducted.

Value the partnership brought. The AXA partnership provided three things that an unfunded academic project of comparable scope could not have provided: recruitment access to a real working-adult cohort within their employment context, the operational infrastructure to deliver four weeks of passive biometric data collection paired with content modules, and the financial support for participant incentives and device provision. The ecological validity of the workplace context — participants using BioBase during real working days with real workplace stressors — was a direct consequence of the partnership and is a meaningful strength of the study.

Tensions and constraints. Three tensions warrant honest acknowledgement. First, the dual role I held — CEO of the commercial entity and DPhil researcher — created an inherent structural conflict between commercial interests in demonstrating BioBase's value and scientific interests in evaluating it candidly. The published Conflicts of Interest statement made this transparent but did not by itself resolve it. Second, the paid-versus-unpaid recruitment design was shaped in part by commercial logic (AXA evaluating BioBase for

their workforce) rather than by a clean methodological design alone, and as a result the study's engagement and acceptability data are confounded with payment status in ways now explicitly documented in the Engagement subsection above. Third, post-acquisition data accessibility — discussed at the thesis level in s. 8.3.4 — was not contractually anticipated at study inception, with consequences for what could be done in these revisions in response to the panel's request for an engagers-versus-non-engagers descriptive comparison.

Learning for future industry-academic collaborations. Several principles follow from this experience. Pre-specified analytic plans (preferably pre-registered) should be lodged at study inception rather than developed during analysis, providing a structural defence against commercially convenient framing. Where commercial and scientific roles co-occur in the same individual, an independent statistical analyst — outside the commercial entity — substantially improves the credibility of the resulting analysis. Data-sharing arrangements should be drafted explicitly with company acquisition, dissolution, and infrastructure-decommissioning scenarios in mind; the present limitation under s. 8.3.4 would have been avoided by such an arrangement. Pre-agreed reporting standards (CONSORT 2025 for trials, and equivalent reporting guidelines for other study types) constrain post-hoc commercial framing of results. Conflict-of-interest disclosure should be detailed at the level of specific role, financial arrangement, and equity holding, rather than at the general level of company affiliation. A general lesson: research conducted from inside a commercial entity is a legitimate and often necessary route into real-world evaluation, but it requires methodological safeguards that exceed those routinely required for independent academic work.

Chapter conclusion

Read against the feasibility and acceptability research questions stated in s. 3.1, the present study contributes the following to this thesis. *Feasibility:* The BioBase platform can be delivered to working adults under elevated stress, with retention rates supporting interpretation of subsequent outcomes and with no reported platform-level failures during the 4-week intervention period. *Acceptability:* Participants who engaged with the feedback questionnaire found the intervention acceptable (70% would-recommend rate among 30 of 55 respondents), but this finding is qualified by a paid-versus-unpaid engagement asymmetry (see the new "Engagement and the limits of the acceptability findings" subsection above; C3-2 revisions) that suggests acceptability evidence drawn only from respondents is biased toward favourable views; future feasibility work should actively recruit non-engagers to characterise barriers (see the new "Future directions: actively recruiting non-engagers" subsection above; Carolan & de Visser, 2018). *Preliminary indicative signal:* Pre-post reductions in self-reported anxiety (STAI) and increases in wellbeing (WEMWBS) of moderate-to-large magnitude were observed, but the single-arm design, the multi-component nature of the intervention (s. 8.3.2; Whitton et al., 2015), and the post-acquisition data-availability constraints (see the new "Reflections on industry-academic collaboration" subsection above and s. 8.3.4) mean these descriptive findings are best read as justification for subsequent controlled evaluation rather than as efficacy evidence. The chapter as a whole positions the BioBase workplace intervention as a candidate component of broader occupational mental-health support pathways, warranting controlled comparison against active alternatives in subsequent MRC-evaluation-phase work, rather than as a standalone efficacious treatment ready for deployment (see s. 8.1.3.1 for the General Discussion summary).

4. Chapter 4: Digital ACT for Generalized Anxiety Disorder: A Person-Based Design Study

Content adapted from: Hemmings NR, Kawadler JM, Whatmough R, Ponzo S, Rossi A, Morelli D, Bird G, Plans D (2021). Development and Feasibility of a Digital Acceptance and Commitment Therapy–Based Intervention for Generalized Anxiety Disorder: Pilot Acceptability Study. JMIR Formative Research. https://doi.org/10.2196/21737

A note on contributions. This study was published with Nicola Hemmings as first author (Hemmings et al., 2021); Nicola joined the team to lead the Person-Based Approach methodology that frames the study. I contributed to adapting the BioBase design for a clinical-anxiety population, programmed the ACT-based therapeutic modules and integrated them into the platform, co-developed the BioBeam firmware for continuous physiological monitoring, analysed the user-engagement metrics and clinical outcome data, and led much of the manuscript writing.

Abstract

Generalized anxiety disorder (GAD) is characterized by excessive worry and high comorbidity with depression, yet individuals face long wait times for diagnosis and accessibility barriers to treatment. This study applied the person-based approach to develop and pilot test a digital acceptance and commitment therapy (ACT) intervention for GAD delivered through the BioBase platform described in previous chapters. Phase 1 involved creating guiding principles, intervention design objectives, and key features based on qualitative research with the target population, resulting in six ACT modules targeting psychological flexibility through awareness, openness, and values-based action. In Phase 2, ten adults with diagnosed GAD (mean age 31.2 years, SD = 8.98) received the app-based intervention and paired wearable device for two weeks. Results showed significant improvements in the primary outcome of psychological flexibility measured by the Acceptance and Action Questionnaire-II (baseline median = 30.5, exit median = 29.0; W = 50, p = .11, effect size = 0.512), though this did not reach statistical significance due to the small sample. Secondary outcomes demonstrated significant reductions in GAD-7 anxiety scores (baseline median = 14.0, exit median = 9.0; W = 34.0, p = .03, effect size = 0.691) and

PHQ-9 depression scores (baseline median = 11.5, exit median = 9.5; W = 45, p = .008, effect size = 0.840). The intervention achieved excellent usability ratings (System Usability Scale mean = 83.4, IQR 75-96.9) with 77% of participants completing all modules. Qualitative feedback highlighted the relevance of ACT concepts, particularly the body scan and mood tracking features for developing present-moment awareness. This pilot study demonstrates that ACT-based digital interventions developed using person-based approaches are acceptable and feasible for individuals with GAD, warranting larger-scale evaluation.

Introduction

Background

Generalized anxiety disorder (GAD) is diagnosed if an individual has excessive worry that is difficult to control for more days than not over a period of 6 months . GAD is associated with an increased reactivity to, and avoidance of, internal experiences (Roemer et al., 2005). There is high comorbidity of GAD with other anxiety (51.7%) and mood (63%) disorders (Ruscio et al., 2017), especially with depression (r = .62) (Kessler et al., 2005).

Just under half of individuals with GAD suffer for 2 years before correctly being diagnosed (Baldwin et al., 2012), and even after diagnosis, waiting times for treatment are up to 18 weeks (England, 2020).

A digital therapeutic tool has the potential to increase accessibility and availability of treatment for those suffering with GAD by reducing barriers such as waiting times before treatment (Bandelow et al., 2017), perceived stigma (Goetter et al., 2020; Marques et al., 2010), geographical distance, financial costs, and lack of time due to, for example, work commitments or caring responsibilities (Andrade et al., 2014). Digital interventions, whether therapist-guided or self-guided, have been shown to be as effective as face-to-face interventions (Carlbring et al., 2018; Lattie et al., 2019) when based on the same core

processes of face-to-face treatment with only the mode of delivery changing (Hedman et al., 2012).

Acceptance and Commitment Therapy as a Treatment for GAD

The choice of ACT for this study was reached through clinical and pragmatic consideration rather than a systematic head-to-head comparison of candidate treatments. The decision space included the principal evidence-based options for generalised anxiety disorder (GAD): cognitive behavioural therapy (CBT), and in particular disorder-specific protocols for GAD targeting intolerance of uncertainty and worry; ACT; mindfulness-based approaches such as mindfulness-based stress reduction and mindfulness-based cognitive therapy; behavioural activation; and metacognitive therapy. CBT is the first-line treatment for GAD in NICE guidance (NICE, 2011) and has the largest evidence base, but the theoretical landscape for GAD remains contested, with multiple competing process models (Behar et al., 2009).

Several considerations pointed towards ACT. First, ACT places explicit emphasis on present-moment bodily awareness and on the willingness to experience internal sensations rather than avoid them — both of which align directly with the thesis' interoception focus, in a way that disorder-specific CBT protocols (which target worry content and cognitive appraisal) do not. Second, ACT has a transdiagnostic effect profile (Swain et al., 2013; Newby et al., 2015), attractive given the high comorbidity of GAD with depression already noted in this chapter's introduction. Third, ACT's process-based structure — short metaphor-driven exercises delivered across the Aware, Open, and Action pillars — adapts well to brief, self-guided digital modules, and there is existing evidence for the acceptability and efficacy of digital ACT in anxiety (Kelson et al., 2019).

These principled reasons sat alongside two pragmatic ones that should be stated openly: the supervisory team and the BioBase development team had prior experience and expertise in ACT-based intervention design, and the BioBase platform infrastructure could be extended to

deliver ACT modules with minimal redevelopment. The selection was not, therefore, the outcome of a systematic comparative trial of candidate models, and this is acknowledged as a limitation of the design rationale. This rationale was developed across supervisory discussions and an internal literature review in 2020.

Acceptance and commitment therapy (ACT) is a form of cognitive behavioral therapy, rooted in functional contextualism (Hayes, 2016) and relational frame theory (Tonneau, 2004).

ACT aims to increase psychological flexibility, or the ability to deal with challenging experiences in a flexible way while continuing to act based on one's values (Hayes et al., 2006). Psychological flexibility reduces experiential avoidance and the unwillingness to experience difficult emotions, thoughts, or sensations (Hayes et al., 2006).

The aforesaid reduction of experiential avoidance and increase in psychological flexibility are achieved via 6 core processes of change that are closely interlinked. Contact with the present moment is the nonjudgmental present moment awareness of one's internal and external environment. Self-as-context is noticing that one is not just ones' thoughts, emotions, and self-image, but they are also the observer of them. Cognitive defusion includes techniques aiming to treat thoughts and feelings for what they are (mental imagery, streams of words, sensations), not as truths that must be reacted to or get caught up in. Acceptance means allowing unpleasant feelings, rather than trying to change them. Values are knowing what matters in life to provide meaning and direction. Committed action is taking targeting value-based action and doing what it takes, even when difficult (Hayes et al., 2006; Kashdan & Rottenberg, 2010).

Psychological flexibility is proposed to be a fundamental aspect of health and well-being; even though ACT does not aim to specifically eliminate symptoms of anxiety, the ACT model has a transdiagnostic approach and has been consistently shown to reduce symptoms of

anxiety (Swain et al., 2013) and depression (Newby et al., 2015) among other mood disorders via its 6 core processes. This effect could be due to the reduction of “experiential avoidance,” a proposed concept of anxiety (Roemer et al., 2005). Experiential avoidance is the attempt to avoid or control internal experiences, rather than accepting them. This fear of losing control over one’s emotional responses (in particular anxiety) is seen as the opposite of being psychologically flexible; therefore, increasing psychological flexibility can reduce experiential avoidance, and thus increase psychological flexibility.

ACT has been used in digital interventions and has shown to be an efficacious and acceptable treatment for adults with anxiety disorders (J. Kelson et al., 2019), and there is early support for therapist-guided and self-guided digital ACT-based effectiveness reducing anxiety in populations with anxiety disorders (M. Brown et al., 2016; Hoffmann et al., 2018; Ivanova et al., 2016) and the general population (Ahtinen et al., 2013; J. N. Kelson et al., 2017; Levin et al., 2014).

However, poor engagement rates are a well-documented and an ongoing concern in the design, development, and evaluation of digital interventions (Batterham et al., 2021; Eysenbach, 2005; Linardon & Fuller-Tyszkiewicz, 2020; O’Connor et al., 2016). Self-guided interventions suffer from even greater dropout than therapist-guided interventions (Domhardt et al., 2019), potentially due to the increased support that these can offer (Kelders et al., 2012). However, reasons for poor engagement and dropout are not consistently reported in the literature; a systematic review from 2010 reported a wide range of adherence rates (2%-83% (Melville et al., 2010)) in 16 studies using internet-based interventions and specifically called for analysis of variables associated with dropout. Consistent reporting of these variables remains an issue in the literature (Beatty & Binnion, 2016; Fleming et al., 2020; Sieverink et al., 2017). To increase engagement, which is linked to efficacy (Donkin et al., 2011), more research is needed to determine whether adopting an accepted framework, such as the person-

based approach (PBA), could serve as a potential first step in designing interventions with greater engagement rates.

Person-Based Approach

The PBA is a systematic framework for designing interventions that addresses and accommodates perspectives of the people who will use them. Adopting a PBA addresses the ongoing concern of a lack of engagement in digital interventions (Yardley et al., 2015, 2016), with the aim of increasing the likelihood of achieving the desired therapeutic outcome. It goes beyond the traditional system of collecting user feedback as it addresses a person's experience of the intended behavior change techniques (Yardley 2015). The approach is conducted over 2 stages to create a persuasive, feasible, and relevant digital intervention . First, the development stage creates a persona, a summary representation of qualitative research conducted with a wide range of people from the target user's population, showcasing a deep understanding of their psychosocial context alongside their views on the proposed intervention. The second stage identifies guiding principles that inform the intervention development in addressing the persona's key context-specific behavioral issues. Understanding the user's views and psychosocial context enables the ability to address barriers to engagement and therefore, increase acceptability, feasibility, and efficacy of the digital intervention (Bradbury et al., 2019).

The PBA is a useful methodology in designing digital interventions; however, to our knowledge, there has been no published research on using the PBA for the development of a suitable self-guided digital intervention for GAD.

Aim

This chapter describes the development of a digital intervention via a mobile app for GAD, based on ACT using the PBA. Furthermore, it reports the results of a pilot study evaluating the prototype digital intervention in a sample of adults with self-referred GAD. The pilot

study is designed to (1) evaluate the acceptability and usability of ACT-based content in a digital format derived using the PBA, and (2) provide a possible trend of the efficacy of this intervention with respect to psychological flexibility, mental well-being, and symptoms of anxiety and depression.

Methods

Pre-registration

Neither the Phase 1 development work (Person-Based Approach) nor the Phase 2 feasibility test reported in this chapter was pre-registered. At the time of conduct (early 2020), pre-registration of mHealth intervention development and feasibility studies was not yet routine practice in the field. The intervention design, target sample, outcome measures and analytic approach reported in this chapter correspond to those documented in the published version of the work (Hemmings et al., 2021). The absence of pre-registration is acknowledged as a transparency limitation, with cross-thesis implications discussed in s. 8.3.

Phase 1: Development

Technical Platform

Building on the biometric feedback interventions described in Chapters 2 and 3, this study addresses a specific clinical population—individuals with generalized anxiety disorder (GAD). As discussed in Section 1.2.3.2, anxiety disorders show consistent associations with altered interoceptive processing. GAD specifically involves excessive worry and experiential avoidance (Roemer et al., 2005), characteristics that may benefit from interventions targeting bodily awareness and psychological flexibility. The BioBase platform's capacity for delivering therapeutic content alongside biometric monitoring provides an ideal foundation for implementing acceptance and commitment therapy (ACT) principles, as digital ACT interventions have shown promise for GAD when they maintain fidelity to core therapeutic

processes while adapting delivery for remote access (Dahlin et al., 2016; Dear et al., 2015; Ivanova et al., 2016).

The ACT-based modules that provided psychoeducational information and activities on the core processes were shown via an in-app course named “Find Your Way.” The course was complemented by in-app BioBase content to support the ACT pillars of “Aware,” “Open,” and “Active”. Mood tracking was made available through an Ecological Momentary Assessment (EMA; (Shiffman et al., 2008)), which allows individuals to increase their awareness of their emotional state by choosing a mood from a list of options and specify any ecological component surrounding the moment they chose to declare their mood (ie, where they were, whether they were alone or with somebody, and what activities they were engaged in). EMAs have been shown to be a valuable mood-tracking tool in the context of digital therapeutics aiming to reduce levels of anxiety and depression (Schueller et al., 2017).

In addition to the mood tracking tool, an in-the-moment breathing exercise for stress reduction and a mindfulness-based progressive relaxation tool (Body Scan) were used to reduce symptoms of anxiety and depression through either heart rate variability biofeedback (Blanck et al., 2018; Schoenberg & David, 2014) or awareness of body sensations intended to modulate interoceptive awareness (Bornemann et al., 2014).

Finally, passive data collection on physical activity (ie, the number of steps performed every 20 seconds) and sleep continuity (eg, hours slept, number of awakenings) was achieved via the paired wearable device, BioBeam, and data on sleep and physical activity history were displayed on in-app dashboards. Awareness of and insight into one’s own sleep and activity patterns have been shown to affect anxiety and well-being (Chou et al., 2015; Stubbs et al., 2017).

Intervention

Stage 1: Development of a GAD Persona

A persona defines the guiding principles of the intervention: a collection of symptoms, desired health behaviors, barriers to treatment, and desired impact of the intervention. Such a persona was created to understand the current psychosocial context of the target population (individuals with GAD).

The Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) (Chou et al., 2015) was consulted to clarify the current diagnostic criteria. Clinically recommended therapies, and their critiques, were investigated alongside models of GAD, self-help strategies, and confounding factors to their effectiveness (Behar et al., 2009; J. Kelson et al., 2019; Swain et al., 2013; Wolitzky-Taylor et al., 2012). This review of the literature highlighted behavioral issues specific to GAD, centered around explicit or voluntary cognitive avoidance strategies, and avoidance of internal experiences.

To create relevant scenarios in which to base an intervention, a persona narrative was created by a psychologist (NH) which was centered around everyday life for a person with GAD. These narratives were based on the psychologist's experience of user feedback from previous BioBase studies with individuals with high anxiety outlined in Chapters 2 and 3 in conjunction with the aforementioned review of the literature. Examples include:

> *I enjoy my job but struggle to concentrate and feel guilty for taking time off when my anxiety is overwhelming*
>
> *I just feel like I am always on edge and it is affecting my work and relationships*
>
> *I don't know what to do, I am just always anxious and can't stop it*
>
> *I am always tired and can never concentrate*
>
> *I feel like there is this looming dread - but I'm not sure where it is coming from*

My whole body is exhausted, I feel constantly tense, all my muscles ache

I can't remember when I last slept well

In addition to the narratives, clear user journeys were created highlighting how a person would receive and interact with the digital intervention. A list of desired outcome behaviors was also collated to inform the intervention design.

This user-centered methodology represents an evolution from the intervention approaches in Chapters 2 and 3, incorporating target population input from the design phase rather than relying solely on post-intervention feedback.

Stage 2: Intervention Design and Creation

A prototype digital intervention was constructed that targeted the desired outcome behaviors and integrated into the existing National Institute for Health and Care Excellence's (NICE) stepped treatment pathways for GAD (NICE, 2011). The NICE stepped treatment pathway organizes the provision of service based on the severity of an individual's symptoms. Initial steps involve monitoring and provision of relevant information. This escalates to self-help or therapist-guided interventions and then to medication and combined interventions.

The intervention design objectives were to reduce symptoms of GAD (ie, excessive worry and anxiety, linked to increased reactivity to and avoidance of internal experiences) by means of increasing psychological flexibility.

The key features that could address these aims are provision of education and guidance on anxiety, specific therapeutics to help the users to be open to and aware of their internal experiences, guidance on defusing and accepting difficult internal sensations, and advice as to how to clarify personal values and reduce barriers to life goals.

The content was curated to target key behaviors, efficacy of ACT processes (Levin et al., 2012), exercises in a digital format (J. N. Kelson et al., 2017; Mattila et al., 2016), and use of visual metaphors for engagement with the program.

The treatment aimed to increase the concept of psychological flexibility through psychoeducation and exercises based on the 3 pillars of ACT: Aware, Open, and Action. Across these 3 pillars, the treatment program is structured around the 6 core processes of ACT: present moment, self-as-context, cognitive defusion, acceptance, values, and committed action (Table 1). This feasibility study investigated a prototype of an initial 6 modules (of 30 total modules designed)—1 module for each core ACT process. Each module contained the same structure: a brief overview of a metaphor that describes a core ACT process, a description of the metaphor (that can be accessed via text or audio file), and an activity to complete with a free text box in which the users can apply the therapeutic concept to their current scenario.

Table 1. The structure of the digital intervention prototype. The outline evidences the 3 core pillars of ACTa alongside the 6 ACT processes and links the intervention design objectives with the intervention key features (module name and exercise).

Module number	Pillar	Process	Intervention design objective	Key feature (Exercise)	Module name
1	Aware	Present moment	The users have increased awareness of what is driving their decisions in life (thoughts, feelings, emotions, and urges).	Passengers on the bus. Notice what thoughts, emotions, feelings, and sensations you are carrying with you.	In your driving seat
2	Aware	Self-as-context	The users have a greater awareness of their internal narrative directed by their thoughts and its effect on their behavior and decisions throughout the day. They check that they are basing decisions and behavior on reality, not the story their thoughts are telling them.	Noticing if the story your thoughts are telling you is different from reality.	The storyteller
3	Open	Cognitive defusion	The users are better able to notice and distinguish between thoughts that are either helping them or preventing them from reaching their goals and then build a positive mindset by attending to helpful thoughts.	Labeling thoughts as helpful or unhelpful.	Is this helpful?
4	Open	Acceptance	The users are aware of difficult emotions, thoughts, or urges they are experiencing, and the short-term gains and long-term costs of their avoidance behaviors.	Journaling the short- and long-term impact of current avoidance behaviors.	What are you avoiding?
5	Active	Values	The users connect with their values and what really matters to them to live a meaningful, valued life.	Journaling what a loved one says about you during an anniversary speech.	Your attention, please
6	Active	Committed Action	The users state that they are willing to experience initial discomfort in order to achieve meaningful goals for them.	Acknowledging barriers and creating an action plan.	ACTion plan

The 6 modules, each taking less than 5 minutes to read/listen, could be accessed and completed in any order, as previous research has shown no consistent user preference for guided or unstructured user journey (Viskovich & Pakenham, 2018). The user was free to spend as long as necessary on the activities as in other digital interventions (Ahtinen et al., 2013). A schedule of 1 module per day was recommended to participants. Each module contained a follow-up notification on completion, which reminded the users to be cognizant of the concept that they had worked through.

To safely address readability and clarity of the modules, iterative testing was conducted with 10 paid user testers with no pre-existing anxiety. Modifications were made to simplify the language and to improve the clarity of what is addressed by the activity. The app was then

further reviewed by a clinical psychologist (RW), who has experience using ACT-based approaches in GAD, whose feedback was fed into the program design via the PBA framework.

To meet the specific behavioral context of GAD (such as excessive worry) the overall design of the app was intended to be relaxing and calming. The home screen changes color based on the time of day to anchor the persons to their present environment. The tone and style of the language used were accessible, nonjudgmental, friendly, and supportive and avoided explicitly mentioning any specific medical conditions or diagnosis (unless necessary within therapeutic Step 1: “Information on anxiety”). All modules had a Flesch Kincaid reading ease score (Thomas et al., 1975) of 71.5 or above suggesting it should be easily understood by individuals aged 13 and above.

Phase 2: Feasibility Testing

Overview

A feasibility test of the initial prototype app–based therapeutic (testing acceptability, usability, and efficacy) was conducted in a sample of individuals who have reported a diagnosis of GAD from a mental health professional as the first step in an iterative cycle of development and evaluation. The study took place in February-March 2020.

Participants

Participants were recruited from a pool of subjects excluded from the study in Chapter 2 because of a declaration of diagnosed mental health disorder. Based on similar feasibility studies (Hoffmann et al., 2018), up to 20 participants were contacted. Participants were paid £20 (~US $27) for their time to take part.

Inclusion criteria were age over 18, self-reported clinical diagnosis of GAD from a general practitioner (GP) or mental health professional, able to read and understand English, and access to an iPhone (6 or above). All participants were given a participant information sheet

and provided consent via the consent form. Ethical approval was obtained from the University of Exeter Research Ethics Committee (eUEBS003011).

Procedure

At baseline, demographic information and the baseline questionnaire data were collected. Participants were then given a unique activation code and sent a video and study information pack to provide a clear overview of what they would be required to do (interact with the app every day for at least 5 minutes a day, using whichever tools they feel comfortable with, and complete all 6 ACT modules in the 2-week period), as well as relevant contacts for research-related questions and mental health emergencies. At 2 weeks after app download, exit questionnaire data were collected and access to the app ceased.

Treatment dosage was defined as completion of all 6 ACT modules and participants were excluded if they did not complete all modules. Participants were given the option to take part in semistructured interviews with an experienced psychologist (JK) after the exit questionnaire was completed.

Duty of Care

As this population is at risk of severe mental health issues, a duty of care protocol was implemented if the research team felt there was a significant concern for the participant's welfare. In the demographics section, each participant was required to input details of an emergency contact and his/her GP; participants were excluded from the study if they failed to provide this information. If on either the baseline or exit questionnaire participants stated they had thoughts of self-harm or suicide or both over the past 2 weeks, the research team would contact the at-risk participant via email with relevant helpline information, the participant's emergency contact by email with relevant information on how to support someone during a mental health crisis, and also the participant's GP via a letter with information on the study and questionnaire results.

Outcome Measures

Psychological flexibility, or the ability to be aware of and deal with difficult emotions while acting in accordance with one's values, was measured by the 7-item Acceptance and Action Questionnaire-II (AAQ-II), which utilizes a 7-point Likert response scale (scores range 10-70 with lower scores indicating greater psychological flexibility) (Bond et al., 2011).

Mental well-being was assessed using the 14-item Warwick-Edinburgh Mental Well-being Scale (WEMWBS), which utilizes a 5-point Likert scale for responses (scores range from 14 to 70, with higher scores indicating greater mental well-being) (Tennant et al., 2007).

The 7-item Generalized Anxiety Disorder Assessment (GAD-7) [66] and the 9-item Patient Health Questionnaire (PHQ-9) [67] are clinical assessment tools for measuring anxiety and depression, respectively, over the previous 2 weeks. GAD-7 scores range from 0 to 21 and PHQ-9 scores from 0 to 27; both have a cutoff of 10, indicating moderate symptoms which warrant referral to a mental health professional.

Feedback questions were asked as part of the exit questionnaire, regarding learned concepts, any behavior change, and any benefits received as a result of the intervention. The System Usability Scale (SUS) was included as part of the exit questionnaire to assess usability of the app; it assesses the ability of the participant to effectively and efficiently complete tasks using the system, as well as the participant's satisfaction in using the app. The scale provides a single number and corresponding grade (Lewis, 2018).

Optional semistructured interviews were conducted to further explore feasibility of the intervention in the participant's own words, in terms of usability of the app and acceptability of the ACT-based content (ie, if the participants felt the content was relevant to the anxiety they struggle with). Questions were based around downloading the app and setting up the paired wearable, accessing the modules, experience using the tools and features (ie, usability),

and experience during and after completion of the modules and if any aspect of the content triggered anxiety (ie, acceptability).

Statistical Analysis

To explore Aim 1 (evaluate the acceptability and usability), the SUS was reported, and a thematic analysis of the feedback questionnaires and semistructured interviews were conducted.

To explore Aim 2 (exploratory analysis to evaluate the preliminary efficacy of this intervention), differences between baseline and exit scores of the AAQ-II, WEMWBS, GAD-7, and PHQ-9 were analyzed. Because of small sample size, nonparametric paired-sample Wilcoxon signed-rank tests were carried out using the R statistical language. Differences were considered significant if $P<.05$. Effect sizes were calculated as $Z/\sqrt{N}$ and interpreted in accordance with Cohen's classification of effect sizes (ie, 0.2 [small effect], 0.5 [moderate effect], 0.8 [large effect]).

Patient and Public Involvement

Reported here against the GRIPP-2 Short Form checklist (Staniszewska et al., 2017). This study used the Person-Based Approach (PBA; Yardley et al., 2015) to ensure the ACT-based digital intervention was relevant, acceptable, and usable to adults with diagnosed Generalised Anxiety Disorder. Phase 1 of the study comprised qualitative interviews and think-aloud sessions with members of the target population. Findings were used to generate personas representing the target users and a set of guiding principles that informed the design of the six ACT modules and their integration with the BioBase platform; specific design choices shaped by this work included the modular ACT content structure, the language and framing of psychoeducational material, and the integration of body-awareness exercises with passive biometric monitoring. The PBA differs from patient and public involvement in the formal GRIPP-2 sense: it generates personas and guiding principles from target-user qualitative work, but does not give users decision-making authority over intervention design (Bate & Robert, 2006). This study represents the closest the empirical work of this thesis came to formal user involvement, but it remains short of the GRIPP-2-aligned standard; experience-based co-design (s. 9.4) is proposed as a more participatory alternative for future iterations. See s. 8.5.5 for the thesis-level reflection.

Results

Participants

A total of 13 participants met inclusion criteria, were eligible to take part in the study, and agreed to do so. Seven participants failed to meet inclusion criteria, as they ceased communication before the study could begin (n=3) or did not download the app (n=4). Three participants were excluded from the analysis because they did not complete all 6 modules (1 participant completed no modules and 2 participants completed 1 module). The final sample consisted of 10 participants; of these 10, 7 participated in the semistructured interview after the exit questionnaire. CONSORT flow chart is presented in Figure 1. Demographics are presented in Table 2.

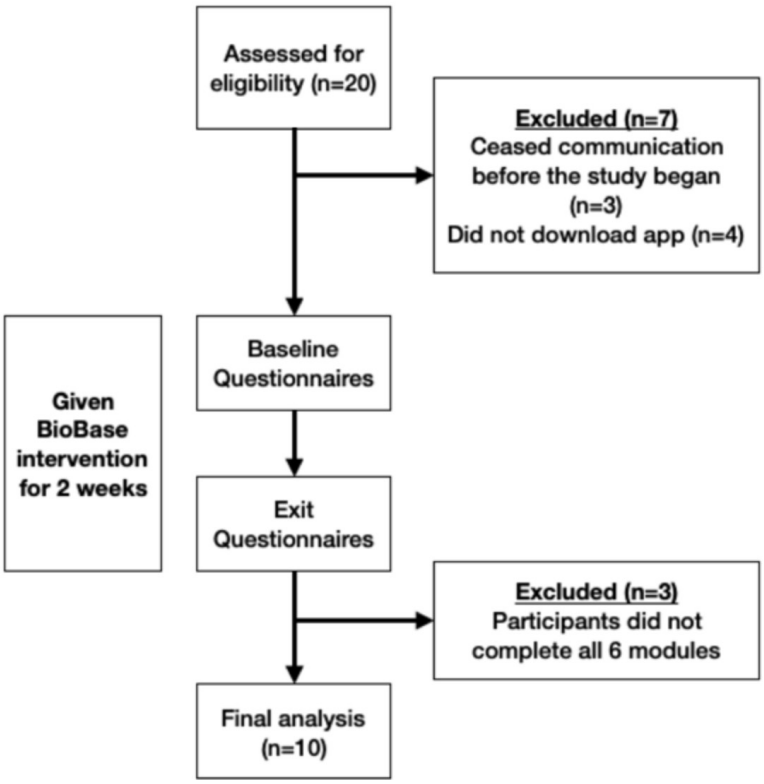


Figure 1. Study flow chart.

Variable	Value
Age (years), mean (range)	25.2 (19-48)
Gender, n	
Female	10
Male	0
Education, n (%)	
School to age 16	2 (20)
College/A-levels to age 18	7 (70)
Undergraduate degree	1 (10)
Undertaking therapy/counseling, n (%)	
Yes	2 (20)
No	8 (80)
On medication for anxiety, n (%)	
Yes	8 (80)
No	2 (20)

Table 2. Baseline characteristics of the study sample (N=10).

Duty of Care

Seven participants responded to item 9 of the baseline or exit PHQ-9, indicating they had thoughts of self-harm/suicide in the previous 2 weeks; the participant, the participant's emergency contact, and the participant's GP were then contacted as per the duty of care protocol.

Engagement

All participants onboarded on to the app over 2 days and engaged with the app on average 5.8 minutes per day (median 6.0, IQR 4.9-7.4 minutes). The instructions to participants were to complete all 6 modules within 2 weeks, and there was considerable variability in the spacing between module completion (Figure 2); some participants completing 1 or 2 modules per day (n=6), some completed up to 3 in 1 day (n=2), and some completed all 6 modules in 1 day (n=2).

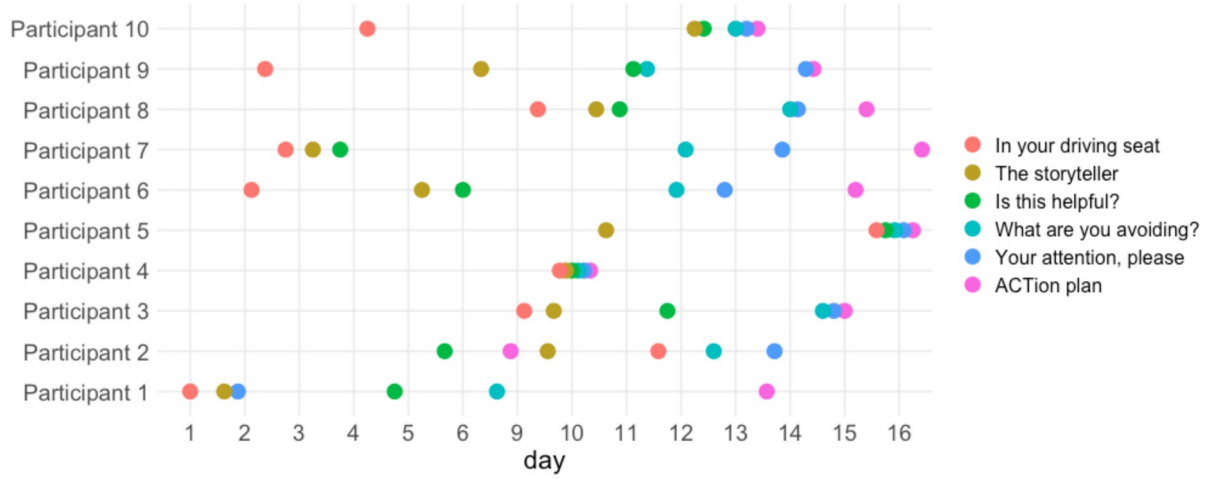


Figure 2. Completion order of ACT-based modules by participant. Participants onboarded onto the app over 2 days and had access to the app for 14 days.

The most-used tool was the mood declaration (EMA); every participant completed at least seven declarations over the study (mean 13.1, median 13, IQR 9.3-14). As much as 9/10 participants (90%) also completed the deep breathing exercise (mean 2.4, median 2, IQR 1.3-2.8) and 6/10 participants (60%) completed the body scan exercise (mean 1.4, median 1, IQR 0-1.8).

Usability

Results from the SUS indicate an average A rating (mean 83.4, median 88.7, IQR 75-96.9). All 7 participants that had an exit interview stated downloading the app and going through the app onboarding process including pairing the wearable were simple and none reported problems. Three participants stated they felt they needed to wear the wearable a bit too tight or that it was uncomfortable at times.

Acceptability

All 7 participants that engaged in exit interviews stated that the ACT modules were relevant to the anxiety they currently struggled with. Two participants stated they had engaged in therapy before and, therefore, the ACT modules were not new concepts. All others stated the modules were relevant to, helpful, and useful for their anxiety. One participant noted the ease

of use with each module being “quite short and didn’t take long to complete.” Two participants stated addressing overthinking and challenging thoughts (“The storyteller” and “Is this helpful?”) were particularly useful. Another 2 participants mentioned “What are you avoiding?” as the most useful module. One participant preferred the journaling activities on paper, rather than in the app.

Participants stated the tools were also relevant, useful, and helpful to their anxiety. Three participants stated the deep breathing exercise was therapeutic, especially before bed. Two participants specifically found the mood tracker helpful for their anxiety, stating “it made me stop and think a bit more about how I’m feeling.” One participant mentioned the body scan as the most useful tool as she preferred body awareness mindfulness techniques over breathing-related relaxation exercises.

Overall, 2 participants stated that by “using an app every day to check in with myself, it’s much easier to control my anxiety” and that “having goals set by the app and a module to complete each day was helping my mental state ... makes me look at [a feeling], acknowledge it and move on from it easier.”

Preliminary Efficacy

Outcome Measures

Table 3 shows the descriptive statistics for baseline and exit for scores on psychological flexibility (AAQ-II), mental well-being (WEMWBS), and symptoms of anxiety (GAD-7) and depression (PHQ-9).

On average, participants had lower median scores for GAD-7 and PHQ-9 at exit than at baseline, indicating fewer anxious and depressive symptoms, respectively, and these differences were statistically significant (P=.03 and .008, respectively, for GAD-7 and PHQ-9; Table 3). For the GAD-7, the median exit score (9.8) fell below the threshold for moderate anxiety (score of 10) from a median baseline score of 12.8 (Figure 3).

Preliminary efficacy in participants (N=10) of the digital intervention. The clinical cutoff for GAD-7 and PHQ-9 for moderate anxiety or depression, respectively, is 10.

Measure	Baseline	Exit	Statistical significance
AAQ-II[a], median (range)	30.5 (23.0-48.0)	29.0 (17.0-47.0)	W[e]=30; *P*=.11; ES[f]=0.512
WEMWBS[b], median (range)	36.5 (17.0-58.0)	38.0 (22.0-56.0)	W=41.5; *P*=.55; ES=0.630
GAD-7[c], median (range)	14.0 (0-20)	9.0 (1-19)	W=34.0; *P*=.03[g]; ES=0.691
PHQ-9[d], median (range)	11.5 (1.0-26.0)	9.5 (0-25.0)	W=45; *P*=.008[h]; ES=0.840

[a]AAQ-II: Acceptance and Action Questionnaire-II.
[b]WEMWBS: Warwick-Edinburgh Mental Well-being Scale.
[c]GAD-7: 7-item Generalized Anxiety Disorder Assessment.
[d]PHQ-9: 9-item Patient Health Questionnaire.
[e]W: paired-sample Wilcoxon signed-rank tests.
[f]ES: effect size.
[g]*P*<.05.
[h]*P*<.01.

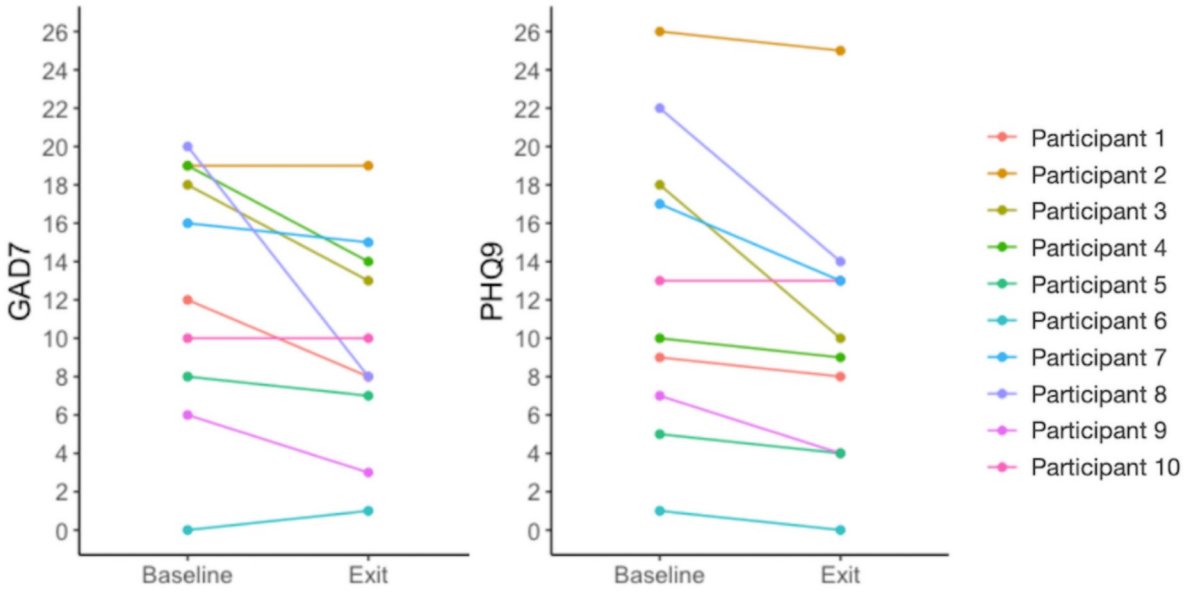


Figure 3. GAD-7 and PHQ-9 results in participants (N=10) of the digital intervention. The clinical cutoff for moderate anxiety or depression, respectively, is 10.

Participants had lower median AAQ-II scores at exit than at baseline, indicating more psychological flexibility and better functioning, and higher median mental well-being scores, although this difference was not statistically significant (P=.11 and .55 for AAQ-II and WEMWBS, respectively; Table 3).

An additional analysis was conducted on 13 participants, which included 3 participants that did not complete all 6 modules, to determine if excluding these participants had a significant impact on the results. However, these results remain significant for change in GAD-7 (n=13; baseline median 12, exit median 8, Wilcoxon W=53, P=.011, d=0.81) and PHQ-9 (n=13; baseline median 10, exit median 8, Wilcoxon W=45, P=.008, d=0.84).

The large effect sizes observed (d = 0.69-0.84) are comparable to those found in the student and workplace populations (Chapters 2-3), suggesting consistent therapeutic benefits across different clinical and non-clinical groups.

Interviews

Using the exit questionnaires, participants reported that the primary learned concept was that mental health consists of a number of factors; 3 participants specifically mentioned learning how much their sleep patterns play an important role. This was investigated further in interviews with participants.

Two participants mentioned they learned the most from the “Is this helpful?” module; 1 participant said, “how to divide your thoughts into those that are helpful and not helpful ... a way I haven’t looked at it before… even now, since I’ve stopped using it, if I have a thought, I think ‘what have you just gained from that thought?’ and if it’s not a productive thought, I need to either eradicate it or turn it into one. And I just found that really useful for a positive mindset”.

In terms of behavior change after the intervention, participants mentioned being more aware of their sleep patterns, consistently doing deep breathing and body scan exercises (without the guide on the app), checking in more with how they are feeling, and trying to make their goals more specific to avoid being overwhelmed.

Discussion

Principal Findings

This pilot study aimed to describe the development of a self-guided ACT-based digital prototype for individuals with GAD and to provide a feasibility analysis (ie, evaluate the acceptability, usability and preliminary efficacy) of the digital intervention. The study found that the content was acceptable to this population, and highly regarded due to its relevance for the participants' struggle with anxiety. The intervention was judged to be usable in a digital format, with an "A" rating on the SUS scale. Additionally, preliminary evidence suggests that this intervention may reduce symptoms of anxiety and depression.

Acceptability

Participants stated that they found the app and ACT-based content useful, easy to complete, and relevant to the anxiety they struggle with. Modules from all 3 pillars of ACT (Aware, Open, Active) were also mentioned as being useful. The therapeutic tools for mood tracking, deep breathing, and mindfulness alongside the novel ACT content were also mentioned as being helpful for managing symptoms. Additionally, using the app daily made it easier for participants to control their anxiety, acknowledge their mental state, and take actions that move them closer to achieving their goals. It is important to note the participants had been diagnosed with GAD and were not treatment naïve; 2 participants did not find the concepts novel. However, being reminded of these concepts was found to be helpful, thereby demonstrating the utility of the intervention.

Usability

This study found the app-based therapeutic to have an SUS rating of 83.5, which equates to an A rating, representing programs that people are likely to recommend to their friends (Lewis, 2018). This is comparable to web-based ACT programs (J. N. Kelson et al., 2017; Levin et al., 2014; Viskovich & Pakenham, 2018). Three participants commented that the wrist-worn

wearable was uncomfortable at times, but this did not affect their judgment of the intervention as simple and easy to use.

Engagement

Of the 13 initial participants, 10 (77%) were included in the final analysis (3 participants did not complete all modules). This was comparable to another self-guided study using a 9-module, 2-week online program (75%; (J. N. Kelson et al., 2017)) with similar financial incentives for the participants to take part, and on the higher end of other guided internet-based treatment programs (average adherence rates of 31%, range of 2%-83%) (Melville et al., 2010). The dropout rate of this self-guided digital intervention (23%) is slightly higher than those of face-to-face, therapist-led ACT programs (15.8% (Ong et al., 2018) and 17.35% (Karekla et al., 2019)).

Adherences rates in self-guided app-based ACT interventions are not consistently reported, therefore putting engagement results within the context of the wider literature is difficult. Across internet-based ACT treatments, 1 review found the average attrition rate was 19.2% (J. Kelson et al., 2019), but made no separate analysis between guided (n=13) and unguided (n=5) interventions. Where higher dropout rates are observed (Batterham et al., 2021; Eysenbach, 2005), these could be explained by a lack of technical knowledge, lower motivation to engage with treatment, and poor usability of the system (Hoffmann et al., 2018). While this study found 65% (13/20) uptake rate, encouragingly, this pilot study indicates system usability was high and the adherence rate was on the higher end of the expected range.

Analysis of Initial Efficacy

Although only a pilot acceptability study, the results provide preliminary evidence that this app-based therapeutic may be efficacious in reducing clinical symptoms in GAD. In this small sample, the results show trends of increasing psychological flexibility and well-being and decreasing symptoms of anxiety and depression. The large effect sizes found in this study are

comparable to similar studies investigating digital ACT-based interventions on anxiety and depression (Dahlin et al., 2016; J. N. Kelson et al., 2017; Levin et al., 2014), but due to the low sample size these results are in need of replication with larger sample sizes.

Symptoms of anxiety and depression were significantly decreased after 2 weeks (P=.03 and .008 for GAD-7 and PHQ-9, respectively). Mean GAD-7 scores decreased by 3 points (median 5 points) and mean PHQ-9 scores decreased by 2.8 points (median 2 points), but this may not represent a clinically significant change, as participants on average are at the cusp of the cutoff of 10, indicating moderate symptoms and a recommendation of a referral to a mental health professional. However, the reduction in anxiety and depression scores allude to the transdiagnostic effect of the treatment on comorbid symptoms (Hayes et al., 2016) and highlights that a therapeutic solution for GAD should address the high comorbidity of anxiety and depression.

This study did not find a statistically significant increase in psychological flexibility (AAQ-II, P=.11), which is the primary objective of ACT-based interventions; however, given the small sample size, it is likely the study was underpowered to detect this effect. Future research with an increased sample size could investigate the mediating effect of psychological flexibility on anxiety and depression scores.

Limitations

The feasibility phase of this study was limited by its small sample size and relatively short intervention period, and so conclusions are made with caution. Even though GAD is more common in females (Ruscio et al., 2017), the study was limited by a sample composed only of females. Future research should include measures of ethnicity (Borgogna et al., 2020) and aim for a sample more representative of the population with GAD. Compared with other PBA studies (Bradbury et al., 2019; Howarth et al., 2019; Rowsell et al., 2015), the small sample size used in this study to represent the target population means that study may not have

covered all context-specific behaviors that users with GAD may experience. In addition, the intervention was a prototype (6 modules) of a larger intervention (30 modules), and therefore only investigates the initial feasibility of the intervention concept. The intervention also comprises features that are known to decrease symptoms of anxiety, and these effects cannot be interpreted separately. Further research with a larger feasibility study on the 30-module intervention, that is more generalizable to a wider population of individuals with GAD, is warranted.

In addition, participants were paid to take part, meaning that engagement rates with the ACT modules in uncompensated participants are unknown. In a previous study using BioBase content related to workplace stressors, paid participants showed higher engagement in the program. However, there were no differences seen in the effect on outcome measures of well-being and anxiety. Future research on engagement should aim to use a similar incentive scheme to particular usage cases.

Although the prototype design process listed desired outcome behaviors of the intervention, the approach lacked clearly stated behavior change techniques for each of the desired behaviors (Abraham & Michie, 2008). Future iterations of the program will need to address this to increase engagement and efficacy of the intervention. In addition, barriers to use need to be further investigated, including the restrictions of using a technology-based solution as a therapeutic. For example, individuals with GAD may find technology useful to help soothe anxiety symptoms; however, technology can also be a tool to provide unhelpful distraction from difficult thoughts and feelings. In addition, technology can be used to connect with supportive social networks; however, feelings of being overwhelmed due to social media use can also increase symptoms of anxiety.

While this study demonstrated feasibility and acceptability, we did not directly measure interoceptive accuracy or sensibility, which retrospectively appears to be a critical oversight given the ACT emphasis on present-moment bodily awareness. This limitation informed the subsequent development of interoception measures described in later chapters.

Conclusions

The PBA approach is designed to complement and enrich the evidence-based approach to intervention design. This chapter shows that the design and development of a persona and incorporation of intervention design objectives and key features alongside the app of an ACT-based model can be a feasible solution within an app-based therapeutic. However, in keeping with the PBA approach, more research is needed to further define the engagement criteria for an efficacious product and further develop and tailor the ACT intervention to the target GAD population in keeping with the iterative development–evaluation–development cycles of the PBA. The emphasis on body awareness through features like the body scan exercise and real-time biometric feedback may have enhanced participants' interoceptive awareness—a mechanism that became increasingly central to understanding the intervention's effects and ultimately led to the interoception measurement studies described in Chapters 5-7.

5. Chapter 5: CARED Task: Ecological Assessment of Cardiac Interoception

Content adapted from: Ponzo S, Morelli D, Suksasilp C, Cairo M, Plans D (2021). Measuring Interoception: The CARdiac Elevation Detection Task. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2021.712896

A note on contributions. This study was published with Sonia Ponzo as first author (Ponzo et al., 2021), reflecting her lead on the writing. I co-conceived the ecological momentary assessment approach used in the CARED task, contributed to the programming of the smartphone application — including the heart-rate detection algorithm and the notification system — designed the data-collection infrastructure for continuous monitoring over four weeks, conducted the statistical analyses including the signal-detection-theory calculations, and led the writing of the methodological sections of the manuscript.

Abstract

Interoception has increasingly become the focus of psychiatric research due to its hypothesized role in mental health, yet existing interoceptive tasks suffer from methodological limitations that impair their validity and limit their usage in vulnerable populations. This study presents the development and validation of the CARdiac Elevation Detection (CARED) task, designed to measure interoceptive accuracy ecologically using smartphone-delivered notifications combined with continuous heart rate monitoring via wearable devices. Thirty healthy participants wore a smartwatch collecting continuous heart rate data over four weeks while receiving notifications asking them to judge whether their heart rate was higher than usual, rate their confidence, and describe their recent activities. Notifications were sent according to an algorithm that inversely weighted probability by empirical heart rate frequency, ensuring adequate sampling across different heart rate states. Responses associated with high-intensity physical or emotional activities were excluded to minimize inference-based judgments. Using Mann-Whitney U test analyses comparing performance to chance distributions, approximately one-third of the sample (N=30) was classified as interoceptive, consistent with previous literature. Participants demonstrated

overall good insight into their interoceptive abilities with AUROC values significantly different from chance ($t_{29} = 4.96$, $p < 0.001$, $d = 5.28$; $M = 0.58$, $SD = 0.09$). Interoceptive awareness, measured through confidence ratings, also showed above-chance performance. No correlations were found between objective CARED performance and subjective interoceptive measures (Interoceptive Accuracy Scale and Body Perception Questionnaire), supporting the dissociation between objective and subjective components of interoception. The CARED task provides a significant methodological advance for psychiatric and developmental research, demonstrating that cardiac interoceptive accuracy can be reliably assessed in naturalistic environments.

Introduction

As discussed in Section 1.2.1 of this thesis, existing cardiac interoceptive tasks suffer from methodological limitations that impair their validity and limit their usage in vulnerable populations. The therapeutic interventions described in Chapters 2-4, which combined biometric feedback with psychological content, highlighted the importance of understanding individual differences in how people perceive and interpret their bodily signals—yet no ecologically valid measures existed to assess these differences in real-world contexts.

While psychophysics variants such as the method of constant stimuli (MCS) and 6-alternative-forced-choice (6AFC) designs have attempted to address some limitations (Brener & Kluvitse, 1988; Brener & Ring, 2016; Clemens, 1984; Ring & Brener, 2018; Yates et al., 1985), none of these tasks is suitable for measuring interoception over time in an ecologically valid fashion. Critically, these laboratory-based tasks only assess interoception in one state of physiological arousal, typically at rest. Given the evidence for intra-individual covariation in cardiodynamics and interoceptive accuracy (G. E. Jones & Hollandsworth, 1981; Schandry et al., 1993), a priority in interoception research is to develop a task that captures individual differences across an ecological range of cardiac activity rather than at complete rest.

Moreover, remotely deployed and simple procedures for measuring interoception present a particularly valuable tool for clinical and developmental research, where complex tasks like the MCS are difficult to administer.

Furthermore, tasks measuring interoception over time allow us to capture different dimensions of interoceptive abilities. A recent model proposed by Murphy et al. (Murphy, Catmur, et al., 2019) highlights the theoretical importance of distinguishing between subjective and objective interoceptive measurements as well as accuracy and attention, proposing that the underlying structure of interoception may be a 2 × 2 factorial one. Traditional tasks aim to measure only interoceptive accuracy; however, none incorporates a measure of interoceptive attention—the extent to which interoceptive signals are the object of one's attention (also referred to as interoceptive "sensibility"; (Garfinkel et al., 2015; Khalsa et al., 2018)). While measuring individuals' abilities to accurately detect and classify interoceptive signals is crucial for assessing potential interoceptive disruptions, capturing natural tendencies to pay attention to interoceptive stimuli may prove particularly useful for training paradigms aimed at improving interoceptive abilities in everyday life.

This chapter presents data from a relatively small sample (N = 30) of healthy participants to illustrate a novel ecological interoceptive task, the CARdiac Elevation Detection (CARED) Task, aiming to capture different components of objective interoceptive accuracy and elements of attention throughout a period lasting a minimum of 4 weeks. Participants wore a device collecting continuous heart rate data for the whole study duration. This device (a smartwatch) was connected to a mobile application that sent notifications to participants according to a predefined algorithm. Participants were asked to state whether their heart rate was higher than usual, report their confidence in their judgments, and freely describe their activities in the 30 minutes prior to receiving the notification. Responses related to high-intensity activities (e.g., exercise) or highly emotional states (e.g., crying or fighting with a

loved one) were discarded from subsequent analyses to minimize the confounding effect of participants potentially using knowledge about their own heart rate rather than basing judgments on perception.

Participants were classified as either interoceptive or non-interoceptive according to how their performance compared to a distribution of reference. Consistent with the literature (Brener & Ring, 2016), it was predicted that approximately one-third of the sample would be classified as interoceptive. Their insight into their own performance (interoceptive awareness) was assessed using confidence ratings to predict correct and incorrect responses, with the expectation that higher confidence would be associated with correct responses. Furthermore, potential dissociations between objective and subjective interoceptive abilities were explored using established self-report measures: the Interoceptive Accuracy Scale (Murphy et al., 2020) and the Body Perception Questionnaire–awareness subscale (Porges, 1993). Finally, relationships between mental health components (anxiety, depression, and stress) and interoceptive abilities were explored to better define potential interoceptive disruption in subclinical manifestations.

Materials and Methods

Pre-registration

This study was not pre-registered. At the time of conduct (recruitment 2019–2020), pre-registration of interoception measurement studies was not yet routine practice in the field, and the study was conducted as a first-pass development of a novel ecological measure (the CARED task) for which the analytic strategy — including the AUROC-based confidence–accuracy analysis — was iteratively refined alongside the task itself. The research questions, outcome measures and analytic approach reported in this chapter correspond to those documented in the published version of the work (Ponzo et al., 2021). The absence of pre-

registration is acknowledged as a transparency limitation, with cross-thesis implications discussed in s. 8.3.

Participants

Fifty-two healthy participants were recruited via social media adverts. Participants were excluded if they were pregnant, or with diagnosed psychiatric or neurological disorders. Of the recruited participants, 10 became unresponsive prior to receiving the wearable device and 12 withdrew their participation. Hence, the final sample comprised 30 participants (16 females, age range: 18–51, M = 27.43, SD = 9.01). All participants signed electronic informed consent and the study was approved by a local ethics committee.

Materials

A smartphone application was developed in React Native (TypeScript/JavaScript) for the User Interface and native Android code (Kotlin) for Bluetooth communication with the wearable device. The wearable device (manufactured at Transtek Medical Electronics) collected heart rate data by photoplethysmography sensors (see (Cropley et al., 2017; Morelli et al., 2018) for validation data), with the average heart rate over 1 min of heart activity being measured every 10 min. Finally, at baseline, participants completed a demographics questionnaire (age, gender, height, and weight) as well as the following questionnaires: the Interoceptive Accuracy Scale (IAS;(Murphy et al., 2020)), the Body Perception Questionnaire (BPQ – awareness subscale; Porges 1993), the State-Trait Anxiety Inventory 6 items short form (STAI-6 SF; (Marteau & Bekker, 1992)), and the Depression Anxiety Stress Scale 21 items (DASS-21; (P. F. Lovibond & Lovibond, 1995)).

Design and Procedure

Participants were asked to continuously wear the wrist device for the whole duration of the study and were informed that they would be required to pay attention to the app notifications

between 9 am and 9 pm. When a notification was triggered by the algorithm (the frequency of which is explained below), participants were prompted to first answer a yes/no question (i.e., “Is your heart beating faster than usual?”), followed by a confidence judgment expressed via a 10-points Likert scale (Figure 1). They were then asked to freely describe what they were doing in the half-hour prior to answering the question. The answers to the free text were coded according to the activity participants declared they were engaged in and were either labeled as high intensity (e.g., physical intensity, such as “workout” or “exercise,” as well as emotional intensity, such as “stressed,” “upset”). These labeled answers were then excluded from the analysis to avoid any confounding effect deriving from knowledge of one’s own heart rate. Participants were made aware that notifications would occur both when their heart-rate was elevated as well as when it was normal.

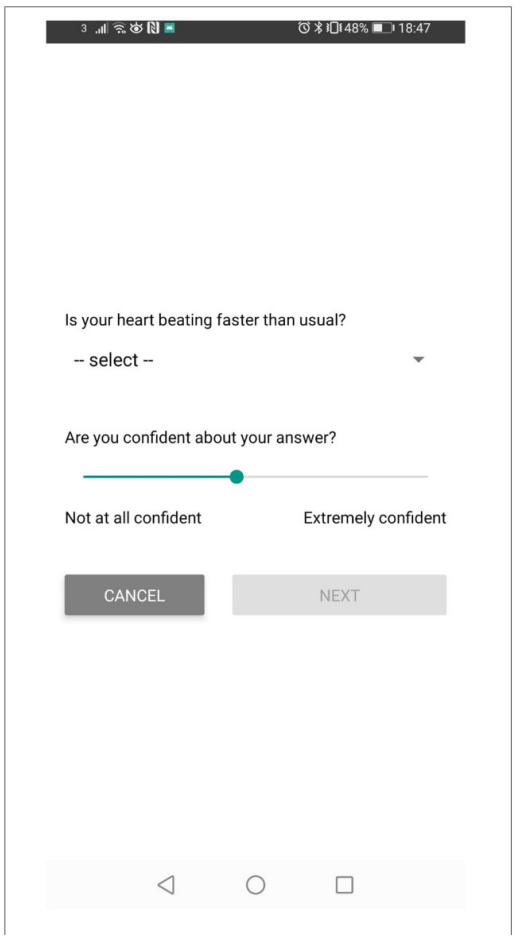


Figure 1: Example of a question screen comprising the heart rate judgment and the confidence rating. Following this screen, participants were presented with a free text box and asked to report what they were doing in the 30 min preceding the notification.

Notifications were sent to the notification center of the participant's mobile at most once every 10 min (i.e., when the heart rate data was obtained from the wearable device). The probability of receiving a notification was dependent on a set of parameters: notifications sent each day to participants should not exceed six per day on average and they should be representative of the full range of heart rate values throughout the day. To this end, the probability was set to be inversely proportional to the empirical frequency of the current heart rate for that user, given the time of the day. Specifically, each day was split into segments of 4 h each and the possible values of heart rate were divided into buckets, each corresponding to a range of 5 beats per minute (BPM) each. Each new heart rate measurement was then assigned to the corresponding segment of the day and heart rate bucket. The empirical frequency of the new heart rate measurement within each segment and bucket was then computed as follows:

$$empirical\ frequency = \frac{heart\ rate\ measurements\ within\ day\ segment\ and\ bucket}{heart\ rate\ measurements\ within\ day\ segment}$$

The probability of sending a notification to the participant was then calculated as being inversely proportional to the empirical frequency (if the resulting value was greater than one, the notification was sent with probability equal to one). To obtain an average of approximately six notifications per day, the constant of proportionality was fixed to a base probability of 20% (a value obtained heuristically) multiplied by the ratio of the number of desired notifications (6) and the number of heart rate measurements per day (72, given by six measurements per hour multiplied by the 12 h included in the 9 am to 9 pm range). Therefore, the final probability of sending a notification to a participant was computed as:

$$P = 20\% \times \frac{6}{6 \times 12} \times \frac{1}{\textit{empirical frequency}} = 1.67\% \times \frac{\textit{heart rate measurements within day segment}}{\textit{heart rate measurements within day segment and bucket}}$$

To compute sensible probabilities at the beginning of the experiment, each participant started with heart rate measurements set to an a priori value. Specifically, in each segment of the day, the count was initialized to 1 measurement per bucket, for all buckets from 40 BPM to 90 BPM, and 0 measurements for all other buckets. These values were then updated with each new empirical measurement.

At the end of the study, participants were asked to answer a set of questions aimed at assessing participants' strategy when replying to the questions throughout the study (see section "End of Study Questions" in Appendix B).

Data Analysis

Responses to the heart rate judgment were analyzed separately for each participant to obtain a summary statistic using the Mann–Whitney U test. Specifically, given a response (A) where a participant reports perceiving their heart rate as higher than usual and a second response (B) where the same participant reports their heart rate as normal, if the recorded heart rate associated with the answer A is higher than the recorded heart rate associated with response B, then A and B formed a concordant pair. In the opposite case, i.e., where the recorded heart rate associated with A is lower than the one associated with B, these two responses constituted a discordant pair. The Mann–Whitney U test does not require any particular assumption about the probability distribution of heart rate values and participants' answers, as long as they are independent under the null hypothesis and higher heart rate values increase the probability of positive answers under the alternative hypothesis. This is especially relevant since heart rate values were not sampled uniformly: the sampling depended both on the

notifications schedule as well as on the number of completed trials. The test used the proportion of concordant versus discordant pairs (effect size), and compared it against the distribution of reference (according to the null hypothesis in which answers were independent of heart rate, and given the total number of pairs), generating a p-value for each participant. If such value was lower than 0.05 (a threshold yielding 5% false positive rate) that participant was deemed interoceptive. Effect sizes were calculated as the difference between concordant and discordant pairs, divided by the total number of pairs, and constituted participants' interoceptive scores. Empirical cumulative distributions were calculated and plotted to illustrate the results.

As for confidence judgments, a threshold approach was implemented given no threshold-free method was identified. Participants' average heart rate throughout the study was obtained. Data points higher than the computed standard deviation from averaged heart rate for each participant were considered higher than usual (and normal otherwise). Hits, false alarms, correct rejections, and misses were calculated by comparing whether participants' judgments (i.e., the yes or no response to the question "Is your heart beating faster than usual?") were in line with measured heart rate (i.e., higher than usual or normal; see Table 1).

Table 1: Summary of type 1 responses based on participants' answers to the question "Is your heart beating faster than usual?"(yes/no answer) and whether their heart rate was detected as higher than usual.

	Yes	**No**
Higher heart rate	Hit	Miss
Normal heart rate	False alarm	Correct rejection

The response type (i.e., correct when participants' answer was a hit or a correct rejection, incorrect otherwise) and the confidence judgments (as a predictor) were used as arguments for

the ROC function (pROC package for R) and area under the curve (AUC) values were extracted. A one-sample t-test was then conducted to check whether these AUROC values significantly differed from 0.5 (i.e., chance). Empirical cumulative distributions were calculated and plotted to illustrate the results.

To explore the relationship between subjective and objective interoceptive accuracy, a Pearson correlation between IAS scores and the effect sizes obtained from the Mann–Whitney U test described above was carried out. Furthermore, a correlation between such values and BPQ scores was conducted in order to investigate the relationship between subjective and objective interoceptive attention.

Finally, exploratory correlational analyses with Pearson correlations, aimed at identifying potential relationship between interoceptive abilities and mental health (i.e., DASS-21 and WEMWBS), as well as investigating the assumed relationship between interoception and BMI, were carried out.

Data was analyzed using Python and R, and plotted using base R.

Patient and Public Involvement

No formal patient and public involvement (PPI) was undertaken in the development of the CARdiac Elevation Detection (CARED) task, reported here against the GRIPP-2 Short Form checklist (Staniszewska et al., 2017). The CARED task was designed by the research team to address methodological limitations of laboratory-based interoception measures identified in the existing literature (s. 1.2.1). No structured consultation with people with lived experience of conditions associated with disrupted interoceptive processing — anxiety, depression, eating disorders, or alexithymia — was undertaken at the design or validation stage. The absence of PPI in measurement-tool development is recognised as a limitation, particularly given the thesis's framing of these measures as potentially informing personalised digital therapeutics for these populations (s. 8.2.3 and s. 9.4). See s. 8.5.5 for the thesis-level reflection.

Results

Out of 1576 total trials, 181 trials, associated with high intensity activities, were excluded from the analyses. The minimum number of trials per participant in the overall sample was

eight and the maximum 149 (M = 58.57, SD = 37.99), with only three participants answering fewer than 15 trials.

In terms of interoceptive accuracy, 1/3 of the sample (9/30 participants) obtained a p-value below the 0.05 threshold (Figure 2A). This result is in line with previous findings showing that approximately 1/3 of healthy participants is classified as interoceptive in interoceptive accuracy tasks (Brener and Ring, 2016). Given that participants had varying amounts of trials, the p-value (which takes into consideration the number of trials) was used to determine whether participants were interoceptive or not rather than relying on effect sizes. A small effect size was detected in slightly more than 1/3 of the sample and a medium effect size in 4/30 participants (Figure 2B).

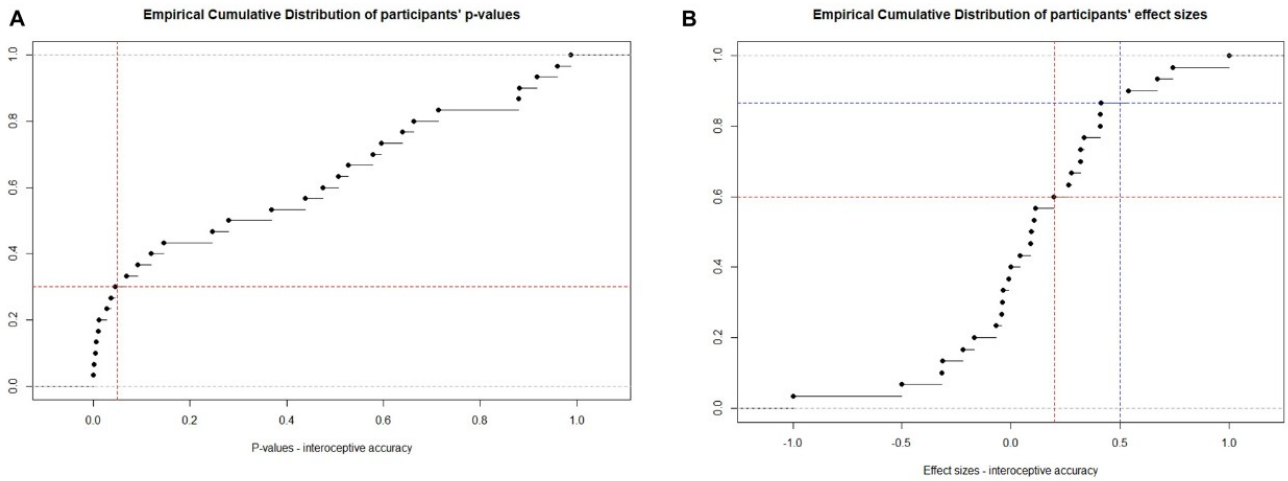


Figure 2: Empirical cumulative distribution (ECD) of (A) p-values obtained after comparing the Mann–Whitney U statistics against the population of reference and (B) effect sizes. (A) Red line vertical = 0.05, horizontal = 0.05 of the ECD; (B) Red line vertical = 0.2 (small effect size), horizontal = 0.2 of the ECD; Blue line vertical = 0.5 (medium effect size), horizontal = 0.5 of the ECD. Gray lines: upper and lower boundaries of p-values.

A one sample t-test revealed that participants' interoceptive awareness, indexed via AUROC values, was statistically different from chance ($t29 = 4.96$, $p < 0.001$, $d = 5.28$; $M = 0.58$, $SD = 0.09$). As can be observed in Figure 3, in 4/30 participants, higher confidence ratings signaled decreased likelihood of correctly judging their heart rate as being higher than usual. As for the rest of the sample, their AUROC values suggested that higher confidence ratings were able to predict correct responses, in line with the notion that participants have insight into their interoceptive processes.

None of the correlations between interoceptive accuracy (effect sizes) and subjective interoception (IAS and BPQ), BMI as well as self-reported measures of mental health was significant (see Table 2, also including descriptive statistics of all measures).

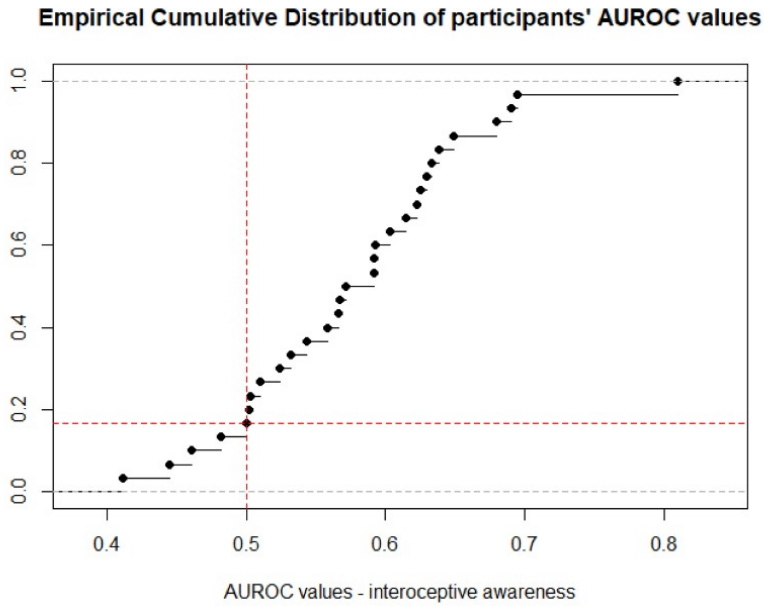


Figure 3: Empirical cumulative distribution of participants AUROC values obtained by applying the ROC function to correct and incorrect responses predicted by confidence ratings in such answers. Red lines: vertical = 0.5, horizontal = 0.5 of the ECD. Gray lines: upper and lower boundaries of effect sizes.

Variable	M	*SD*	R	*p*
BMI	22.12	3.23	-0.35	0.06
BPQ	68.77	18.11	-0.29	0.12
IAS	81.63	9.95	-0.18	0.33
WEMWBS	48.67	6.90	0.07	0.70
DASS-21 - Stress	6	3.43	-0.33	0.07
DASS-21 - Anxiety	3.4	2.19	-0.17	0.37
DASS_21 - Depression	4.3	3.72	-0.30	0.10
STAI	12.47	3.17	0.27	0.15
CARED p values	0.37	0.12	-	-
CARED effect sizes	0.34	0.39	-	-

Table 2: Behavioural and questionnaires descriptive statistics and correlations between interoceptive accuracy (effect sizes) and BMI and self-reported measures of interoception and mental health.

Discussion

The current study presented the development and proof of concept testing of an ecological test of interoceptive abilities, the CARED task. Approximately 1/3 of the participants, in line with the literature (Brener and Ring, 2016), were classified as interoceptive. In terms of interoceptive awareness, the majority of the sample showed insights into their performance, with higher confidence predicting increased likelihood of correct responses. Exploratory analyses aimed at identifying relationships between objective and subjective components of interoception, as well as self-reported mental health components, did not yield significant results.

We classified participants as interoceptive following a threshold-based approach due to the novelty of the task and the varying number of trials participants completed. Accordingly, the comparison of the obtained Mann–Whitney U statistic to the distribution of reference is dependent on the number of trials, whereby fewer trials make it more difficult for participants to be classified as interoceptive (correctly or not). However, instead of an arbitrary threshold, methods such as Gaussian mixture models or clustering algorithms may be preferred in future studies with a bigger sample size, allowing for a data-driven classification of participants between interoceptive and non-interoceptive.

By sampling cardiac judgments over time and "day-to-day" fluctuations in the cardiovascular system, this novel task captures interoceptive ability across an ecological range of afferent cardiac signals. This is a strength over existing procedures, such as the HCT and HDT, which could only index interoceptive accuracy at particular states of the cardiovascular system (e.g., at complete rest or after exercise). The CARED task thus successfully demonstrated that cardiac interoceptive accuracy can be measured ecologically, addressing the critical limitations identified in Section 1.2 of this thesis. Despite this, research on the role of interoception in psychopathology has largely relied on the HCT and HDT. Therefore, a priority for future research is to assess whether existing associations between interoception and emotional constructs still hold when interoception is assessed in ecological conditions.

Hypotheses Emerging from the AUROC Distributions

The distribution of AUROC values warrants closer attention than the central tendency alone permits, because the shape of the distribution generates testable hypotheses about how interoceptive awareness — the meta-cognitive correspondence between confidence and accuracy (Garfinkel et al., 2015) — might vary systematically across individuals and clinical groups.

First, the four participants whose AUROC values fell below .5 represent an "inverse mapping" profile in which higher confidence predicted incorrect heart-rate judgements rather than correct ones. Whether this profile is noise (potentially driven by the limited number of valid trials some participants completed) or a stable individual difference cannot be determined here, but it is theoretically intriguing. Inverse mapping is what one would predict for individuals who are hypervigilant to bodily sensations but systematically misinterpret them — a candidate signature of anxious interoceptive processing as theorised by Paulus and Stein (2010), and a plausible profile for presentations in which subjective awareness becomes uncoupled from objective signal. The hypothesis is that this AUROC < .5 subgroup, if replicated in a larger sample, would be enriched for clinical anxiety, panic disorder, or related dissociation-spectrum symptoms relative to the well-calibrated majority.

Second, the AUROC profile itself is a candidate phenotyping variable. Future studies could compare AUROC distributions across clinical groups — anxiety disorders, depression, eating disorders, alexithymia — to test whether inverse-mapping rates differ in theoretically motivated directions, and whether the shape of the AUROC distribution (rather than mean accuracy) carries diagnostic information that group-level accuracy comparisons miss.

Third, the AUROC profile may itself predict response to biometric feedback interventions of the kind evaluated in Chapters 2–4. Individuals whose confidence is well-calibrated to their accuracy may benefit from biometric feedback because the feedback maps coherently onto their existing self-model; individuals with inverse mapping may, paradoxically, experience increased confusion or distress when feedback contradicts their confident misjudgements. This is a falsifiable hypothesis that connects the measurement work of this chapter to the broader thesis argument about interoceptive phenotyping for personalised digital therapeutics (developed further in s. 8.2).

The present sample (n = 30) is far too small for any of these subgroup or between-groups analyses; the observations above are intended only to generate hypotheses for future, adequately powered work.

In order to exclude the potential confounding factor of inferring one's own heart rate according to the activity one is involved in or the emotional state one is experiencing, we eliminated all answers associated with free text entries indicating high physical or emotional intensity in the 30 min preceding participants' judgment. Furthermore, asking questions throughout the day and with an unpredictable pattern makes it difficult for participants to make their judgments solely based on estimation. Even though these countermeasures ensured adequate control for obvious activities and emotional states, it is not possible to completely exclude the possibility that participants' judgment was not entirely based on perception but also on inference of one's own states. However, the trade-off between measurement accuracy and the ecological nature of the task makes the CARED a valid method to continuously collect interoceptive data with minimal input from participants. The obvious advantage of such a simple methodology is the ease of use in clinical populations as well as in developmental studies.

In terms of interoceptive awareness, out of 30 participants, four presented with a profile indicating an inverse relationship between confidence and accuracy. The rest of the sample, however, behaved in line with the expectation that higher confidence is associated with correct responses (and vice-versa). In the current study, due to the low sample size, it wasn't possible to investigate differences in interoceptive awareness in interoceptive versus non-interoceptive participants (e.g., Garfinkel et al., 2015). Future research is needed to elucidate the relationship between interoceptive performance and insight.

The lack of correlation between objective and subjective components of interoception corroborates the idea that these components dissociate and tap into different aspects of interoception (Murphy et al., 2019, 2020). The CARED task taps into both accuracy, as well as attention, components of interoception due to its ecological nature, thus making it more difficult to fully explore the relationship between the four components of the model. Future studies could focus on better distinguishing between objective accuracy and attention, perhaps by pairing the CARED with a task capturing time spent focusing on interoceptive sensations (e.g., experience sampling methods; Murphy et al., 2020). It has to be noted, however, that these findings may be due to the small sample size of the current study and hence further research is needed.

We did not find any correlation between interoceptive accuracy and self-reported measures of subjective well-being nor mental health dimensions. Whilst interoceptive disruption has been hypothesised as a candidate mechanism in several mental and physical health disorders (Dunn et al., 2010; Caspi et al., 2014; Quattrocki and Friston, 2014; Barrett and Simmons, 2015; Murphy et al., 2017; Khalsa et al., 2018), the empirical evidence remains mixed — most notably, a recent meta-analysis (Adams et al., 2022) failed to find an association between anxiety and cardiac interoceptive accuracy as measured by the heartbeat counting and detection tasks — and further research is needed to clarify the exact relationship between interoceptive accuracy and subclinical manifestations.Furthermore, in the current study, participants were not screened for mental health disturbances prior to taking part in the study. Whilst this ensured a sample of healthy participants, it led to smaller between-subjects variations in mental health profiles. Future studies could employ stratification techniques to ensure subgroups of participants in different mental health categories. Furthermore, data-driven classifications between interoceptive and non-interoceptive participants could shed more light into differences in self-reported mental health subgroups.

When we looked at differences in heart rate judgments between participants who reported checking their pulse throughout the study and participants who did not, we found that only one of the interoceptive participants declared having checked their pulse a minimal portion of the time (roughly 10%). Nevertheless, manually checking one's own pulse may not necessarily be related to one's ability to detect changes in heart rate throughout the day and further research is needed to assess the impact of such strategies on interoceptive accuracy.

One direction for future research using the CARED task is to investigate whether prior familiarity with knowledge of one's heart rate may improve task accuracy, speculatively by providing a more precise reference against which to compare heart rates during each trial. Secondly, some studies have reported that interoceptive ability can be trained, observing gains in accuracy following a heartbeat detection task with feedback (Quadt et al., 2021; Sugawara et al., 2020). Whether the CARED task accuracy can improve with either prior experience with one's heart rate, or frequent monitoring of heart rate over several weeks or months, remains to be empirically observed.

In conclusion, we presented preliminary evidence that the CARED task can measure interoceptive abilities ecologically and represents a potential asset for psychiatric and developmental research. This advance is particularly relevant for understanding the role of interoception in digital therapeutic interventions like those described in Chapters 2-4, where individual differences in bodily awareness may moderate treatment response. Future research should examine whether CARED performance predicts therapeutic outcomes and whether interoceptive accuracy can be improved through targeted training.

6. Chapter 6: The Phase Adjustment Task: Smartphone-Based Assessment of Cardiac Interoception

Adapted from: ***Plans, D.****, Ponzo, S., Morelli, D., Cairo, M., Ring, C., Keating, C. T., Cunningham, A. C., Catmur, C., Murphy, J., & Bird, G. (2021). Measuring interoception: The phase adjustment task. Biological Psychology, 165, 108171. https://doi.org/10.1016/j.biopsycho.2021.108171*

A note on contributions. I led this study. I was the primary developer of the Phase Adjustment Task concept and methodology, created the smartphone application implementing the cardiac-phase-adjustment interface, developed the photoplethysmography-based heartbeat-detection system, performed the data analyses including the consistency-scoring algorithms, and was the primary author of the manuscript.

Abstract

As outlined in the introduction to this thesis, interoception has increasingly been the focus of psychiatric research due to its hypothesized role in mental health, but existing interoceptive tasks either suffer from important methodological limitations, impacting their validity, or are burdensome and require specialized equipment, limiting their usage in vulnerable populations. This study reports the development and validation of the Phase Adjustment Task (PAT), a novel measure of cardiac interoceptive accuracy. Participants' heart rate is recorded using a wearable device connected to a mobile application, and auditory tones are presented that are triggered by heartbeats but out of phase with them. Participants adjust a virtual dial to synchronize the tones with their perceived heartbeat across multiple trials. Study 1 (N=124) demonstrated that PAT performance, assessed through the consistency of phase selections, could reliably distinguish between interoceptive and non-interoceptive individuals using Bayesian classification approaches, with 34-48% of participants classified as interoceptive depending on the statistical threshold. Critically, PAT performance was unaffected by physiological characteristics including heart rate variability, heart rate, body mass index, or the bodily location where heartbeats were perceived. Study 2 (N=90) revealed moderate test-retest reliability (ICC=0.508) over one week, consistent with theoretical models proposing both trait and state components of interoceptive ability. Given its ease of administration and

accessibility via smartphone technology, the PAT has significant potential as an asset for psychiatric and psychological research investigating interoception's role in mental health.

Introduction

As discussed comprehensively in Section 1.2.1 of this thesis, existing cardiac interoceptive tasks—particularly the heartbeat counting task (HCT; (Dale & Anderson, 1978); (Schandry, 1981)) and heartbeat detection tasks (HDT)—suffer from important methodological limitations that impair their validity and limit their usage in vulnerable populations. The role of interoception in mental health processes, including its proposed function as a transdiagnostic risk factor (the P-factor hypothesis discussed in Section 1.2.3.2; see also (Brewer et al., 2021)), underscores the critical need for improved measurement approaches.

Despite the importance of accurately measuring interoception for understanding its role in psychopathology (as reviewed in Sections 1.2.3.2 and 1.3), current tasks remain problematic. The limitations of the HCT regarding belief influences and physiological confounds (detailed in Section 1.2.1) and the challenges with 2AFC heartbeat detection tasks—particularly their failure to account for individual differences in preferred synchrony delays—necessitate novel approaches. Furthermore, as discussed in Section 1.3.3, the need for ecologically valid, remotely deployable measures has become increasingly apparent, particularly for clinical populations who may struggle with laboratory-based assessments.

The Phase Adjustment Task (PAT) was developed to address these specific limitations through several technical innovations. Unlike tasks that rely on counting or binary synchrony judgments, the PAT presents participants with auditory tones that are triggered by their heartbeats but out of phase with them. Participants adjust a virtual dial to change the phase relationship between the tones and their heartbeat until they perceive synchrony. This approach eliminates several critical confounds: participants cannot rely on knowledge of their resting heart rate (as in the HCT), are not required to count or track heartbeats (avoiding

mathematical confounds), and the task accommodates individual differences in preferred synchrony timing by allowing continuous phase adjustment rather than forcing binary choices at predetermined delays.

The PAT's implementation via smartphone technology offers significant advantages for psychiatric and developmental research. The task requires only a smartphone with a camera capable of photoplethysmography—technology now ubiquitous in modern devices—making it accessible to diverse populations without specialized equipment. This accessibility is particularly important for vulnerable clinical populations who may find laboratory visits challenging due to anxiety, mobility limitations, or geographical constraints. Furthermore, the smartphone-based approach enables ecological momentary assessment of interoceptive abilities in naturalistic settings, addressing the ecological validity concerns raised in Section 1.3.3 regarding laboratory-based measures.

We conducted two studies to validate the PAT. Study 1 examined whether the PAT could reliably distinguish between interoceptive and non-interoceptive individuals using a data-driven classification approach. We hypothesized that PAT performance, assessed through the consistency of phase selections across trials, would identify a distinct subgroup of interoceptive individuals. Critically, we predicted that PAT performance would be independent of physiological variables that confound traditional measures, including heart rate variability (HRV), heart rate, body mass index, and the bodily location where heartbeats were perceived. Study 2 assessed test-retest reliability over one week, with the hypothesis that the PAT would show moderate stability consistent with theoretical models proposing both trait and state components of interoceptive ability (Garfinkel et al., 2015).

Given the mixed evidence regarding relationships between interoception and psychopathology in non-clinical samples (reviewed in Section 1.2.3.2), we also explored associations between

PAT performance and measures of anxiety, depression, and stress. Based on recent meta-analytic findings suggesting no simple linear relationship between cardiac interoceptive accuracy and anxiety in non-clinical populations (Adams et al., 2022), we predicted no significant associations between PAT performance and psychopathology measures in our non-clinical sample, though we remained open to the possibility that the PAT's improved methodology might reveal relationships obscured by confounds in traditional tasks.

Study 1

Pre-registration

Neither of the two studies reported in this chapter (Study 1: validation and confound analysis; Study 2: one-week test–retest) was pre-registered. At the time of the original PAT development work, pre-registration of novel interoception measurement studies was not yet routine practice in the field, and the analytic approach to PAT scoring evolved over the course of the work — the updated scoring procedure used here is described in Palmer et al. (2025). The hypotheses, design, outcome measures and analytic approach reported in this chapter correspond to those documented in the published version of the work (Palmer et al., 2025). The absence of pre-registration is acknowledged as a transparency limitation, with cross-thesis implications discussed in s. 8.3.

Materials and methods

Participants

124 healthy participants (excluding those pregnant, or with psychiatric or neurological diagnoses) were recruited via Prolific (prolific.co) and successfully completed the task (70 females, age range: 18–63, M = 26.5, SD = 8.02). Ethical approval was obtained from the local ethics committee and informed consent obtained from participants.

Test delivery

A smartphone application developed in swiftUI by BioBeats Ltd (now Huma Therapeutics) was used to administer the PAT task. The application uses a camera-driven photoplethysmogram sensor that detects heartbeats when participants place their finger over the camera (for details and validation see (Cropley et al., 2017); (Morelli et al., 2018)). In addition to technical validation, this method has been demonstrated to be of equivalent accuracy when compared to lab-based methods (electrocardiogram and pulse oximeter) when recording heartbeats and does not produce a heartbeat sensation in the finger due to pressure as certain pulse oximeters may do (Murphy, Brewer, et al., 2019).

Questionnaire measures were completed via an online platform (Qualtrics) and consisted of the Depression, Anxiety and Stress Scale - 21 items (DASS-21; Lovibond & Lovibond, 1996), Body Perception Questionnaire - Awareness subscale (BPQ-A; Porges, 1993), Interoceptive Accuracy Scale (IAS; Murphy et al., 2020) and Toronto Empathy Questionnaire (TEQ; (Spreng et al., 2009)). The DASS-21 has three subscales assessing feelings of depression, anxiety and stress over the previous week and the total score ranges from 0 to 21. Each subscale contains 7 items scored on a 4-point scale, ranging from 0 (“Did not apply to me at all”) to 3 (”Applied to me very much or most of the time”). The DASS-21 has been shown to have acceptable reliability in a non-clinical US sample (Sinclair et al., 2012) and good reliability, construct and structure validity in an international cohort (Bibi et al., 2020). The Body Perception questionnaire - Awareness subscale is part of a longer tool assessing several components of body perception. The awareness subscale contains 26 items, scored on a 5-point Likert scale ranging from 1 (“Never”) to 5 (“Always”). The Awareness subscale is part of the BPQ-Short Form, containing the Awareness and Autonomic nervous system reactivity subscales, which has strong reliability and validity (Cabrera et al., 2018). The Interoceptive Accuracy Scale aims to capture self-reported accuracy of the perception of

interoceptive signals. It contains 21 items rated on a scale from 5 (“Strongly agree”) to 1 (“Strongly disagree”), with scores ranging from 21 to 105 (where higher scores indicate better interoceptive accuracy). The IAS has good test-retest reliability and excellent internal consistency (Murphy et al., 2020). The Toronto Empathy Questionnaire, assessing individuals’ self-reported empathic abilities, contains 16-items scored on a 5-point scale ranging from 0 (“Never”) to 5 (“Always”). The TEQ has been shown to have good validity, internal consistency and test-retest reliability (Spreng et al., 2009). Demographic data, including age, gender and height and weight (used to calculate BMI) were also collected.

PAT task: design and procedure

Participants completed the entirety of the task remotely. On each trial participants listen to a series of tones (duration = 156 ms) at the same frequency as their heart rate. They are asked to rotate a virtual dial on their smartphone to advance or retard the tones until they perceive the tones to be synchronous with their heartbeats (see Supplementary Materials in Appendix C for full instructions). This is equivalent to allowing the participant to continuously adjust the delay between their R-wave and the tone in order to indicate the delay perceived as synchronous, but does not require the R-wave to be measured (just the pulse at the finger using the smartphone camera (Bacciu et al., 2018; Morelli et al., 2018). Interoceptive participants (those able to perceive their heartbeats) are identified by consistency in the delay selected across trials, as those who cannot perceive their heartbeat will choose delays at random.

Importantly, heartbeats were not used to directly trigger each tone (individuals may perceive their heartbeat before it is recorded at the finger due to the delay between heart contraction and arrival of the pulse at the finger). Instead, participants were asked to sit still with their finger on the camera while their heart rate was continuously estimated from beat-to-beat intervals every 3 s throughout the whole of the task, and this estimate was used to predict

when the next heartbeat would occur. If participants moved their finger, an on-screen prompt required them to reposition their finger before recording resumed.

The tones were presented at the heart rate frequency but out of phase with heartbeats, so that a single tone was played between each heartbeat and the next, and the delay between each heartbeat and the tone was adjusted by the participant. The starting delay (or phase asynchrony between heartbeats and tones) was randomly determined at the beginning of each trial. Adjustments to the dial resulted in a change of delay that was customized for each participant based on their heart rate - each turn of the dial advanced or delayed the tones by a set proportion of the time between heartbeats (so that those with slower heart rates would not have to work harder to make the tones synchronous with their perceived heartbeats). Participants were free to make as many adjustments to the dial as they wished, and pressed a button to confirm when they perceived the tones as synchronous with their heartbeat (Fig. 1A).

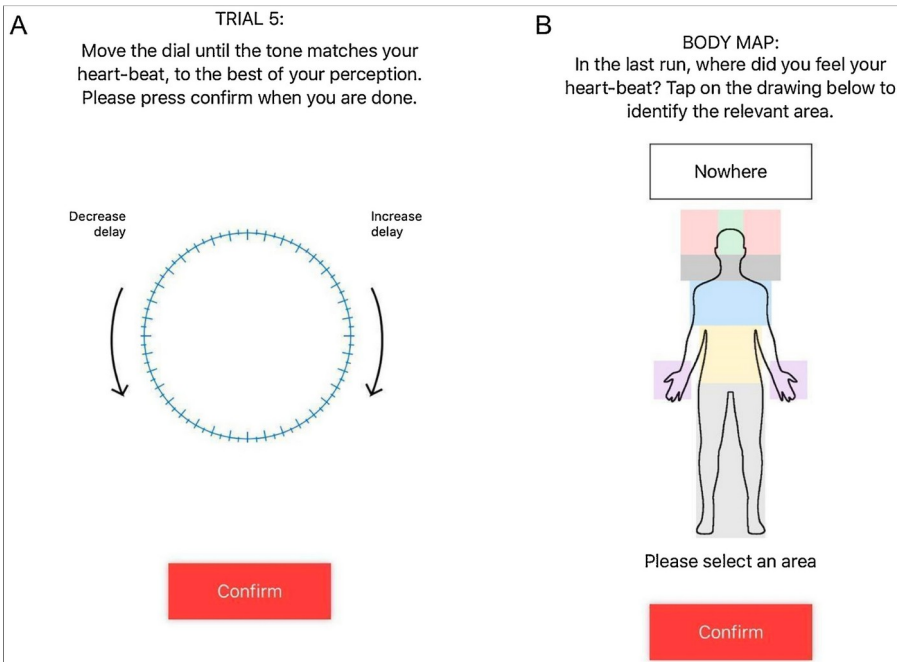


Figure 1: Screenshots of the PAT task mobile application. A) Screenshot of one of the 20 trials and B) body map screen.

Simulated data were used to assess the impact of the number of trials on expected consistency of selected delays. 20 trials constituted a good compromise between task duration and the expected consistency profile, such that the risk of false negatives was minimised (see SM for details). These 20 trials were preceded by 2 practice trials. After selecting their preferred delay on each trial, participants rated their confidence in having successfully completed the trial on a 10-point Visual-Analogue Scale (1="Not at all confident", 10="Extremely confident"). In addition, every 5 trials, participants indicated the location from which they felt their heartbeat on the previous trial using a body map (Fig. 1B). In total, the task took approximately 15−20 min.

PAT task: scoring

Note that, as in the 6AFC and MCS tasks, there is not a 'correct' response in the PAT task. Participants may perceive their heartbeat at any one of a range of delays. The measure used to discriminate between interoceptive and non-interoceptive participants is therefore the consistency of the selected phase relationship or 'delay' across trials.

Given the periodic nature of heartbeats, analysis of the consistency of selected delays should not be conducted using the standard deviation of selected delays. If, for example, a participant completes two trials at 60bpm (interbeat interval 1000 ms) and selects delays of −100 ms and 900 ms, a linear analysis of the mean and standard deviation would yield an average of 500 ms and high dispersion, with the participant classified as inconsistent (non-interoceptive). However, −100 ms and 900 ms are

the same value if the period is 1000 ms (e.g. 100 ms before heartbeats), indicating high consistency. This problem is solved if the selected delay (d) is expressed as a periodic function of the period (p) so that a delay of 0 ms is equivalent to a delay that is equal to the period $d = p$. This is implemented by expressing the delay as an arc of length $2\pi d/p$ on a unitary circumference. Mapping each participant to a standardised circumference allows individual variabilities in heart rate to be accounted for, as the space that delays are mapped onto is not influenced by differences in period length. This formalisation allows the similarity of angles to be used to define cross-trial consistency. In the previous example, delays of −100 ms and 900 ms at 60bpm would result in the same point on the circumference. In the PAT task, each movement of the dial corresponds to:

$$\alpha = d + 2\pi d/p$$

and a measure of angular similarity can therefore be used to estimate the consistency of selected delays. By expressing angles as complex numbers of modulus = 1 and argument equal to 2πd over p, consistency can be computed as the modulus of the sum of the angles divided by the number of considered angles:

$$consistency(d,p) = \frac{1}{n} mod(\sum_{j=1}^{n} e^{i2\pi \frac{d_j}{p_j}})$$

If angles are close to one another, the modulus of their average will be close to 1; if they are randomly positioned the modulus will approach 0. This consistency formula is equivalent to the Phase Locking Value (PLV; (Lachaux et al., 1999), a technique used to measure similarity between phases of periodic signals.

Tooling and data analysis

Acquisition of all data was performed using a smartphone application written in Objective-C and Swift programming languages within the XCode (version 12.5) environment and deployed using Microsoft App Centre for distribution to participants. The operating system target was iOS 10 and above, and all participant phones were iPhone 7 (and above) models by Apple (Cupertino, USA). The algorithm used for PPG detection is an implementation of the open source WABP algorithm, available on physionet (Soria, 2015). As documented by Zong et al. (2003) WABP can detect 99.31 % of the 368,364 beats annotated in an ECG, and 96.41 % of the 39,848 beats in the reference database, whereby the difference between the manually-edited and algorithm-determined ABP pulse onset was less than or equal to 20 ms.

Data were analyzed and plotted using the tydiverse, ggplot2 (Wickham, 2016), RHRV (Rodríguez-Liñares et al., 2011), AdaptGauss (Thrun et al., 2019), rstatix, ggpubr and purrr packages for R within the RStudio environment. T-tests with Welch's correction for degrees of freedom were employed to compare interoceptive and non-interoceptive participants. P-values are reported uncorrected but whether comparisons survive Bonferroni correction for multiple comparisons is noted.

Comparison against random responding

To test the null hypothesis that participants' responses reflect random behaviour, 5000 participants responding at random were simulated and the resultant distribution compared to the distribution of real responses via a Wilcoxon-signed rank test.

Classification of participants as interoceptive or non-interoceptive

In order to classify participants as either interoceptive or non-interoceptive, a gaussian mixture model with 2 mixtures was applied to the consistency values following the assumption that the population is made of two subpopulations; interoceptive and non-interoceptive participants. The gaussian mixture model returned two distributions, one for

non-interoceptive and one for interoceptive participants by means of an expectation-maximisation algorithm (Fig. 3). Z-scores were calculated for each participant for non-interoceptive and interoceptive distributions separately:

$$(z = x - \underline{x}/s)$$

The obtained z-scores were used to calculate the probability of each participant belonging to the non-interoceptive and interoceptive distributions. Specifically, the probability of being non-interoceptive or interoceptive was calculated as

$$P_{NI}(z) = 1 - F_{NI}(|z|) + F_{NI}(-|z|)$$

and

$$P_I(z) = 1 - F_I(|z|) + F_I(-|z|)$$

respectively, with F_{NI} and F_I being the cumulative distribution functions with a mean of 0 and unitary variance. $P_{NI}(z)$ and $P_I(z)$ provide the probability of obtaining a z-score at least as extreme as the value z that was observed for each participant, and $F_{NI}(z)$ and $F_I(z)$ provide the probability lower than z.

Comparison of the probability that each participant is part of the interoceptive and non-interoceptive distributions allows a Bayes Factor (BF) to be calculated as the ratio of the probability of belonging to one of the two distributions over the probability of belonging to the other distribution. Each participant can therefore be classified as more likely to be interoceptive or not interoceptive, with the BF reflecting the strength of evidence in favour of a classification. In order to allow for participants to be categorised one can apply a threshold to BFs such that participants can be classified as interoceptive, not interoceptive, or unclassified (i.e. where there is not sufficient evidence to classify a participant as either interoceptive or not). By convention, there are three thresholds that might be used for

classification; BFs greater than 3 (providing moderate evidence that a participant is interoceptive or not interoceptive), greater than 10 (providing strong evidence), and greater than 30 (providing very strong evidence). We therefore present three analyses when assessing whether interoception predicts individual difference variables relating to mental health and empathy, using each of the thresholds identified above (BFs >3, >10, >30).

Bodily locations, average delays, and task engagement

The modal location of the four bodily location reports by each participant was used to derive a summary statistic for each participant. The mode of the most selected location was used to check whether the average delay chosen by interoceptive participants corresponded to previously reported delays between tones and perceived heartbeats. Mean delays were calculated using selected delays and heartbeat periods to obtain a complex number signalling the mean delay (calculating the modulus as period/2π). To examine the impact of participant response profiles on consistency, consistency scores were correlated with the mean time spent on trials and the number of dial movements.

Physiological variables

After the heartbeat occurrences from the software photoplethysmograph have been detected, the HRV time series is calculated using the RHRV package in R. Following extraction of beat-to-beat intervals (RR), core time-domain methods are deployed to calculate time analysis statistics (HRV, SDNN, RMSSD and pNN50) as measures of heart rate variability (HRV). These are calculated from a pre-testing baseline period heart rate recording of 120 s in duration.

In order to check whether physiological variables such as heart rate, HRV and BMI differed between interoceptors and non-interoceptors at the different Bayes Factor thresholds, t-tests were conducted.

Task confidence and self-reported measures

T-tests between interoceptors and non-interoceptors at the different BF threshold were conducted to explore differences in task confidence, self-reported interoceptive accuracy, anxiety, stress, depression and empathy.

Patient and Public Involvement

No formal patient and public involvement (PPI) was undertaken in the development of the Phase Adjustment Task (PAT), reported here against the GRIPP-2 Short Form checklist (Staniszewska et al., 2017). The PAT was designed by the research team to address specific methodological confounds in existing cardiac interoception measures (heart rate knowledge, counting strategies, physiological variables). Iterative testing of the smartphone interface and task instructions during development was conducted to refine the user interface, but this constituted usability testing rather than formal PPI consultation in the GRIPP-2 sense. As with Chapter 5, the absence of formal involvement of people with lived experience in the design of this measurement tool is a recognised limitation, particularly relevant given the thesis's clinical application framing (s. 8.4.4, s. 9.4). See s. 8.5.5 for the thesis-level reflection.

Results

Comparison against random responding

Whilst the simulated data were normally distributed (Fig. 2, red line), the distribution of real consistency values was binomial (Fig. 2, black line) and only partially overlapped the simulated distribution. A Wilcoxon-signed rank test revealed the simulated and real response distributions differed significantly ($Z = 10.40$, $p < 0.0001$, $r = 0.15$), confirming that, as a group, real participants were not responding randomly. This result was confirmed using a subsequent simulation of 100,000 random responders which produced identical results.

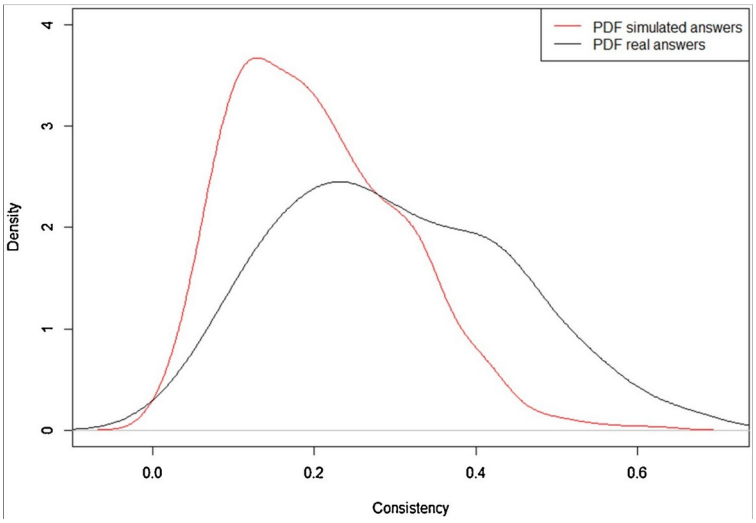


Figure 2: Probability density function of data from simulated participants responding at random (red line) and real participants' data (black line) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

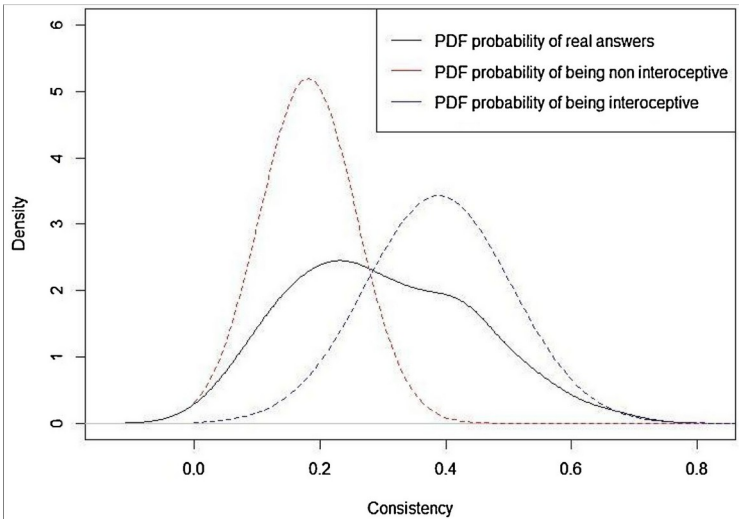


Figure 3: Probability density function of real participants' data (black line), estimated distribution of non-interoceptive participants (red line) and interoceptive one (blue line) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

Classification of participants as interoceptive or non-interoceptive

At the different Bayes Factor thresholds, 60 participants were classified as interoceptive and 44 participants classified as non-interoceptive (BF > 3), 51 participants classified as interoceptive and 30 participants classified as non-interoceptive (BF > 10), and 42 participants classified as interoceptive and 0 participants classified as non-interoceptive (BF > 30), respectively. Due to the lack of participants meeting the BF threshold of 30 to be considered non-interoceptive, the third analysis compared participants for whom there was very strong evidence they were interoceptive with all other participants. The ratio of participants classified as interoceptive with Bayes Factor thresholds of 10 and 30 is consistent with previous estimates that approximately ⅓ of healthy participants are interoceptive on multi-delay HDT tasks (Brener & Ring, 2016).

Bodily locations, average delays, and task engagement

The means and standard deviations of consistency values in the 3 most commonly selected bodily locations did not differ (chest: N = 64, M = 0.30, SD = 0.16; neck: N = 18, M = 0.32, SD = 0.14; fingers: N = 170, M = 0.28, SD = 0.13). Additionally, no difference was found in consistency values between participants declaring one bodily location and those declaring multiple locations ($t_{105.81} = 0.75$, $p = 0.45$, $d = 0.14$).

The mean delay for the most selected location by interoceptive participants (chest) was 13 ms, 23 ms, and 11 ms, for Bayes Factor thresholds of 3, 10 and 30, respectively. When added to the R-wave to finger pulse interval of 200 ms at seated rest (Obata et al., 2017), these delays fit with previous findings from the MCS that tones 200−300 ms after the R-wave are perceived as synchronous (Brener et al., 1993; Brener & Kluvitse, 1988).

Neither the mean time spent per trial, nor the number of dial movements, were significantly correlated with consistency in the overall sample ($rs<.08$, all $ps>0.42$), nor in the interoceptive subgroups ($rs<+/-0.25$, all $ps>0.121$). Furthermore, t-tests comparing interoceptive and non-interoceptive participants for these variables at each BF threshold were not significant (see Table 1, Table 2).

Table 1: Mean and standard deviations of the physiological, task engagement, self-reported interoception, and mental health variables for interoceptive and non-interoceptive participants, when participants are classified as interoceptive and non-interoceptive at Bayes Factor thresholds of 3, 10, and 30 in Study 1.

Variable	BF3	BF10	BF30
HR	Interoceptive: M = 78.33, *SD* = 9.30 Non-interoceptive: M = 75.63, *SD*=7.97	Interoceptive: M = 78.48, *SD* = 9.5 Non-interoceptive: M = 76.33, *SD*=8.61	Interoceptive: M = 78.33, *SD* = 9.97 Non-interoceptive: M = 76.83, *SD*=8.35
SDNN	Interoceptive: M = 149.19, *SD* = 87.36 Non-interoceptive: M = 175.53, *SD*=108.97	Interoceptive: M = 150.17, *SD* = 89.4 Non-interoceptive: M = 174.37, *SD*=110.81	Interoceptive: M = 152.18, *SD* = 62.62 Non-interoceptive: M = 172.26, *SD*=101.38
pNN50	Interoceptive: M = 48.11, *SD* = 19.42 Non-interoceptive: M = 44.09, *SD*=17.67	Interoceptive: M = 48.14, *SD* = 19.53 Non-interoceptive: M = 41.92, *SD*=17.78	Interoceptive: M = 47.87, *SD* = 19.45 Non-interoceptive: M = 44.66, *SD*=18.89
RMSSD	Interoceptive: M = 125.18, *SD* = 66.93 Non-interoceptive: M = 114.61, *SD*=65.33	Interoceptive: M = 124.87, *SD* = 66.85 Non-interoceptive: M = 107.9, *SD*=58.52	Interoceptive: M = 125.28, *SD* = 67.72 Non-interoceptive: M = 120.21, *SD*=63.61
BMI	Interoceptive: M = 23.28, *SD* = 3.44 Non-interoceptive: M = 23.66, *SD*=5.25	Interoceptive: M = 23.46, *SD* = 3.54 Non-interoceptive: M = 23.43, *SD*=5.59	Interoceptive: M = 23.61, *SD* = 3.81 Non-interoceptive: M = 23.68, *SD*=6.26
Dial movements	Interoceptive: M = 5.08, *SD* = 3.67 Non-interoceptive: M = 4.77, *SD*=2.69	Interoceptive: M = 5.27, *SD* = 3.82 Non-interoceptive: M = 5.05, *SD*=2.88	Interoceptive: M = 5.65, *SD* = 4.1 Non-interoceptive: M = 4.81, *SD*=3.21
Time	Interoceptive: M = 19.8, *SD* = 12.55 Non-interoceptive: M = 17.69, *SD*=10.71	Interoceptive: M = 20.24, *SD* = 13.05 Non-interoceptive: M = 18.62, *SD*=12.26	Interoceptive: M = 21.15, *SD* = 13.96 Non-interoceptive: M = 17.37, *SD*=9.86
Confidence	Interoceptive: M = 5.67, *SD* = 1.86 Non-interoceptive: M = 5.6, *SD*=1.76	Interoceptive: M = 5.52, *SD* = 1.92 Non-interoceptive: M = 5.39, *SD*=2.03	Interoceptive: M = 5.4, *SD* = 2.04 Non-interoceptive: M = 5.81, *SD*=1.6
IAS	Interoceptive: M = 82.68, *SD* = 8.33 Non-interoceptive: M = 84.57, *SD*=9.87	Interoceptive: M = 82.84, *SD* = 8.05 Non-interoceptive: M = 85.87, *SD*=8.02	Interoceptive: M = 83.26, *SD* = 7.53 Non-interoceptive: M = 83.21, *SD*=10.21
BPQ	Interoceptive: M = 131.4, *SD* = 40.37 Non-interoceptive: M = 127.16, *SD*=37.72	Interoceptive: M = 131.35, *SD* = 41.79 Non-interoceptive: M = 133.43, *SD*=37.5	Interoceptive: M = 130.74, *SD* = 40.17 Non-interoceptive: M = 129.17, *SD*=38.17
DASS-21 Anxiety	Interoceptive: M = 3.17, *SD* = 3.32 Non-interoceptive: M = 3.77, *SD*=4.05	Interoceptive: M = 2.71, *SD* = 2.69 Non-interoceptive: M = 4.2, *SD*=4.48	Interoceptive: M = 2.71, *SD* = 2.72 Non-interoceptive: M = 3.84, *SD*=3.83
DASS-21 Stress	Interoceptive: M = 6.12, *SD* = 3.95 Non-interoceptive: M = 6.16, *SD*=4.4	Interoceptive: M = 5.55, *SD* = 3.78 Non-interoceptive: M = 6.5, *SD*=4.86	Interoceptive: M = 5.36, *SD* = 3.73 Non-interoceptive: M = 6.56, *SD*=4.23
DASS-21 Depression	Interoceptive: M = 4.68, *SD* = 4.75 Non-interoceptive: M = 5.96, *SD*=4.98	Interoceptive: M = 4.37, *SD* = 4.28 Non-interoceptive: M = 6.13, *SD*=5.26	Interoceptive: M = 4.38, *SD* = 4.62 Non-interoceptive: M = 5.68, *SD*=4.71
TEQ	Interoceptive: M = 42.1, *SD* = 4.35 Non-interoceptive: M = 39.14, *SD*=7.06	Interoceptive: M = 42.25, *SD* = 4.47 Non-interoceptive: M = 38.1, *SD*=7.08	Interoceptive: M = 42.38, *SD* = 4.31 Non-interoceptive: M = 40.51, *SD*=6.53

Heart rate (HR); Heart Rate Variability components: Standard Deviation of NN intervals (SDNN), proportion of NN50 (the number of times successive heartbeat intervals exceed 50 ms) divided by the total number of NN (R-R) intervals (pNN50) and Root Mean Square of Standard Deviations (RMSSD); Questionnaires: Interoceptive Accuracy Scale (IAS); Body Perception Questionnaire (BPQ) - awareness subscale; Depression, Anxiety, Stress Scale - 21 items (DASS-21); Toronto Empathy Questionnaire (TEQ); Dial movements: mean dial movements throughout the task; Time: mean time spent on trials. M: mean; SD: standard deviation.

Table 2: T-tests of comparison between interoceptive and non-interoceptive participants on physiological, task engagement, self-report interoception and mental health variables, when participants are classified as interoceptive and non-interoceptive at Bayes Factor thresholds of 3, 10, and 30 in Study 1.

Variable	BF3	BF10	BF30
HR	$t_{99.467} = 1.592$, p = 0.115, $d = 0.31$	$t_{65.82} = 1.04$, p = 0.30, $d = 0.23$	$t_{71.107} = 1.116$, p = 0.27, $d = 0.22$
SDNN	$t_{80.114}$=-1.322, p = 0.19, $d = 0.27$	$t_{51.11}$=-1.017, p = 0.314, $d = 0.25$	$t_{89.67}$=-1.106, p = 0.272, $d = 0.2$
pNN50	$t_{97.29} = 1.1$, p = 0.27, $d = 0.22$	$t_{65.582} = 1.466$, p = 0.147, $d = 0.33$	$t_{80.651} = 0.878$, p = 0.38, $d = 0.17$
RMSSD	$t_{94.055} = 0.807$, p = 0.422, $d = 0.16$	$t_{67.532} = 1.195$, p = 0.236, $d = 0.27$	$t_{78.334} = 0.403$, p = 0.69, $d = 0.08$
BMI	$t_{69.309}$=-0.426, p = 0.67, $d = 0.09$	$t_{42.937} = 0.029$, p = 0.98, $d = 0.007$	$t_{118.36}$=-0.074, p = 0.94, $d = 0.01$
Dial movements	$t_{102} = 0.487$, p = 0.627, $d = 0.09$	$t_{74.023} = 0.303$, p = 0.763, $d = 0.07$	$t_{67.431} = 1.161$, p = 0.25, $d = 0.24$
Time	$t_{99.617} = 0.925$, p = 0.358, $d = 0.18$	$t_{64.072} = 0.55$, p = 0.58, $d = 0.12$	$t_{62.554} = 1.566$, p = 0.122, $d = 0.33$
Confidence	$t_{95.537} = 0.20$, p = 0.842, $d = 0.04$	$t_{58.19} = 0.287$, p = 0.775, $d = 0.067$	$t_{67.501}$=-1.136, p = 0.26, $d = 0.23$
IAS	$t_{83.161}$=-1.027, p = 0.31, $d = 0.21$	$t_{61.154}$=-1.637, p = 0.107, $d = 0.98$	$t_{106.68} = 0.034$, p = 0.97, $d = 0.006$
BPQ	$t_{96.146} = 0.554$, p = 0.58, $d = 0.11$	$t_{66.312}$=-0.231, p = 0.82, $d = 0.05$	$t_{79.138} = 0.209$, p = 0.835, $d = 0.04$
DASS-21 Anxiety	$t_{81.48}$=-0.81, p = 0.42, $d = 0.17$	$t_{41.575}$=-1.66, p = 0.104, $d = 0.43$	$t_{109.54}$=-1.893, p = 0.061, $d = 0.32$
DASS-21 Stress	$t_{86.76}$=-0.051, p = 0.96, $d = 0.01$	$t_{49.654}$=-0.92, p = 0.36, $d = 0.23$	$t_{92.626}$=-1.624, p = 0.108, $d = 0.30$
DASS-21 Depression	$t_{90.156}$=-1.311, p = 0.19, $d = 0.26$	$t_{51.483}$=-1.556, p = 0.126, $d = 0.38$	$t_{84.347}$=-1.475, p = 0.144, $d = 0.28$
TEQ	**$t_{66.519} = 2.464$, p = 0.016, $d = 0.52$**	**$t_{42.846} = 2.89$, p = 0.006, $d = 0.75$**	$t_{114.29} = 1.905$, p = 0.059, $d = 0.32$

Heart rate (HR); Heart Rate Variability components: Standard Deviation of NN intervals (SDNN), proportion of NN50 (the number of times successive heartbeat intervals exceed 50 ms) divided by the total number of NN (R-R) intervals (pNN50) and Root Mean Square of Standard Deviations (RMSSD); Questionnaires: Interoceptive Accuracy Scale (IAS); Body Perception Questionnaire (BPQ) - awareness subscale; Depression, Anxiety, Stress Scale - 21 items (DASS-21); Toronto Empathy Questionnaire (TEQ); Dial movements: mean dial movements throughout the task; Time: mean time spent on trials. T(df); t values and degrees of freedom. p; uncorrected p values, note none of these comparisons were significant after multiple comparisons correction. d; Cohen's d value.

Physiological variables

Physiological variables such as heart rate, HRV and BMI did not differ significantly between interoceptive and non-interoceptive participants under any of the BF thresholds (see Table 1 for means and SDs for interoceptive and non-interoceptive participants under each of the three BF thresholds, and Table 2 for the results of t-tests comparing interoceptive and non-interoceptive participants). No difference in consistency was observed between males and females ($t111.2 = 0.906$, $p = 0.3668$, $d = 0.17$).

Task confidence and self-report interoception measures

There was no difference between task confidence reported by interoceptive and non-interoceptive participants, nor was there a difference on the Awareness subscale of the Body Perception Questionnaire or the Interoceptive Accuracy Scale (see Table 1, Table 2).

Empathy and mental health

There were clear differences, in the expected direction, between interoceptive and non-interoceptive participants in empathy, with significantly higher TEQ scores for interoceptive participants across all Bayes Factor thresholds. Note, however, that these were not seen in study 2 (which it should be noted had a far smaller sample size), and thus require further investigation in subsequent studies. The mental health data were less clear; although interoceptive participants were less stressed, anxious and depressed than non-interoceptive participants, the difference between groups of participants only approached significance for anxiety under the Bayes Factor > 10 ($p = .104$) and > 30 ($p = .061$) thresholds, and stress under the Bayes Factor > 30 threshold ($p = .108$; see Table 1, Table 2 for details). It should be noted, however, that these are 2-tailed p values, and it could be argued that 1-tailed p values are more appropriate. While this argument could be made for stress, the mixed findings in the literature for anxiety (see Introduction) make this questionable for anxiety.

Study 2: consistency of PAT scores after a one week delay

Materials and methods

Participants and procedure

90 Prolific users were invited to participate in the Study. Prior to completing the PAT, participants completed a screening task that was identical to the PAT, except that each participant's heartbeats were sonified, and participants were required to synchronize the two auditory signals as in the PAT task (by changing the delay, or phase asynchrony, between the heartbeat tone and the delayed tone). This task was included in order to ensure that individual differences on the PAT could not be attributed to poor understanding of the task instructions, ability to synchronise signals, or poor motivation and/or attention. The screening task was analysed in the same way as the PAT task. 73 participants completed the screening task and were invited through Prolific to take part in two measurement occasions (T1, T2) of the PAT task. 60 accepted the invitation. Of these 60, 59 completed the PAT task at T1, and 43 at T2. After accounting for missing data and skipped trials, two further participants' datasets were discarded, resulting in 41 usable datasets (30 female, 11 male, ages 19–58, mean age = 29.7). After filtering for users who passed the screening task successfully, the population was reduced to 23 usable datasets (scr + population, 17 female, 6 male, ages 19–53, mean age = 30.6) with data at T1 and T2. The mean delay between T1 and T2 was 7.1 days. Participants were given identical instructions as in Study 1. For all participants, usage of the app was checked to ensure that the task was not completed between testing occasions.

Results

Bayes factor classification

Table 3 shows numbers of participants across Bayes Factor classifications (interoceptive, non-interoceptive, unknown) for BF > 3, BF > 10 and BF > 30 across the 23 participants in the scr + population.

Table 3: Bayes factor classification of participants successfully passing the screening test for Study 2.

	T1	T2
Bayes Factor >3		
Interoceptive participants	8 (34.8 %)	13 (56.5 %)
Non-interoceptive participants	14 (60.9 %)	9 (39.1 %)
Unknown	1 (4.3 %)	1 (4.3 %)
Bayes Factor >10		
Interoceptive participants	7 (30.4 %)	13 (56.5 %)
Non-interoceptive participants	13 (56.5 %)	8 (34.8 %)
Unknown	3 (13.0 %)	2 (8.7 %)
Bayes Factor >30		
Interoceptive participants	6 (26.1 %)	13 (56.5 %)
Non-interoceptive participants	13 (56.5 %)	2 (8.7 %)
Unknown	4 (17.4 %)	8 (34.8 %)

Consistency in PAT similarity scores from T1 to T2

Calculation of the Intraclass Correlation Coefficient was performed using the icc functions in the irr package in R (Gamer et al., 2019) across the two measurement points in those that successfully passed the screening task suggested moderate consistency in PAT scores of 0.508 (note that in the complete sample of those with usable data in T1 and T2 – i.e., including those who did NOT pass the screening task – ICC was 0.618). Paired t-test results `show a significant difference between T1 and T2, suggesting a potential practice effect ($T(22) = -2.44$, $p = .022$, $d = 0.51$), with a mean similarity score of 0.33 at T1 (SD = 0.12) and 0.42 at T2 (SD = 0.18).

Classifications across measurement points

This moderate consistency in PAT similarity scores was reflected in consistency in classification, for example, at BF > 3, across T1 and T2 five participants were consistently interoceptive, four were consistently non-interoceptive, five switched from interoceptive to non-interoceptive, five switched from non-interoceptive to interoceptive, one switched from

non-interoceptive to unknown, three from unknown to interoceptive, for a total of nine consistent participants, and fourteen inconsistent participants. Tables S1 and S2 present physiological and task-engagement variables separately for participants who remain consistent in their classification as interoceptive or not across T1 and T2, vs those who are inconsistent (Table S1), and the same variables separately for those who are consistently interoceptive (n = 4), consistently non-interoceptive (n = 4), who change from interoceptive to non-interoceptive (n = 4), from non-interoceptive to interoceptive (n = 9), from non-interoceptive to unknown (n = 1), and from unknown to non-interoceptive (n = 1) (Table S2) at BF > 3. Small sample sizes render these data unsuitable for statistical analysis.

General discussion

Study 1 outlined the development of a novel measure of interoceptive ability - the PAT task. Participants' consistency scores differed significantly from random responses, with between 34 % and 48 % of participants classified as interoceptive depending on the threshold adopted. Consistency scores did not differ according to the bodily location from which the heartbeat was perceived, nor the number of locations from which it was perceived. Neither physiological characteristics of the participants (including HRV and heart rate), nor their response profiles, related to interoceptive ability. Study 2 explored the consistency in PAT scores across a one-week delay. The Intraclass Correlation Coefficient indicated moderate consistency in scores, and this was reflected in moderate consistency in classification of participants as interoceptive or not.

No relationship between performance on the PAT task and confidence was observed, but it is worth noting that the confidence question was poorly specified; future versions of the PAT task will employ separate questions concerning confidence in heartbeat perception and confidence in task performance. Notably, previous studies of the relationship between self-reported and objective measures of interoception produced mixed results (Murphy, Catmur, et

al., 2019), and a lack of a relationship is consistent with proposals that interoception is a multidimensional construct, with dissociations between subjective and objective aspects (Garfinkel et al., 2015; Murphy, Catmur, et al., 2019). This may also explain why Interoceptive Accuracy Scores and Body Perception Questionnaire did not dissociate between interoceptive and non-interoceptive participants, although a recent study has highlighted problems with how participants interpret questions on the Body Perception Questionnaire (Gabriele et al., 2022), which may also explain a lack of association with PAT scores on this measure.

Currently, the PAT task classifies participants as interoceptive, not interoceptive, or unclassifiable. Individual differences in consistency scores are not treated as meaningful, except for the evidence they provide for participant classification. This is because if there is no statistical evidence that a participant can perceive their heartbeat (i.e., if the evidence that they are interoceptive is not sufficiently greater than that they are not-interoceptive), then whether that participant's consistency score is greater than another participant's consistency score (for whom there is also insufficient evidence that they are either interoceptive or not-interoceptive) is not useful information. Conversely, if two participants achieve a consistency score indicating that they are interoceptive, there is currently no way of knowing whether the difference in their scores reflects anything other than motivation, attention to the task, or timing ability (note that these factors cannot produce interoceptive performance in the absence of heartbeat perception). The validity of classifying participants as interoceptive or not - instead of using continuous consistency scores - is supported by the proportion of participants who cannot be classified under each threshold. These participants vary in their consistency scores but not to a degree that enables them to be classified as either interoceptive or not with a sufficient degree of certainty. The classification of participants as either interoceptive or not is motivated by the fact that, when tested at rest, participants can either

feel their heartbeat or they cannot. Unless the heartbeat is at the very limit of being perceptible, it is unclear why some heartbeats should be perceived and some not (the basis of individual differences in HCT scores). Instead, we suggest that individual differences in interoception should be considered as the percentage of time, or variety of circumstances, in which one can perceive interoceptive signals in everyday life. Such a suggestion is consistent with evidence of state effects on interoception (Murphy, Catmur, et al., 2019; Wittkamp et al., 2018), and the variance in PAT classification observed in Study 2. Future research should therefore seek to examine performance on the PAT across situations and time, coupled with tests of interoception in domains other than cardiac (e.g., respiratory, physiological arousal). It should be clear then, that our use of the terms interoceptive and non-interoceptive relate to a participant's performance at a particular time, and for the cardiac signal alone. It should also be noted that interoceptive participants could perceive their heartbeats when completing the PAT task which provides an external signal of their heartbeats, whether they can perceive their heartbeats without such a signal is unknown. Indeed, the consistency of an individual's interoceptive ability across time, situations, and domains is likely to be more useful than exactly how many heartbeats they count in a given period when testing theorised relationships between interoception, health and cognition. It is of interest therefore, that a significant correlation was observed between the change in PAT similarity scores from T1 to T2 in Study 2 and change in HRV(RMSSD) in those participants who were inconsistently classified as interoceptive or not ($r(12) = .573$ $p = 0.034$). The change in RMSSD suggests a state change between T1 and T2 occasions, as RMSSD is the primary time-domain measure used to estimate the vagally-mediated changes reflected in HRV, and more accurately represents parasympathetic nervous system activity than other HRV variables. This potential state change and its relationship to cardiac interoception warrants further investigation with much larger sample sizes. It is especially important to determine how much of the inconsistency in

PAT scores is a product of the test itself, or reflective of true variation in the ability to perceive one's heartbeat across time (of course, ideally this would require an additional valid test of cardioception). Of interest also is testing the stability of PAT scores when tested under laboratory conditions using a smartphone as instantiated here, and using an ECG to trigger tones. Furthermore, it would be interesting to compare stability when administered remotely via smartphone to participants who are more naive than the somewhat 'professional' participants available for recruitment via Prolific.co. Once the test-retest reliability of the PAT task is established, this will inform future studies of links between cardioception and individual differences variables so that they are appropriately powered.

With respect to task validation, the absence of an effect of physiological factors is encouraging. This is particularly true for heart rate variability which, if especially extreme, could have been problematic as it would have made heartbeats difficult to predict. Previous studies of interoception have controlled for physiological factors statistically when they correlate with interoceptive ability (Brener & Ring, 2016; Murphy et al., 2018) - but the absence of physiological effects on the PAT task makes this unnecessary. The smartphone data collection has been extensively validated (Cropley et al., 2017; Morelli et al., 2018; Murphy, Brewer, et al., 2019), however the PAT task can be conducted using any objective method of heartbeat detection (for example an electrocardiogram) in situations in which smartphones cannot be used.

The absence of robust correlations with mental health variables, or with self-reported interoception, could be considered evidence that the PAT task has low validity. However, as noted above, current evidence suggests that self-report measures of interoception do not typically predict interoceptive accuracy (see Murphy et al., 2020 for discussion), and studies of the link between interoception and mental health have produced variable results (see Introduction for review of studies concerning anxiety, and also (Desmedt, Van Den Houte, et

al., 2020), and these studies have typically used interoception tests of questionable validity. For example, when it comes to the link between interoception and depression, although depression is often associated with poor cardiac interoception (Furman et al., 2013; Lackner & Fresco, 2016; Pollatos et al., 2009; Terhaar et al., 2012); (for a review see (Eggart et al., 2019)), this relationship may be non-linear (Dunn et al., 2007), and there have been some failures to replicate this association (e.g.,(Dunn et al., 2010; Mussgay et al., 1999; Shah et al., 2016; Van der Does et al., 1997)). It must also be acknowledged that the lack of a relationship between PAT scores and mental health variables may reflect the fact that a community sample was tested, and none of the participants had scores in the clinical range on anxiety and depression measures - a limitation of this study.

With respect to task validity, it is unlikely that the PAT incorrectly classifies non-interoceptive participants as interoceptive, but classifying interoceptive participants as non-interoceptive is possible on any task if a participant has low motivation or attention. Furthermore, individuals with poor timing ability may also perform poorly on the PAT task even if they can perceive their heartbeat (although timing ability would have to be very poor indeed to be problematic). The development of the screening task for Study 2 guards against these possibilities, however, by requiring participants to successfully complete a task that is structurally identical to the PAT task, but which does not require any ability to perceive heartbeats. Any problems with attention or motivation, or poor timing ability, would mean that participants would not be able to pass the screening task. Given that the PAT task is designed to enable remote testing, any experimenters with concerns even with participants who successfully pass the screening task may wish to observe task completion via video chat software.

Problematically, given the questionable validity of the HCT and HDT tasks, they cannot be used to validate the PAT task. It is therefore encouraging that the delays at which participants

judge heartbeats as synchronous with the tones on the PAT task matched those previously observed with the MCS and 6AFC tasks (Brener et al., 1993; Brener & Kluvitse, 1988; Yates et al., 1985). The task is also far more scalable and accessible for clinical populations than currently available measures.

The PAT directly addresses the measurement gap identified through the therapeutic interventions in Chapters 2-4. The informal feedback from BioBase users who could not reconcile biometric data with their emotional states highlighted a critical need: to measure individual differences in interoceptive ability that might moderate treatment response. The PAT's smartphone implementation enables this assessment to be integrated directly into digital therapeutic platforms, allowing for baseline interoceptive profiling and personalized intervention strategies.

Crucially, the PAT's independence from heart rate variability is particularly relevant for understanding responses to biometric feedback interventions. BioBase provided users with HRV data, yet traditional interoceptive measures are themselves confounded by HRV. The PAT thus enables researchers to distinguish genuine interoceptive difficulties from physiological artifacts—a distinction essential for understanding why some participants struggled with biometric feedback while others did not.

The ecological validity of the PAT also aligns with the person-based intervention approach used in developing digital ACT (Chapter 4). Just as these interventions were designed to work within users' daily lives, the PAT enables interoceptive assessment in naturalistic contexts where therapeutic interventions are actually deployed. This methodological coherence between assessment and intervention represents a significant advance for digital mental health research.

Most importantly, the PAT provides the tool needed to test whether interoceptive ability moderates treatment response to biometric-based interventions—the hypothesis that emerged from observing differential responses to BioBase. Future studies can now examine whether baseline PAT performance predicts who will benefit from biometric feedback versus who requires alternative therapeutic approaches, enabling more targeted intervention allocation. The convergent validity established in Chapter 7 further supports using the PAT for such clinical stratification.

7. Chapter 7: Validating the PAT: Convergent and Discriminant Validity Evidence.

*Adapted from **Plans, D.**, Bird, J. M., Morelli, D., Tam, R., Keating, C. T., Palmer, R., Adams, K. L., Piercy, T., Bird, G., & Murphy, J. (in preparation). Assessing the relationship between the Phase Adjustment Task and the Novel Phase Selection Task.*

A note on contributions. This study was a collaborative effort with co-authors from the Murphy and Bird laboratories. I co-designed the validation study protocol, including the pre-registration, programmed both the PAT and the Phase Selection Task applications for cross-validation, developed the exteroceptive control task used to assess discriminant validity, conducted the correlation analyses and the non-parametric statistical tests, and contributed equally to the preparation of the manuscript with co-authors.

Abstract

The Phase Adjustment Task (PAT) has been described as one of the most promising new measures of cardiac interoceptive accuracy, yet its relationship with other cardiac interoceptive accuracy tasks remained unexamined. This pre-registered study explored the relationship between the PAT and a novel measure of cardiac interoceptive accuracy—the Phase Selection Task (PST)—where participants are presented with tones at three delays following the R-wave (0ms, 250ms, 500ms) and asked to rank the delays according to their perception of heartbeat-tone synchrony. In a sample of 59 participants, results revealed a significant moderate association between the PAT and PST ($r(57) = .252$, $p = .027$), which strengthened when restricting analysis to participants who passed the exteroceptive control task ($r(39) = .349$, $p = .013$). Crucially, no relationship was observed between either interoceptive task and an exteroceptive control task matched in format to the PST (all ps $> .250$), providing evidence for discriminant validity. Additionally, participants classified as interoceptive on the PAT demonstrated significantly higher PST scores compared to those who failed the PAT. However, the association between tasks was smaller than anticipated, highlighting that task-specific factors strongly influence performance on different measures of cardiac interoceptive accuracy. These results support the validity of the PAT as a measure of

cardiac interoceptive accuracy while emphasizing the importance of including control tasks in interoceptive research and recognizing that different measures may capture distinct aspects of interoceptive processing rather than a unitary construct.

Introduction

As demonstrated in Chapter 6, the Phase Adjustment Task (PAT) represents a significant methodological advance in measuring cardiac interoceptive accuracy, overcoming many of the limitations of existing measures detailed in Section 1.2.1 of this thesis. Although the PAT has been successfully employed in several studies (Spooner, Bird, Clemente, et al., 2024; Spooner, Bird, Irigoras Izagirre, et al., 2024; Todd et al., 2024), it has not yet been compared to other measures of cardiac interoceptive accuracy. Such comparison is critical for establishing construct validity, given evidence that different measures of interoceptive accuracy often show only poor-to-moderate correspondence with each other (Brener et al., 1993; Hickman et al., 2020).

While Section 1.2.1 outlined the fundamental limitations of the two-alternative forced choice heartbeat detection task (2AFC HDT)—including its susceptibility to false negatives when individuals perceive synchrony at delays other than the predefined 200-250ms and 500-550ms intervals—additional complexities warrant consideration. Evidence indicates that heartbeats are typically perceived between 0-400ms following the R-wave (Brener et al., 1993; Clemens, 1984; Yates et al., 1985). This broad range creates problems even when analysis strategies are modified to accept either delay as evidence of interoceptive ability (Brener et al., 1993; Garfinkel, Manassei, et al., 2016; Herman et al., 2019). An interoceptive individual whose perceived heartbeat falls between the two presented delays may experience both tones as asynchronous, resulting in random responding and false classification as non-interoceptive.

Furthermore, recent evidence suggests a more fundamental challenge: a subgroup of individuals may perceive heartbeat-tone synchrony at a particular phase of their cardiac cycle

rather than at a specific delay following heartbeat contraction (Palmer, R., Morelli, D., Plans, D., Todd, J., Aspell, J. E., Leganes, M., Bird, G., & Murphy, J., n.d.; Palmer, R., Murphy, J., Bird, J., Piercy, T., & Adams, K.., n.d.). While heartbeat perception at a fixed phase coincides with a consistent delay when heart rate is stable, natural heart rate fluctuations mean the same cardiac phase maps onto multiple delays. For instance, 25% of a 1000ms interbeat interval corresponds to 250ms, but only 125ms with a 500ms interval. For phase-based responders, delay selections will vary with heart rate changes, potentially causing interoceptive individuals to appear non-interoceptive when only delay consistency is considered.

Alternative approaches such as the Method of Constant Stimuli (MCS) and newer scoring approaches for the heartbeat tapping task (HTT; (Körmendi et al., 2022); (Smith et al., 2020, 2021)) partially address these issues by inferring interoceptive ability from response consistency while presenting stimuli at multiple delays or allowing broader response windows. However, as discussed in Section 1.2.1, the MCS remains impractical due to administration length, presents delays only within 0-500ms (failing to cover the full cardiac cycle except at very fast heart rates), and typically uses 100ms bins despite evidence that individuals may resolve smaller intervals. The HTT, while providing wider response windows, still requires responses within predefined parameters and may not capture phase-based responders with variable heart rates.

The aim of this pre-registered study was to examine the degree of correspondence between the PAT and another task assessing cardiac interoceptive accuracy. Given these limitations and practical constraints, we designed a 3-alternative forced choice task—the Phase Selection Task (PST). Unlike existing tasks that typically present a series of tones at one delay and then ask participants to make a binary decision (synchronous or asynchronous), in the PST, on each trial, participants are presented with tones at three delays following objective heartbeat detection (0ms, 250ms, and 500ms). The order of the three delays is randomized across

presentations. Participants are told that tones follow heartbeats (as per (Wiens & Palmer, 2001)) and are asked to rank the tones according to their delay from the heartbeat. The consistency with which participants rank tones presented at each delay serves as evidence of interoceptive ability.

The PST design addresses multiple limitations of existing measures. First, it does not assume a specific delay at which participants will perceive synchrony—participants rank tones by their delay from the heartbeat rather than making binary synchronous/asynchronous judgments. Second, the use of three delays with ranking instructions overcomes the problem of participants who perceive their heartbeat between two delays. Even if such participants perceive tones at two delays as equally synchronous and alternate between ranking them first and second, they should consistently rank the third delay as least synchronous. Since each ranking position is analyzed independently and consistency at any position indicates interoceptive ability, participants perceiving their heartbeat between delays will still be correctly classified. These design features make the PST valid for participants who perceive their heartbeat at any point in the cardiac cycle, though like other delay-based measures, it may not fully capture phase-based responders with high heart rate variability.

The presentation of three tones per delay represents a balance between accuracy and practicality. Evidence indicates that participants benefit from greater exposure when comparing one to five tones, but show no additional performance benefits when comparing five to ten tones (Brener et al., 1994). This design choice ensures adequate sampling of heartbeat-tone relationships while maintaining reasonable task duration.

Given evidence that ~25% of the variance in heartbeat-tone synchronicity judgments may be accounted for by light-tone synchronicity judgments (Knapp et al., 1997), we designed a structurally identical control task for the PST. In this control task, participants rank tones

presented at the same delays following a flashing shape rather than heartbeats, allowing us to distinguish cardiac interoceptive ability from general multisensory integration capabilities. Participants completed all tasks on smartphones in counterbalanced and randomized order, supervised via videoconferencing software.

As pre-registered, we predicted that scores on the PAT would correspond positively with scores on the PST, and that this correspondence between the two cardiac interoceptive accuracy tasks would be greater than the correspondence between either interoceptive task and the control task. Although we predicted no relationship between the PAT and PST Control Task given their differing task demands and focus (interoception vs. exteroception), given the evidence regarding light-tone simultaneity judgments (Knapp et al., 1997), we predicted a relationship between the PST and PST Control Task due to their structural similarity.

Methods

Ethical approval was granted by the Medical Sciences Inter-Divisional Research Ethics Committee (IDREC) at the University of Oxford [R62582/RE004]. Participants provided informed consent and were fully debriefed on completion of the study. Prior to the collection of data, the protocol, hypotheses and analysis plan were pre-registered and can be accessed at (https://osf.io/vbequ/registrations). Analysis scripts can be accessed via this link (https://osf.io/vbequ/files/osfstorage). All deviations from the pre-registered protocol are reported in text.

Participants

Sample size was determined a priori anticipating a medium effect size (r = .50) based on reported relationships between other cardiac interoceptive accuracy multisensory integration tasks (Brener et al., 1993). Power analyses suggested that a sample size of 36 participants

would provide >80% power to detect medium effect size for the analysis examining the relationship between the interoception tasks.

Participants were recruited from Prolific.ac.uk. Inclusion was restricted to those who were over 18 years of age, spoke English as their first language, had never participated in a study employing any of the interoception tasks, and who had an iPhone meeting software requirements (iOS 14 or later; iPhone version 8 and above). 348 participants attempted the PAT Screener Task in Phase 1 (for a description of the method see Materials). As pre-registered, those who had too few valid trials on the Screener Task were not invited to the second phase of the experiment. This was initially pre-registered as 17 valid trials but following piloting we also extended inclusion to participants with at least 15 valid trials to retain a greater number of participants. This change is justifiable to retain participants, particularly as simulations suggest 15–20 trials provides a reasonable balance between task length and the reliability of consistency scores (Palmer et al., under review) with a maximum of a 0.6% change in consistency score with two fewer trials. Participants with a PAT Screener score <.80 using the original scoring method (Plans et al., 2021) were also not invited. Note that during piloting this pre-registered threshold was changed from > .42 to > .80 as examination of the data suggested that those scoring < .80 did not demonstrate sufficient task engagement. All participants invited to Phase 2 met the >.80 threshold using the original scoring method at the time of screening; however, in the final dataset some participants may appear to have screener scores below .80 because we applied the updated PAT scoring method (Palmer et al., under review) for consistency with the main PAT analysis. The screening decisions themselves were correctly made using the > .80 threshold, ensuring that only participants who demonstrated adequate task engagement during the screener were invited to continue. 181 participants passed these thresholds and were invited back to complete additional tasks under video supervision.

86 eligible participants completed the interoception tasks in Phase 2 (see Materials for a description of methods). As pre-registered, we excluded participants who had too few valid trials (<17 for the PAT; <15 for the PAT Screener; <18 for the PST or PST Control Task), and individuals whose heart rate was > 90 bpm for the Phase Selection Task (see rationale below). Scores on the PAT Screener Task were reanalysed using the updated scoring approach for the PAT (Palmer et al., under review), and only those who passed this task were included in the final sample. This resulted in a final sample size of 59 participants (Mage = 34.58 years, SDage=12.72 years, 30 females) with valid data across all four tasks.

Materials

Phase Adjustment Task (PAT)

Task Description

Participants completed the PAT as described in (Palmer et al., 2025) via a bespoke smartphone application. The procedure utilised photoplethysmography to capture cardiac activity, with participants placing a finger on the smartphone's camera and flash. Following a baseline recording period to capture a minimum of 180 heartbeats to recover heart rate parameters, participants were provided with detailed task instructions (see Palmer et al., 2025). In the PAT, participants are presented with a series of auditory tones that are triggered by their heartbeat but are out of phase with their heartbeat. Participants are asked to rotate a virtual dial to either advance or delay the tones until they subjectively perceive the auditory signals to be synchronised with their heartbeat. The task included an initial familiarisation phase, comprising two trials requiring synchronisation of two tones, and two trials where participants were asked to align tones with their heartbeat. This preparatory phase was succeeded by 20 main task trials, during which participants were asked to synchronise the tones to their heartbeat.

Since the original description of the PAT (Plans et al., 2021), we have made several refinements to the measure, and the present study makes use of this optimised version (Palmer et al., under review). Of relevance here is that in the present study tones were triggered by heartbeats and the cheat avoidance mechanism was used. However, as participants were observed during the task via video conferencing software as outlined in our updated pre-registration, no exclusions were made based on the number of times the cheat-aversion mechanism was triggered as it was possible to determine visually whether participants attempted to feel for their pulse. During the task, participants are explicitly informed to focus on one bodily location throughout. Locations from which heartbeats are perceived are collected using a body map that is presented before and after the task.

Scoring

PAT data were initially exported in JSON format. We imported these data into R Studio (v2023.12.1) and organised the data using the tidyr package. We extracted the beat-to-beat (RR) intervals from the baseline heart rate data. To increase the integrity of heart rate data to get a stable estimate of heart rate, a structured cleaning and smoothing protocol was implemented. First, raw photoplethysmographic (PPG) signals were assessed for artifacts, including implausible values (e.g., heart rates below 30 bpm or above 180 bpm) and segments corrupted by movement or incomplete readings, which were excluded. Erroneous data points, such as those recorded as -1 due to device misreads, were systematically removed. Following automated artifact rejection, a rolling median filter with a window size of 5 was applied to reduce random noise and smooth the IBI series, enhancing signal clarity. Subsequently, a second-order low-pass Butterworth filter with a cutoff frequency of 0.4 Hz was employed to suppress high-

frequency noise while preserving the physiological rhythm of the signal. These preprocessing steps meant the resultant dataset accurately reflected participants' heart rate dynamics during the task. After removing each participant's practice trials, we checked several quality filters, so that trials with 0 dial movements or ≤ 4 heart rate values were removed, owing to a lack of user engagement. As described above, we excluded participants who had registered <17 valid trials (as pre-registered). To ensure that participants had equal trial numbers, we selected each participant's first 17 trials (while discarding additional trials) to allow for the computation of consistency scores. Example code for all tasks can be found at https://github.com/Interolab/Interoception.

Due to updates to the PAT scoring procedure following pre-registration, the present study utilises the revised scoring approach described in Palmer et al., (under review). This deviates from our planned approach but is a significant improvement in scoring method. In this revised approach, consistency is defined using delay values as periodic functions of the heartbeat period, calculated as:

$$consistency\ (d,p) = \frac{1}{n} mod\left(\sum_{j=1}^{n} e^{i2\pi\frac{d_j}{p_j}}\right)$$

where d is the delay and p is the heartbeat period.

Phase-based. To calculate phase-based consistency, we expressed delays as phase angles relative to each trial's median IBI at the final dial position (p). In other words, delays are interpreted as a proportion of the trial IBI. To determine whether performance is above chance, a group-level cutoff was derived from 100,000 simulations of random delay sequences and randomly sampled heartbeat periods (0.5–1.5s). The 95th percentile of this null

distribution serves as the threshold for classifying responses as non-random (i.e., interoceptive).

Delay-based. To calculate delay-based consistency, fixed-period scoring was utilised. Here, all delays are interpreted as an absolute value using a constant, participant-specific period p, calculated as the participant's overall median of the final dial position median IBIs. Because individual variability in IBI affects the null distribution under this model (see Palmer et al., under review), participant-specific thresholds are generated from 10,000 simulated delay sequences per participant. The 95th percentile of each individual's null distribution is used as the cutoff for classifying their responses as interoceptive.

Classification. Participants are classified as interoceptive if they score above the chance threshold under either scoring method, resulting in a binary score. For those classified as interoceptive under the delay-based (fixed period) method, a continuous score is also provided, with a score of 1 indicating absolute consistency. Notably, in our sample no participants who met our inclusion criteria were classified as phase-based responders.

PAT Screener Task

Task Description

The PAT Screener Task is structurally identical to the PAT, except that participants are asked to synchronise two tones, with one of these tones (unbeknownst to participants) triggered by their heartbeat. Like the PAT, the task was implemented on a purpose-built smartphone application, and heartbeats are captured using the same method (photoplethysmography). The aim of this screener task was to familiarise participants with the task procedure. Participants completed two practice trials followed by 20 main task trials.

Scoring

Scoring of the Screener Task was identical to the PAT (described above), but only those scoring above chance thresholds were included in the final sample. For 90.3% of participants,

scores were based on the completion of 17 valid trials, as pre-registered. However, 6 participants' scores were based on the completion of 15 valid trials to help retain participants. Note that the pattern of results reported below is broadly similar when these 6 participants were excluded.

Phase Selection Task (PST)

Task Description

All participants completed the PST implemented using a purpose-built smartphone application, with heartbeats captured using photoplethysmography in the same manner as the PAT and PAT Screener Task. Before the task, a 2 min baseline recording was used to capture heartbeat dynamics. In the PST, participants hear three sets of tones (with each set randomly allocated one of ten notes separated by whole tones between E7 and G#5 without replacement on each trial), with each set containing three tones that were triggered by their heartbeats. Presentation of three tones was considered a reasonable trade-off between accuracy and task length considering evidence that participants benefit from greater exposure to heartbeats when presented with one tone to five tones, but exhibit no performance benefits when comparing five tones to ten tones (Brener et al., 1994). One of the series of tones had a 0ms delay from when the pulse was detected at the participant’s finger by the smartphone application, one had a 250ms delay, and the final set of tones had a 500ms delay. The delays used were determined by piloting, but broadly correspond to the delays typically used for the 2AFC heartbeat detection task (Hickman et al., 2020) after accounting for average pulse transit time. After each tone series was presented, there was a delay of 1000ms + half the average IBI before the next series was presented. The following heartbeat triggered the next series of tones. Each series of tones was presented alongside a distinct coloured shape continually present throughout the tones. The specific combinations of shapes and colours and the orders in which they appeared varied randomly across participants. Participants were asked to order the tones

in terms of how close they perceived each series of tones to be to the time of their heartbeats using the shapes (“Rank the tones in order following your heartbeat. The tone closest to your heartbeat should be first, and the tone furthest after your heartbeat should be last. Tap an icon to add or remove it to the list.”; for full instructions please see the Supplement). Participants could amend their answer as many times as they wished before clicking a button to submit their response which triggered the next trial. Between each trial there was a 500ms delay plus a randomly selected delay between 270 and 1350ms.

To discourage cheating, before each trial participants had to place the index finger of one hand over the flash and camera lens and put five digits of the other hand on a blue box (5.8 by 4.0cm as measured on an iPhone 8) on the screen. Once their digits had been recognised as having been placed on the screen, they were able to slide them over the screen to a comfortable position but could not lift them off the screen. This was to prevent participants from measuring their pulse with their hand. If participants lifted their fingers off the screen during the trial, the anti-cheat mechanism would be activated, and that trial would be aborted. That order of tone series would be entered back into the pool of potential upcoming trials and would be repeated. If the anti-cheat mechanism was triggered too many times (> 20 times during the task), then the task was ended. Participants were able to take breaks if needed between the trials and did not need to hold their fingers on the screen during this time. As pre-registered, no exclusions were made based on the number of times the anti-cheat mechanism was activated as participants were observed during the task via video conferencing software. In total, participants completed 2 practice trials followed by 18 main task trials.

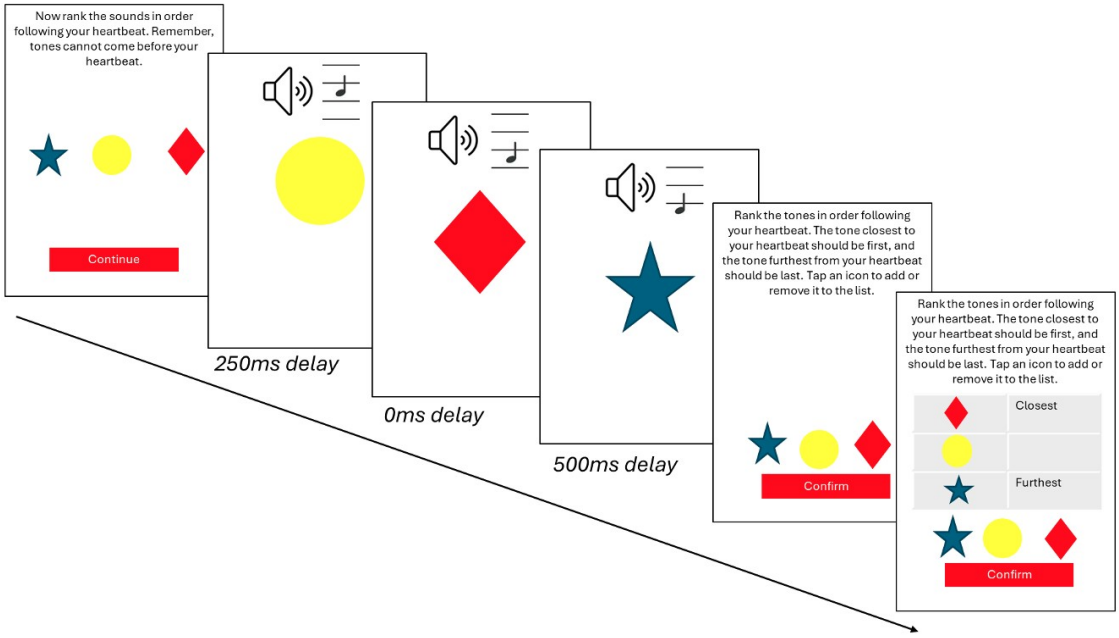


Figure 1. Example trial structure for the Phase Selection Task. After receiving instructions (see Supplement) participants are presented with a series of tones at one of three different delays (0ms, 250ms or 500ms after the detection of the pulse wave at the finger; presentation of delays is randomised across trials). Each delay is accompanied by a unique tone and each series is accompanied by a random colour-shape combination that changed on each trial. After listening to tones, participants are asked to rank the tones according to their delay from the heartbeat before confirming their response. The Phase Selection Control Task was identical to the Phase Selection Task, except that participants were required to match tones to the flash of the shape, with the same delays used for presentation.

Scoring

Heartrate data from the baseline recording was cleaned in the same manner to the PAT data (described above). PST data were initially exported in JSON format. We imported these data into R Studio (v2023.12.1) and organised the data using the tidyr package.

As pre-registered, participants with a heart rate >90 bpm were excluded to ensure that there was no overlap between tones triggered by the target heartbeat and the next heartbeat. Participants with <18 valid trials were excluded from further analyses as pre-registered. In accordance with the pre-registration, an individual was deemed to be interoceptive (i.e., to be able to perceive their heartbeat) on the Phase Selection Task if they selected the same position (least delay following heartbeat, intermediate delay, longest delay) for a certain delay on

>=11 trials (representing above chance performance according to a binomial distribution) regardless of whether this was in the first position[1] . For example, if a person placed 250ms in position 3 in >=11 trials, this participant would be deemed to be interoceptive. This scoring method resulted in a binary score for each participant reflecting consistent ('interoceptive') or inconsistent ('non-interoceptive') responses over trials, as well as a continuous score reflecting the maximum number of times delays were placed in the same position. This resulted in a score from 1–18 with higher values reflecting greater consistency.

Phase Selection Task – Control Task (PST Control)

Task Description

The PST Control task was identical to the PST except participants were asked to compare the timing of a flashing shape to the timing of tones. Before the task, a baseline recording of 2 min was taken to capture heart rate dynamics. For a consistent participant experience across the PST Control task and PST, heart rate was taken throughout the task using photoplethysmography and the anti-cheat mechanism was employed (described above). During the PST Control task, both flashes and tones were presented for 40ms. To match the variability of heart rates on average, during the PST Control Task, flashes were generated using a normal distribution with a mean of 0.8, standard deviation of 0.066 and a cut-off of 2 standard deviations.

Scoring

Scoring for the PST Control Task was identical to the PST.

Procedure

Participants were initially invited to complete the PAT Screener Task in an unsupervised fashion. Those who passed (criteria described above; see Participants) were invited to

1 Note that this was originally pre-registered as 10/18 but subsequently updated in the pre-registration documents to the correct number of 11/18 that represents above chance responding.

complete the other tasks whilst supervised. Note that this change from unsupervised to supervised testing was included in our updated pre-registration. Supervision was achieved via video conferencing software, such that participants were observed during the task. This decision was made because during piloting it was observed that ~93% of participants passed the PST Control Task when supervised during testing, but only ~27% of participants passed the PST Control task when unsupervised. This differs from the PAT Screener (and PAT) where pass rates are generally consistent regardless of supervision and likely reflects the increased difficulty of the PST demands (Palmer et al., under review; Spooner, Bird, Clemente, et al., 2024). To reduce the burden on researcher time, we supervised participants in groups of between 5–10 (on average this resulted in groups of 4–6 participants). Before completing tasks, participants were asked to watch a presentation given by a trained researcher, with slides explaining the nature of the tasks and the logistical details of performing the tasks on their smartphones. After this, they completed the tasks on their smartphones supervised by the researchers. The order of the PAT and PST were counterbalanced, with the order of the PST and PST Control Task randomised by the smartphone application. Whilst we pre-registered our intention to invite participants completing the PST back at a second time point to examine test-retest reliability, this was not completed due to resource constraints.

Analysis Plan

We pre-registered our intention to use logistic regression to examine the relationship between the PAT, PST and PST Control classifications. Our intention was to convert the odds ratios obtained from the logistic regressions into correlation coefficients and compare the effect sizes between the different tasks using R-to-Z tests to test the proposal that the cardiac interoceptive accuracy tasks should correspond better with each other than they do with the control task. However, due to the small number of interoceptive individuals observed across

both tasks (see Results), we took the decision to simplify the analysis. In place of logistic regression, we utilise chi-square tests, reporting odds ratios as well as the results of both Pearson chi-square tests and Fisher's exact tests due to small expected counts ($N<5$). This revised analysis is conceptually similar to the logistic regression but is better placed to handle a small number of observations. Then as planned and pre-registered, we converted the odds ratios into correlation coefficients and compared these using R-to-Z tests.

However, as this approach still results in a substantial loss of power in comparison to continuous scores (Cohen, 1983), to increase power we now also include the relationship between continuous scores on both tasks. For the PST and PST Control Task, this continuous score reflects the maximum number of times delays are placed in the same position. For the PAT, we utilised the consistency scores from the delay method (described above) as no individuals in this sample who met our inclusion criteria were classified as phase-based responders. Whilst we generally recommend against the use of continuous scores (Palmer et al., under review), this provides the best logical comparison with the PST, as the PST focuses on the consistency of delay selections which is analogous to the PAT delay scoring approach which also focuses on the consistency of selected delays. As continuous scores for all tasks did not follow a normal distribution, Spearman correlations were used to examine associations, which were then compared using R-to-Z tests as planned. As a complementary analysis, Mann-Whitney U tests were used to explore whether PST continuous scores were greater in those who achieved above chance performance on the PAT than those who did not, and vice versa. We also assessed the relationship between the interoceptive tasks after accounting for control task performance, using a multiple regression approach. All analyses examining PST and PAT associations were completed before and after excluding participants who failed the PST control task, as pre-registered.

Analyses were completed in SPSS, except odds ratios, which were converted to correlation coefficients using an online effect size converter (www.escal.site/) and correlation comparisons were made using an online calculator (https://www.quantpsy.org/corrtest/corrtest2.htm). As pre-registered, one-tailed p-values were used for directional predictions.

Patient and Public Involvement

No formal patient and public involvement (PPI) was undertaken for this pre-registered validation study, reported here against the GRIPP-2 Short Form checklist (Staniszewska et al., 2017). The Phase Selection Task (PST) was designed by the research team as a comparator measure for the PAT (see Methods, Phase Selection Task). The study protocol was pre-registered on OSF prior to data collection (see G2 declaration in this chapter), and no PPI consultation was undertaken at the design, recruitment, or analysis stages. As with Chapters 5 and 6, this is a recognised limitation of the measurement-development work in this thesis. Future applications of the PAT and PST in clinical populations should incorporate structured PPI consultation, particularly given the thesis's interest in remote deployment to under-researched groups (s. 9.4). See s. 8.5.5 for the thesis-level reflection.

Results

Descriptives

Descriptive Statistics for all four tasks can be found in Table 1.

Table 1. Presents mean and standard deviations of consistency scores on all tasks and the N of participants above chance cut-off.

	Mean	SD	N above chance
Phase Adjustment Task	.25	.15	8
Phase Selection Task	9.17	1.44	10
Phase Selection Task (Control)	13.03	3.52	41

Total N=59.

Continuous scores

In the full sample, a significant association was observed between continuous scores on the PAT and PST ($r(57) = .252$, $p=.027$). No association was observed between the PST and the PST Control Task ($r(57) = -.030$, $p>.250$), or the PST Control Task and the PAT ($r(57) =.015$, $p>.250$). The size of the correlation between the PAT and PST was not greater than the relationship between the PST and PST Control Task ($Z=1.54$, $p=.060$), or the relationship between the PST Control Task and the PAT ($Z=1.26$, $p=.10$).

After excluding those who failed the PST Control Task, the association between the PAT and PST increased ($r(39) =.349$, $p=.013$). No relationship was observed between the PST Control Task and PST ($r(39) =-.233$, $p=.072$), or the PST Control task and PAT ($r(39) =.003$, $p>.250$). The size of the correlation between the PAT and PST was not greater than the relationship between the PAT and PST Control Task ($Z=1.42$, $p=.078$) but was greater than the relationship between the PST Control Task and the PST ($Z=2.68$, $p=.004$).

In the full sample, those who were deemed interoceptive on the PAT had a significantly higher score on the PST ($M=10.00$, $SD=1.20$) than those who failed the PAT ($M=9.04$, $SD=1.44$; $U=1.93$, $p=.027$). There was no significant difference in the mean PAT score for individuals who passed ($M=.30$, $SD=.21$) or failed ($M=.24$, $SD=.13$) the PST ($U=.566$, $p>.250$), though scores were numerically higher. When focusing only on those who passed the PST Control Task, those who were deemed interoceptive on the PAT had a significantly higher score on the PST ($M=10.33$, $SD=1.21$) than those who failed the PAT ($M=9.09$, $SD=1.54$; $U=2.06$, $p=.021$). Likewise, those who were deemed interoceptive on the PST had a significantly higher score on the PAT ($M=.35$, $SD=.20$) than those who failed the PST ($M=.22$, $SD=.13$; $U=1.77$, $p=.038$).

Finally, when predicting PST scores from PAT continuous scores, controlling for PST Control Scores using a regression model, only PAT scores were a significant predictor

(β=.247, t=1.91, p=.031). PST Control Scores were not a significant predictor of PST scores (β=-.080, t=-.620, p>.250). After removal of one outlying data point (>3.5 SD away from the mean of the standardised residuals), the same pattern of results was observed whereby PAT scores (β=.344, t=2.72, p=.009), but not PST Control scores (β=-.049, t=-.389, p>.250), predicted PST scores. For the comparable analysis predicting PAT continuous scores from PST continuous scores whilst controlling for Screener Task scores using a regression model, only PST scores were a significant predictor (β=.249, t=1.90, p=.032). Screener Task scores were not a significant predictor of PAT scores (β=.045, t=.342, p>.250).

Classification scores

Table 2 presents a contingency table comparing PAT and PST classification scores (Table 2a), PAT and PST Control Task classification scores (Table 2b) and PST and PST Control Task classification scores (Table 2c). No association was observed between the PAT and PST (X^2=2.77, p=.096; Fisher's exact = .125, OR=3.77). No association was observed between the PST and PST Control Task (X^2=.627, p>.250; Fisher's exact = .708, OR=1.94). No association was observed between the PAT and PST Control Task (X^2=.132, p>.250; Fishers exact = 1.00, OR=1.37). After excluding those who failed the PST Control Task, an association was observed between the PAT and PST (X^2=4.16, p=.041; Fisher's exact = .077, OR=6.00). In the full sample, after converting odds ratios to r values, the size of the association between the PAT and PST was not significantly greater than the association between the PST and PST Control Task (Z=0.96, p=.17) or the association between the PAT and PST Control Task (Z=1.58, p=.057).

Exploratory analyses

As an exploratory analysis, we descriptively examined whether task engagement and time spent on trials varied across the groups who passed both cardiac interoceptive accuracy tasks or failed one task and passed another (Table 2). We did this by exploring the mean task

length, mean time on trials, mean engagement (dial turns) and the estimated number of heartbeats sampled using the following equation: Baseline HR (BPM) × (Mean time per trial/ 60 seconds). As can be seen in Table 3, the highest engagement and time spent on task was exhibited in those who passed both tasks, followed by those who passed the PAT and failed the PST, and finally by those who passed the PST and failed the PAT.

To examine order effects, we also compared continuous scores on the PAT, PST and PST Control Task as a function of task order (Table 3). As can be seen, PST scores were lowest when the PST was completed last and highest when the PST was completed first. For the PST Control task, scores were greater when the PST Control task was completed before the PST. Finally, although less variability was observed, scores were slightly higher when the PAT was completed first. In those who passed the PST but failed the PAT, 5/7 of these participants completed the PAT last.

Table 2. Classification scores across the three tasks

A)	PST PASS	PST FAIL	B)	PST PASS	PST FAIL	C)	PST-C PASS	PST-C FAIL
PAT PASS	3	5	PST-C PASS	8	33	PAT PASS	6	2
PAT FAIL	7	44	PST-C FAIL	2	16	PAT FAIL	35	16

PST-C = PST Control Task. Note that the Screener Task is not included as (as pre-registered) inclusion was contingent on passing this task.

Table 3. Engagement metrics across the tasks

	PST PASS; PAT PASS	PAT PASS; PST FAIL	PAT FAIL; PST PASS
Mean Engagement (full sample)	28.90 (15.95)	23.09 (13.49)	21.20 (5.40)
Time on Task (full sample)	436.29 (243.69)	293.54 (160.32)	271.01 (57.94)
Mean time per trial (full sample)	25.66 (14.33)	17.27 (9.43)	15.94 (3.41)
Estimated heartbeats (full sample)	30.95 (17.92)	24.76 (13.26)	22.19 (5.87)
Mean Engagement (PST Control Passers)	28.90 (15.95)	25.96 (18.25)	23.06 (5.28)
Time on Task (PST Control Passers)	436.29 (243.69)	323.58 (218.79)	282.42 (65.02)
Mean time per trial (PST Control Passers)	25.66 (14.33)	19.03 (12.87)	16.61 (3.82)
Estimated heartbeats (PST Control Passers)	30.95 (17.92)	26.91 (18.29)	24.27 (5.69)

Note: Estimated heartbeats calculated from mean heart rate estimated during the baseline recording and applied to the time spent on trials using the following equation: estimated heartbeats = HR (BPM) × (Mean time per trial/60 seconds).

Table 3. Task performance as a function of order

	Order 123	Order 132	Order 231	Order 321
PST	8.93 (1.16)	9.17 (1.19)	9.04 (1.46)	9.89 (2.03)
PST Control	13.27 (3.77)	12.92 (3.97)	13.13 (3.65)	12.56 (2.03)
PAT	.28 (.18)	.24 (.10)	.23 (.15)	.23 (.15)

Order. 1 = PAT, 2 = PST Control, 3 = PST

Discussion

This pre-registered study set out to explore the association between the PAT and a novel test of cardiac interoceptive accuracy, the Phase Selection Task (PST), designed to overcome the significant problems with other tests of cardiac accuracy, and a closely matched exteroceptive control task (PST Control). In line with our predictions, we observed a significant relationship between the PAT and the PST. Partially confirming our predictions, the relationship between the PAT and the PST was greater than the relationship between both of the cardiac

interoceptive accuracy tasks and the PST Control task, although the statistical significance of the latter finding varied slightly depending on the analysis strategy employed and was only observed when restricting inclusion to those who had performed above chance on the PST Control Task. The implications of these results are discussed below.

The overall pattern of results supports the convergent and discriminant validity of the PAT as a measure of cardiac interoceptive accuracy. Indeed, a significant relationship was observed only between the measures of cardiac interoceptive accuracy, with no relationship observed between the exteroceptive control task and cardiac interoceptive accuracy tasks. This pattern is striking considering the structural similarities between the PST and PST Control Task, which may be expected to produce a relationship between scores given their similar task demands. However, contrary to previously reported relationships between cardiac interoceptive accuracy tasks and matched exteroceptive control tasks (Knapp et al., 1997), we observed no relationship. These results suggest a separation between the ability to match tones to heartbeats and the ability to match tones to a flashing shape, emphasising the separation between interoception and exteroception. Although such results may be taken to suggest that it is not necessary to control for exteroception when assessing cardiac interoceptive accuracy, it is notable that the strongest relationship between the PAT and PST was observed after restricting inclusion to those who passed both exteroceptive control tasks. Such a result underscores the importance of including control measures to ensure that poor performance on tasks of cardiac interoceptive accuracy can be attributed to poor heartbeat perception and not non-interoceptive features of the task, or features of the participant such as poor motivation or attention (Palmer et al., under review).

It should be noted that the relationship between the PAT and PST (~r=.35) was lower than previously reported relationships between interoceptive tasks that share very similar formats (~r=.50; Brener et al., 1993), although it was notably greater than the relationship between the

two most commonly used tests of cardiac interoceptive accuracy (r=.21; Hickman et al., 2020). Differing task demands may explain this discrepancy. In the PST participants are only exposed to three tones triggered by their heartbeats at each delay. In contrast, the PAT allows participants to sample tones for as long as they wish before submitting their response. It is notable that on the PAT participants typically sample a far greater number of heartbeats (in our sample the mean across trials was 23.47 heartbeats per trial; range 11.53–49.67 heartbeats per trial). It is conceivable that participants may vary in terms of how much they benefit from increased exposure to heartbeats, though the broad range of sampled heartbeats certainly suggests individuals vary in terms of how much they choose to engage with the task when it is self-paced. Although it was not possible to examine this statistically, those who passed the PAT and failed the PST showed greater engagement with the PAT than those who failed the PST and passed the PAT, consistent with the latter argument. Overall, these results reflect patterns observed in the wider literature that differing task demands strongly influence results obtained from different interoceptive accuracy tasks, with those that are closely matched showing stronger relationships than those that diverge (Brener et al., 1993; Hickman et al., 2020). Although controlling for performance in a matched exteroceptive control task may account for some of this variation, arguably no control task will fully capture all variance associated with differing task demands on tasks of cardiac interoceptive accuracy. A second possible explanation is that individuals who pass the PAT, but fail the PST, prefer delays that do not neatly map to those presented in the PST. Although participants were instructed before completing the PST that tones could only follow heartbeats, and to rank the degree to which tones were delayed rather than pick the synchronous tones, it is possible that these instructions are difficult to follow and contribute towards discrepancies across measures. A final possibility is the presence of order effects. Although too few participants to examine statistically, there was some indication that performance was better when tasks were

completed at earlier stages of the study. Indeed, in those who passed the PST but failed the PAT, 5/7 had completed the PAT last, suggesting that some cross task discrepancy may be explained by fatigue effects. Overall, whilst there are many possible reasons for divergence, it is notable that in those who had passed the control tasks, scores were significantly numerically higher in the PAT and PST for those who had above chance performance on the PST and PAT, respectively, confirming correspondence across the two tasks.

In this study we chose to relate the PAT to a newly developed task – the PST – which requires judgements about the timing of heartbeat-tone synchrony but overcomes the significant problems with the 2AFC HDT, is far quicker to administer than the MCS, and can identify heartbeat perception across the entirety of the cardiac cycle. Although the development of the PST enabled comparison with the PAT, we cannot recommend this measure for the assessment of individual differences in cardiac interoceptive accuracy for two reasons. First, like all commonly used tasks of cardiac interoceptive accuracy, the PST focuses on the consistency of participants' selected delays and thus would not adequately capture interoceptive ability in phase-based responders with high levels of IBI variability over trials. Although no participants who met inclusion criteria in this sample exhibited a phase-based preference for the PAT, given evidence that some participants may synchronise tones at a certain phase of their cardiac cycle, rather than at a specific delay (see Palmer et al. under review), we recommend the use of the PAT and the updated analysis approach for the assessment of individual differences in cardiac interoceptive accuracy (Palmer et al., under review). Second, to ensure that tones were only presented within one cardiac cycle, we excluded participants with heart rates >90bpm. Although such an approach should likely be implemented for all cardiac interoceptive accuracy tasks that present tones at specific delays following the R-wave, excluding participants based on heart rate is not ideal, particularly for researchers that wish to examine cardiac interoceptive accuracy under heart rate perturbation

conditions. These factors highlight the benefits of the PAT for the assessment of individual differences in cardiac interoceptive accuracy - unlike the PST, the PAT captures all patterns of non-random responding, can be used with all heart rates, and makes no assumptions regarding the timing at which individuals perceive heartbeat-tone synchrony. The development of a matched exteroceptive control task that captures multisensory integration processes in a similar manner to the PST Control Task (e.g., a tone-flash variant of the PAT Screener Task) would further enhance the utility of this measure. Indeed, although the PAT Screener Task captures a participant's ability to match two stimuli (thus likely controlling for attentional and motivational factors), testing participants' ability to match stimuli across two different perceptual systems would enable us to conclude that failure on the PAT is due to poor heartbeat perception and not difficulties in multisensory integration. Although above chance performance on the PAT cannot be achieved via any mechanism other than heartbeat perception, this refinement would allow separation of those who do not score above chance into those who can perform the task except for the requirement to perceive their heartbeat (i.e., 'non-interoceptive' individuals), and those who cannot perform the task for another reason.

Before discussing general limitations of this study, a foundational limitation must be stated at the outset. The Phase Selection Task (PST) was developed specifically for this study and, like the PAT, has no independent psychometric track record; convergence between the PAT and PST therefore does not establish validity against a known criterion, but only that two novel tasks pattern together. The rationale for this choice was not arbitrary. The most widely used measures of cardiac interoceptive accuracy — the heartbeat counting task (HCT) and the two-alternative forced-choice heartbeat detection task (2AFC HDT) — are subject to the validity concerns reviewed in s. 1.2.1, including the documented contribution of belief and estimation strategies to HCT performance (Desmedt et al., 2018; Murphy, Millgate, et al., 2018; Ring et

al., 2015) and the susceptibility of the 2AFC HDT to false negatives when individuals' preferred synchrony delays fall between the two presented intervals (Brener & Ring, 2016). These measures are therefore poor criteria against which to validate a new task. The underlying difficulty is not, however, specific to this thesis: there is no uncontested gold-standard measure of cardiac interoceptive accuracy in the field, and the absence of one constrains the validation of any new task. The methodological priority for future work is therefore convergence between the PAT and measures with independent evidence of construct validity — for example, refined heartbeat counting protocols using non-tracking instructions (Desmedt, Corneille, et al., 2020), neural correlates such as the heartbeat-evoked potential, and convergence with measures of cardiac vagal tone.

Despite the importance of these data for assessing the validity of the PAT, limitations should be acknowledged. First, for practical and scientific reasons at times we deviated from our pre-registered protocol. Although all these deviations were for justifiable scientific reasons reported in the text (e.g., updates to the analysis procedure for the PAT and PAT Screener Task; issues with power requiring changes to our analytic approach; changes to retain a greater number of participants in our sample), due to these deviations these results require replication. Second, although participants completed the three main tasks in the same session in order to control for state effects (Wittkamp et al. 2018), future replications may benefit from having participants complete tasks on different days to reduce participant fatigue and the possibility of order effects. Third, although this study was powered to detect associations between the tasks, it was likely underpowered to detect differences in the size of associations across tasks. For the planned comparison of two dependent correlations with one common index, power depends on the relationship between the two non-shared variables. Given the absence of the predicted relationship between the PST and PST Control Task (predicted based on results from Knapp et al., 1997), our final sample size only provided ~80% power to detect

a large difference in the size of the correlation across tasks (a difference of $\approx\Delta r$=.40; α=.05; one-tailed). It is therefore premature to conclude that relationships between the PST and PAT are not greater than the relationships between these tasks and the PST Control task, as this study was only powered to detect a large difference. Future research in larger, well-powered samples informed by effect sizes reported in the present study are necessary for adequately testing this hypothesis. Finally, although we assessed the relationship between the PAT and another measure of cardiac interoceptive accuracy, future work would benefit from examining the relationship between the PAT and another closely matched cardiac interoceptive accuracy task that can account for both delay- and phase-based response patterns, is sensitive to heartbeat perception across the entire cardiac cycle, and does not make assumptions regarding the timing at which individuals perceive heartbeat-tone synchrony.

In conclusion, these results support the convergent and discriminate validity of the PAT as a measure of cardiac interoceptive accuracy. Despite convergence with tasks of cardiac interoceptive accuracy, and a lack of a relationship with the exteroceptive control task, observed relationships were smaller than expected, emphasising that task demands strongly influence results on tasks of cardiac interoceptive accuracy. Although such observations underscore the need to control for non-interoceptive factors that may impact performance through the inclusion of control tasks, further work is required to replicate these findings and explore relationships between the PAT and other measures of cardiac interoceptive accuracy.

8. Discussion

8.1. Summary of Empirical Chapters

8.1.1. Brief Summary

This thesis comprises six empirical studies that document the progression from the design, implementation and piloting of biometric-driven therapeutic interventions to the development and validation of novel interoception measures. The early studies (Chapters 2-4) investigated smartphone applications that utilized biometric feedback to support mental health, consistently demonstrating therapeutic efficacy while revealing challenges in user engagement with physiological data such as heart rate variability. The later studies (Chapters 5-7) focused specifically on developing reliable, remotely deployable measures of interoception to address the fundamental question of why individuals with anxiety and depression appear to sometimes have diminished awareness of their bodily signals, and how to effectively measure interoception via telemetrics in therapeutic interventions.

8.1.2. Chapter 2

8.1.2.1. Summary and interpretation

Chapter 2 examined the efficacy of the BioBase digital therapeutic platform in a randomized controlled trial with university students experiencing elevated anxiety and stress. The study employed a waitlist-controlled design with 805 participants screened and 262 randomized to intervention (n=130) or control (n=132) groups. The intervention consisted of a smartphone application paired with a wearable device (BioBeam) providing real-time biometric feedback including heart rate variability and movement data, combined with psychoeducational content based on cognitive behavioral therapy and mindfulness strategies delivered over four weeks.

Results demonstrated significant reductions in anxiety as measured by the primary outcome, the STAI-S-6 (State-Trait Anxiety Inventory). Linear mixed models revealed significant group × time interactions at 4 weeks (estimate = 7.09, 95% CI 2.94 to 11.24, p = .001) and at

6-week follow-up (estimate = 4.56, 95% CI 0.49 to 8.63, p = .03). Planned comparisons showed that participants in the intervention group had significantly lower anxiety scores at 4 weeks (d = 0.67) and 6 weeks (d = 0.81) compared to baseline, while the control group showed no significant changes. Mental well-being measured by WEMWBS showed significant improvements in the intervention group at 4 weeks (d = 0.65) and 6 weeks (d = 1.16). Depression scores measured by the PHQ-9 also showed significant reductions in the intervention group at 4 weeks (d = 0.99) with no significant changes in the control group.

These findings should be interpreted as preliminary indicative evidence that a multi-component digital programme combining biometric feedback, brief CBT- and mindfulness-based modules, and EMA-based mood declarations is associated with reductions in self-reported state anxiety and improvements in wellbeing in a student population with elevated baseline stress and anxiety. The post-intervention effect on the primary outcome was robust across the intention-to-treat sensitivity analyses presented in Appendix E (ITT-equivalent d ≈ 0.47–0.65 at four weeks). The 6-week follow-up effect was less robust (ITT-equivalent d ≈ 0.17–0.35) and should be treated as a question requiring further investigation rather than a confirmed claim of sustained efficacy. The non-significant 2-week interim assessment suggests that longer exposure may be required for clinically meaningful change to emerge, although this remains an inference from a single trial. Because the intervention combined multiple therapeutic components, the present design cannot isolate the specific contribution of biometric feedback to the observed effects, and the chapter therefore positions BioBase as a candidate adjunct to existing support pathways rather than as a standalone efficacious treatment.

8.1.2.2. Strengths and limitations

This study represents one of the largest randomized controlled trials of a biometric feedback intervention for anxiety to date, with robust power calculations and appropriate statistical analyses using linear mixed models. The use of validated, standardized measures (STAI-S-6, WEMWBS, PHQ-9) enhances comparability with other research. The inclusion of both post-intervention and follow-up assessments provides evidence for sustained therapeutic effects, addressing a common limitation in digital intervention research. The focus on university students represents an important target population given high rates of anxiety and depression in this demographic. The study adopted an ecologically valid approach by allowing participants to engage with the intervention naturally without strict usage requirements.

The study was conducted exclusively with university students, limiting generalizability to other populations including clinical samples and working adults. The intervention combined multiple therapeutic components (biometric feedback, psychoeducation, mindfulness, CBT techniques), making it impossible to determine the specific contribution of biometric feedback to the observed effects. The study lacked an active control condition, raising questions about whether the effects were specific to the BioBase intervention or attributable to general factors such as self-monitoring or device novelty. Participants with psychiatric diagnoses were excluded, limiting conclusions about clinical populations who might benefit most from such interventions. The study was unblinded due to logistical constraints, potentially introducing expectancy effects.

8.1.3. Chapter 3

8.1.3.1. Summary and interpretation

Chapter 3 reported a pilot feasibility and acceptability study of the BioBase intervention in working adults experiencing workplace stress. This single-arm study with 55 participants over

four weeks focused on real-world implementation of biometric feedback for stress reduction, targeting a different population from the university students in Chapter 2. The intervention utilized the same BioBase platform with workplace-specific psychoeducational content based on the job demands-resources model (A. B. Bakker & Demerouti, 2007), incorporating elements of mindfulness, cognitive behavioral therapy, and behavioral activation theory.

Results demonstrated significant reductions in anxiety measured by the State-Trait Anxiety Inventory (baseline M = 45.52, SD = 13.2; 4-week M = 39.82, SD = 11.2; t_{54} = -3.51, p < .001, Cohen's d = 0.96) and improvements in mental well-being measured by WEMWBS (baseline M = 48.12, SD = 6.4; 4-week M = 50.4, SD = 6.9; t_{53} = 2.41, p = .019, Cohen's d = 0.66). Consistent with Chapter 2, participants with higher baseline stress showed significantly greater improvements in both anxiety reduction (R^2 = 0.180) and well-being enhancement (R^2 = 0.101). Acceptability data revealed that 70% of participants would recommend the program to a friend, and the majority found the content relevant to workplace stressors.

Read alongside the Chapter 2 trial, the pattern observed in the workplace sample is consistent with the hypothesis that multi-component digital programmes of this type may produce reductions in self-reported anxiety and improvements in wellbeing across non-clinical populations under elevated stress. The single-arm design without a control group means the present study cannot rule out regression to the mean, time effects, or expectancy from being enrolled in a workplace mental-health programme; it should therefore be read as feasibility and acceptability evidence rather than as confirmation of intervention effectiveness. As with Chapter 2, the multi-component nature of the BioBase intervention precludes any claim about which therapeutic element drives the observed change (s. 8.3.4 and Whitton et al., 2015). The findings are best positioned as supporting continued investigation of digitally delivered, workplace-adapted multi-component programmes as candidate components of broader

occupational mental-health support pathways, rather than as standalone efficacious treatments.

8.1.3.2. Strengths and limitations

This study provides important evidence for the feasibility and acceptability of biometric feedback interventions in workplace settings, addressing a critical gap in occupational health research. The use of workplace-specific content based on established models (job demands-resources model, HSE work stressor dimensions) enhances ecological validity and relevance for the target population. The study collected comprehensive acceptability and usability data, providing insights into user experience that are often missing from efficacy trials. The finding that participants with higher baseline stress showed greater improvements is consistent with Chapter 2, suggesting a robust pattern across populations. The 4-week intervention period is pragmatic for workplace implementation while still allowing sufficient time for therapeutic engagement.

The single-arm design without a control group significantly limits the ability to attribute observed changes to the intervention rather than natural fluctuation, regression to the mean, or general factors such as seasonal effects. The sample size (n=55) was relatively small for establishing reliable estimates of intervention effects. Participants were paid to participate, which may have influenced engagement patterns and acceptability ratings in ways that would not generalize to naturalistic use. The study did not assess longer-term sustainability of effects beyond the 4-week intervention period. The workplace context may have introduced unmeasured confounds related to work demands, organizational culture, or concurrent workplace wellness initiatives. Selection bias may have affected results as participants volunteered for a study involving workplace stress, potentially over-representing individuals already motivated to address their stress levels.

8.1.4. Chapter 4

8.1.4.1. Summary and interpretation

Chapter 4 described the development and pilot testing of a digital acceptance and commitment therapy (ACT) intervention for generalized anxiety disorder using the person-based approach (Hemmings et al., 2021). This study represented a methodological evolution from the previous studies by incorporating user-centered design principles and targeting a specific clinical population. The intervention was delivered through the BioBase platform and included modules targeting the core ACT processes of awareness, openness, and active engagement with values-based behavior over a 2-week period.

The person-based approach involved creating detailed user personas based on qualitative research with the target population, followed by the development of guiding principles that informed intervention design. Results from 10 participants diagnosed with GAD showed significant reductions in anxiety measured by GAD-7 (baseline M = 14.7, SD = 4.3; post-intervention M = 11.7, SD = 4.6; $t_9 = 2.66$, p = .03, Cohen's d = 0.69) and depression symptoms measured by PHQ-9 (baseline M = 11.5, SD = 5.9; post-intervention M = 8.7, SD = 5.2; $t_9 = 3.25$, p = .008, Cohen's d = 0.84). The intervention achieved high usability scores (System Usability Scale M = 83.5, equivalent to "A" rating) and good engagement with 77% of participants completing all modules.

These findings provide preliminary indicative evidence that an ACT-based digital intervention developed using the person-based approach is acceptable and feasible for adults with diagnosed GAD, and that further evaluation in a controlled design is warranted. The within-subject effect sizes observed on the GAD-7 (d = 0.69) and PHQ-9 (d = 0.84) are of a similar order of magnitude to those reported for face-to-face ACT, but with n = 10 and no control group these descriptive statistics should be read as hypothesis-generating rather than as evidence of efficacy; regression to the mean, attention effects, and the natural fluctuation of

GAD symptoms cannot be ruled out. The proposal that present-moment bodily awareness in ACT may enhance interoceptive attention is a post-hoc inference (interoception was not measured here) and is best treated as a candidate mechanism motivating the measurement work in Chapters 5–7 rather than as a finding of the present study.

8.1.4.2. Strengths and limitations

This study employed rigorous person-based approach methodology, ensuring that the intervention was developed with substantial input from the target user population. The focus on participants with diagnosed GAD addresses an important clinical population with high comorbidity and substantial functional impairment. The use of ACT as the therapeutic framework is well-suited to digital delivery given its focus on acceptance and mindfulness rather than complex cognitive restructuring. The 2-week intervention period is practical for clinical implementation while still achieving significant therapeutic effects. The study provided detailed acceptability and usability data using validated measures (System Usability Scale), offering insights into user experience. The inclusion of participants already engaged in treatment reflects real-world clinical contexts where digital interventions might serve as adjunctive therapies.

The extremely small sample size (n=10) severely limits statistical power and generalizability of findings. The study lacked a control group, making it impossible to determine whether observed improvements were due to the intervention, natural fluctuation of symptoms, or placebo effects. The sample was entirely female, limiting generalizability to male populations with GAD. Participants were recruited through self-referral, potentially introducing selection bias toward individuals already motivated to engage with digital interventions. The study did not include longer-term follow-up assessment, leaving questions about sustainability of therapeutic gains. No measures of the proposed ACT mechanisms (psychological flexibility, cognitive defusion, values clarification) were included, limiting

understanding of therapeutic processes. The integration of ACT content within the existing BioBase platform may have created methodological confounds between therapeutic approach and delivery platform. Specifically, participants may have attributed therapeutic benefits to the biometric feedback features rather than ACT principles, or vice versa. Future studies should employ dismantling designs to isolate the specific contributions of ACT content versus biometric monitoring to therapeutic outcomes.

8.1.5. Chapter 5

8.1.5.1. Summary and interpretation

Chapter 5 presented the development and validation of the CARdiac Elevation Detection (CARED) task, representing the first novel measure of interoception designed specifically for remote, ecological assessment (Ponzo et al., 2021). This study marked a significant methodological shift from therapeutic intervention research toward fundamental measurement development in interoception science. The CARED task utilized continuous heart rate monitoring via wearable devices combined with ecological momentary assessment delivered through smartphone notifications over four weeks.

The task required participants to judge whether their heart rate was elevated above normal levels when prompted by notifications, providing a naturalistic measure of cardiac interoceptive accuracy in real-world contexts. Data from 30 participants revealed that approximately one-third of the sample was classified as interoceptive using Mann-Whitney U test analyses, with participants presenting overall good insight into their interoceptive abilities as evidenced by AUROC values significantly different from chance ($t_{29} = 4.96$, $p < 0.001$, $d = 5.28$; $M = 0.58$, $SD = 0.09$). Interoceptive awareness, measured by confidence ratings, also demonstrated above-chance performance.

The CARED task established a proof-of-concept for smartphone-based interoception measurement, demonstrating that cardiac interoceptive accuracy could be assessed in naturalistic environments rather than controlled laboratory settings. This represents an important advance for the field, given that most interoception research has relied on laboratory-based measures with questionable ecological validity. The finding that approximately one-third of participants were classified as interoceptive is consistent with previous research using traditional heartbeat counting tasks, providing convergent validity for the approach.

8.1.5.2. Strengths and limitations

This study pioneered the use of ecological momentary assessment for interoception measurement, addressing a major limitation of laboratory-based approaches. The 4-week assessment period provided extensive data on interoceptive accuracy across multiple contexts and time points, enhancing reliability and ecological validity. The use of wearable devices for continuous heart rate monitoring enabled objective verification of physiological states, eliminating potential confounds related to heart rate knowledge. The study provided clear operational definitions and scoring procedures that could be replicated by other researchers. The integration of confidence ratings alongside accuracy measures addressed the important distinction between interoceptive accuracy and interoceptive awareness (Garfinkel et al., 2015). The real-world implementation demonstrates the feasibility of remote interoception assessment for clinical and research applications.

The small sample size (n=30) with significant attrition (22 participants withdrew) raises concerns about statistical power and selection bias. The study lacked validation against established interoception measures, limiting conclusions about convergent validity. No control tasks were included to assess discriminant validity or rule out general factors such as attention or compliance. The reliance on subjective judgments of "elevated" heart rate without

individualized baselines may have introduced measurement error. The study population was limited to healthy young adults, restricting generalizability to clinical populations where interoception abnormalities are hypothesized to be most relevant. Technical issues with wearable devices and smartphone notifications may have affected data quality and participant engagement. The study did not examine test-retest reliability, leaving questions about the stability of CARED task performance over time. A critical limitation is that participants may use contextual knowledge rather than interoceptive awareness to judge heart rate elevation. For instance, participants engaged in exercise when notified would logically infer elevated heart rate without necessarily perceiving bodily signals. While we excluded responses during reported high-intensity activities from analysis, participants may still use cognitive strategies based on recent activities rather than genuine interoceptive perception. Future iterations should include control questions during rest periods or implement signal detection approaches to distinguish interoceptive sensitivity from response bias.

8.1.6. Chapter 6

8.1.6.1. Summary and interpretation

Chapter 6 reported the development and validation of the Phase Adjustment Task (PAT), a smartphone-based measure designed to overcome the methodological limitations of existing interoception tasks (Plans et al., 2021). The PAT required participants to synchronize audio tones with their perceived heartbeat by adjusting the phase relationship between tones and heartbeats using a simple dial interface. This approach eliminated potential confounds related to heart rate knowledge, counting strategies, and physiological variables that plague traditional measures such as the heartbeat counting task.

Study 1 with 124 participants demonstrated that the PAT could reliably distinguish between interoceptive and non-interoceptive individuals using Bayesian classification approaches, with between 34% and 48% of participants classified as interoceptive depending on the statistical

threshold adopted (Bayes Factors >3, >10, or >30). Critically, PAT performance was unaffected by physiological characteristics including heart rate variability (HRV), heart rate, body mass index, or the bodily location where heartbeats were perceived. Study 2 with 90 participants revealed moderate test-retest reliability over one week (ICC = 0.508), consistent with theoretical models proposing both trait and state components of interoceptive ability.

Unlike many previous studies using traditional heartbeat counting tasks, no significant associations were found between PAT performance and measures of anxiety, depression, or stress in this non-clinical sample. This finding challenges previous theoretical assumptions about the relationship between interoception and psychopathology, suggesting that earlier reported associations may have been artifacts of measurement limitations rather than genuine theoretical relationships. The PAT's freedom from physiological confounds represents a significant methodological advance, enabling more valid assessment of interoceptive abilities independent of individual differences in signal strength or physiological parameters.

8.1.6.2. Strengths and limitations

The PAT represents a significant methodological innovation by eliminating many confounds that affect traditional interoception measures. The use of smartphone cameras for heart rate detection makes the task highly accessible and scalable for large-scale research and clinical applications. The Bayesian classification approach provides probabilistic evidence for interoceptive abilities rather than relying on arbitrary cut-off scores. The task's independence from physiological confounds (HRV, heart rate, BMI) is a major advantage over existing measures. The two-study design with different samples enhances confidence in the findings and provides important reliability data. The task duration (15-20 minutes) is practical for clinical and research contexts. The inclusion of confidence ratings and bodily location reports provides additional insights into interoceptive processes beyond simple accuracy measurement.

The absence of associations with anxiety, depression, and stress measures raises questions about the clinical relevance of PAT performance, particularly given theoretical predictions that interoception should relate to psychopathology. The moderate test-retest reliability (ICC = 0.508) suggests substantial state variability, complicating interpretation of individual differences. The study did not include validation against other established interoception measures, limiting conclusions about convergent validity. The reliance on smartphone camera-based heart rate detection may introduce technical artifacts that could affect measurement precision. The sample consisted primarily of young, healthy adults recruited through online platforms, limiting generalizability to clinical populations and older adults. The task assumes that participants can perceive their heartbeat location accurately, which may not be valid for all individuals. The binary classification approach may oversimplify the continuous nature of interoceptive abilities (see The Challenge of Binary Classification in Interoception Research section for further discussion).

8.1.7. Chapter 7

8.1.7.1. Summary and interpretation

Chapter 7 presented a pre-registered validation study examining the relationship between the Phase Adjustment Task (PAT) and a newly developed Phase Selection Task (PST) to establish construct validity for smartphone-based interoception measures. The PST required 59 participants to hear tones presented at three different delays following heartbeat detection (0ms, 250ms, and 500ms) and then rank these tones according to their perceived synchrony with heartbeats. This study was designed to test convergent validity between different approaches to measuring cardiac interoceptive accuracy while also examining discriminant validity through an exteroceptive control task.

Results revealed a significant moderate association between the PAT and PST in the full sample ($r(57) = .252$, $p = .027$), supporting convergent validity between different smartphone-

based approaches to measuring cardiac interoceptive accuracy. When restricting analysis to participants who passed the control task, this association strengthened considerably ($r(39) = .349$, $p = .013$). Crucially, no relationship was observed between either interoceptive task and an exteroceptive control task that was matched in format to the PST (PST and PST Control: $r(57) = -.030$, $p > .250$; PAT and PST Control: $r(57) = .015$, $p > .250$), providing important evidence for discriminant validity. Additionally, participants classified as interoceptive on the PAT had significantly higher PST scores ($M = 10.00$, $SD = 1.20$) compared to those who failed the PAT ($M = 9.04$, $SD = 1.44$; $U = 1.93$, $p = .027$).

However, the association between the PAT and PST was smaller than anticipated, highlighting that task-specific factors strongly influence performance on different measures of cardiac interoceptive accuracy. A further constraint on the interpretation of these findings is that the PST is itself a novel task, developed for this validation work; the present results therefore establish convergence between two new smartphone-based measures rather than validation against an established criterion (see s. 7.7), and validation against an independent gold-standard measure of cardiac interoceptive accuracy remains a priority for future research. The discriminant-validity finding nonetheless reinforces the importance of including matched control conditions in interoceptive research to ensure that observed effects reflect genuine interoceptive abilities rather than general cognitive or perceptual factors, and supports the broader proposition that different interoception measures may capture distinct aspects of interoceptive processing rather than a unitary construct.

8.1.7.2. Strengths and limitations

The pre-registered design enhances methodological rigor and reduces the risk of data-driven analyses. The inclusion of both convergent and discriminant validity assessments provides comprehensive construct validation. The development of a novel comparison task (PST)

specifically designed to test similar interoceptive processes represents an important methodological contribution. The PST uses photoplethysmography to detect actual heartbeats and trigger tones at precise delays, rather than relying on participants to self-pace or estimate timing. This eliminates confounds present in tasks where participants might count seconds or use other timing strategies instead of attending to heartbeat sensations. The study employed appropriate statistical analyses for non-normal data distributions and included planned comparisons following pre-registered hypotheses. The sample size (n=59) was adequate for correlation analyses and provided reasonable power for detecting medium-sized associations.

The smaller than expected correlation between PAT and PST raises questions about whether these tasks measure the same underlying construct or tap into different components of cardiac interoception, such as temporal discrimination (PAT's focus on phase adjustment) versus relative timing judgments (PST's ranking of delays). The study did not include validation against established interoception measures (e.g., heartbeat counting task), limiting broader construct validity assessment. The sample consisted exclusively of healthy young adults, restricting generalizability to clinical populations where interoceptive abnormalities are most theoretically relevant. The study did not examine test-retest reliability of the PST, leaving questions about the stability of this measure. No measures of other interoceptive dimensions (awareness, sensibility) were included, limiting understanding of the tasks' relationships to broader interoceptive constructs. The study did not investigate potential moderators of the task relationships, missing opportunities to understand individual differences in interoceptive measurement (see further discussion of this point in the Distinguishing Interoceptive Abilities from Task-Specific Demands section)

8.2. Overall Implications

The research presented in this thesis has important implications for our understanding of digital mental health interventions that attempt to deploy biometric measures to understand

the role and potential therapeutic application of capturing and analysing autonomous nervous system activity, the measurement of interoception, and the relationship between biometric feedback and therapeutic outcomes. The findings contribute to three key areas of scientific knowledge: the development of ecological interoception measures, the effectiveness of smartphone-based biometric interventions, and the theoretical understanding of why individuals with anxiety and depression might struggle to connect with bodily signals.

8.2.1. Digital Therapeutics and Biometric Feedback

The early studies in this thesis (Chapters 2-4) demonstrate that smartphone-based digital therapeutics incorporating biometric feedback can produce meaningful therapeutic effects for anxiety and depression. The consistent finding across university student and workplace adult populations that the BioBase intervention significantly reduced anxiety (with sustained effects at follow-up) aligns with recent meta-analytic evidence indicating that digital mental health interventions yield small to moderate effect sizes (g = 0.30) across diverse populations (Diel et al., 2024) . However, the informal feedback from users throughout these studies — that participants could not fully reconcile biometric data with their emotional states — suggested that individual differences in interoceptive capacity may moderate response to purely data-driven approaches to mental health intervention. This remains a hypothesis the measurement work of Chapters 5–7 was designed to enable testing of, rather than a confirmed empirical claim; the evidence base for interoceptive atypicality in anxiety and depression, and particularly for cardiac interoceptive accuracy in anxiety, is mixed (Adams et al., 2022).

This finding has important implications for the design of future digital therapeutics. While biometric feedback clearly has therapeutic potential, the effectiveness of such interventions may be fundamentally limited by individuals' interoceptive abilities. The progression from Chapters 2-4 to Chapters 5-7 illustrates that understanding and measuring interoceptive capacities may be essential prerequisites for developing more effective biometric-based

interventions. This suggests that personalized digital therapeutics should assess interoceptive profiles before deploying biometric feedback, potentially offering different intervention approaches for individuals with varying interoceptive abilities.

8.2.2. Measurement Innovation in Interoception Science

The development of the CARED task (Chapter 5), PAT (Chapter 6), and Phase Selection Task (Chapter 7) represent significant methodological contributions to interoception science. These smartphone-based measures address critical limitations identified in traditional laboratory-based assessments, offering several advantages: ecological validity through real-world telemetric deployment, accessibility across diverse populations, and freedom from physiological confounds that affect traditional measures (Desmedt et al., 2023).

Importantly, recent validation research has confirmed that remote administration of smartphone-based interoception tasks yields comparable results to laboratory-based testing, supporting the validity of these approaches for large-scale research and clinical applications (Spooner, Bird, Clemente, et al., 2024). The finding that approximately one-third of participants across the CARED and PAT studies were classified as interoceptive is consistent with previous laboratory-based research, providing convergent validity for these novel approaches.

However, the smaller than expected correlations between different interoception measures (Chapter 7) highlight that task-specific factors strongly influence performance, suggesting that interoception may be more multifaceted than previously conceptualized (see section The Challenge of Binary Classification in Interoception Research). This has important implications for future research design and suggests that multiple measures may be necessary to comprehensively assess interoceptive abilities.

8.2.3. Theoretical Understanding of Interoceptive Difficulties in Mental Health

The findings of this thesis sit within a research literature in which the empirical evidence linking interoception to anxiety and depression is mixed. The meta-analytic null between anxiety and cardiac interoceptive accuracy (Adams et al., 2022, r = .02) constrains the strength of any claim that anxiety is characterised by aberrant cardiac interoception, at least as measured by the heartbeat counting and heartbeat detection tasks. The progression from the therapeutic studies of Chapters 2–4 to the measurement studies of Chapters 5–7 reflects an attempt to make this empirical question more tractable by developing ecologically deployable tasks that are not subject to the limitations of those classical measures — limitations that may in part account for the heterogeneity of findings in the literature. The thesis therefore positions interoception as a candidate mechanism warranting further investigation, with the measurement work providing tools by which the question can be examined more rigorously in future research. It does not position interoception as an established characteristic of anxiety or depression, nor as a confirmed moderator of biometric-feedback intervention response, although both of those remain plausible hypotheses worth pursuing.

8.3. General Limitations

Several limitations affect the interpretation and generalizability of findings across this thesis, relating to methodological approaches, sample characteristics, and measurement considerations.

8.3.1. Sample Limitations

The studies in this thesis were conducted primarily with non-clinical populations of university students and working adults, limiting the generalizability of findings to clinical populations where interoceptive abnormalities are hypothesized to be most relevant. The absence of participants with diagnosed anxiety or depressive disorders in the interoception measurement studies (Chapters 5-7) is particularly limiting, given that these populations are theoretically

expected to show the greatest interoceptive difficulties. Future research must validate these measures in clinical samples to establish their utility for understanding psychopathology.

Moreover, the sample sizes for the measurement development studies were relatively small (n = 30 for CARED, n = 124 for PAT, n = 59 for validation), raising questions about the stability and generalizability of the classification thresholds and psychometric properties. The significant attrition observed in the CARED study (22 participants withdrew from the original sample) suggests potential selection bias toward individuals with higher technological engagement or interoceptive motivation, as well as socio-economic factors in both quality and availability of latest model smartphones capable of running complex, resource-demanding biometric applications.

8.3.2. Methodological Considerations

The therapeutic intervention studies (Chapters 2-4) combined multiple intervention components (biometric feedback, psychoeducation, mindfulness content), making it impossible to isolate the specific contribution of biometric feedback to therapeutic outcomes. This limitation is common in complex digital interventions but prevents definitive conclusions about the mechanisms underlying therapeutic effects (Werntz et al., 2023).

The reliance on self-report measures for mental health outcomes introduces potential bias, particularly given evidence that individuals with anxiety and depression may have altered interoceptive awareness that could affect their ability to accurately report symptoms. The absence of objective physiological or behavioral measures of therapeutic change limits the strength of conclusions about intervention effectiveness.

8.3.3. Measurement Validity Questions

Despite the innovation in interoception measurement in this work, fundamental questions remain about the construct validity of these tasks. The observation that different interoception

measures show only modest correlations (Chapter 7) raises questions about whether they assess the same underlying construct or distinct aspects of interoceptive processing. This issue reflects broader challenges in the field, where traditional measures have faced similar validity concerns (Schulz et al., 2021).

The use of smartphone cameras for heart rate detection, while increasing accessibility, may introduce technical artifacts that could affect measurement precision. Although validation studies have supported the accuracy of photoplethysmographic measurement via smartphones (De Ridder et al., 2018; Gruwez et al., 2024; Morelli et al., 2018), individual differences in device usage, lighting conditions, and movement artifacts may still influence data quality in ways that are difficult to control in remote settings.

8.3.4. Data accessibility in industry-academic partnerships

Chapters 2 and 3 of this thesis report trials conducted as industry-academic collaborations under the commercial sponsorship of BioBeats Ltd (subsequently acquired by Huma Therapeutics Ltd in 2021). The acquisition, and the subsequent decommissioning of the company's research infrastructure, mean that the underlying trial datasets are no longer accessible to me. This has constrained two analyses requested at viva: a sensitivity intention-to-treat re-analysis of the Chapter 2 trial (s. 2.7 and Appendix E) and a formal engagers-versus-non-engagers comparison in the Chapter 3 workplace study (s. 3.7). Both points have been addressed by drawing on published descriptive data from the underlying papers and on the relevant methodological and empirical literature, but neither could be re-run on the raw data. This is a generalisable lesson for the field: industry-academic partnerships in digital mental health should establish data-access provisions for the academic partner at study inception, and analysis-ready de-identified datasets should be deposited alongside pre-registrations (e.g., on the Open Science Framework) at the point of publication, so that post-publication re-analysis remains feasible independently of any single commercial entity. The

measurement-development work in Chapters 5–7 was conducted with this independence in mind: the PAT and PST are open, smartphone-deployable instruments that can be used, replicated, and improved upon by the field without dependence on proprietary infrastructure.

Pre-registration and Transparency

A further transparency limitation cuts across the empirical chapters of this thesis. Of the six empirical studies presented, only the Chapter 2 randomised controlled trial (https://osf.io/2zd45/) and the Chapter 7 PAT/PST validation study (https://osf.io/vbequ/registrations) were pre-registered prior to data collection; the feasibility, development and earlier measurement studies in Chapters 3, 4, 5 and 6 were not. This unevenness reflects the period over which the work was conducted (2018–2022), during which pre-registration moved from a peripheral aspiration to an expected standard in psychological and mental-health research (Nosek et al., 2018). For the non-pre-registered studies, the research questions, outcome measures and analytic approaches reported here correspond closely to those documented in the published versions of the underlying papers (Kawadler et al., 2020; Hemmings et al., 2021; Ponzo et al., 2021; Palmer et al., 2025), providing a degree of post-hoc transparency; nevertheless, prospective pre-registration of all primary analyses should be the default standard for future work in this programme, both because it disciplines the inference (separating confirmatory from exploratory analyses) and because it strengthens reader trust in the reported effects.

8.4. Outstanding Questions

Several critical questions emerge from this research that warrant future investigation and have important implications for both theoretical understanding and clinical application.

8.4.1. Mechanisms of Biometric Feedback Effectiveness

While the therapeutic studies demonstrated that biometric feedback interventions are effective for reducing anxiety and improving well-being, the specific mechanisms underlying these effects remain unclear. Do these interventions work by improving interoceptive accuracy, by providing external validation of internal states, or through entirely different mechanisms such as increased self-monitoring or mindfulness? The informal user feedback suggesting difficulty reconciling biometric data with emotional states indicates that the relationship between objective physiological data and subjective experience is more complex than initially hypothesized.

Future research should directly test whether interoceptive training enhances the effectiveness of biometric feedback interventions, and whether individuals with different interoceptive profiles respond differently to various types of digital therapeutics. This could inform the development of personalized intervention approaches based on interoceptive phenotyping.

8.4.2. The Challenge of Binary Classification in Interoception Research

A fundamental methodological challenge that emerged across the measurement studies in this thesis (Chapters 5-7) concerns the use of binary classification systems to categorize individuals as either "interoceptive" or "non-interoceptive." While such categorical approaches offer practical advantages for clinical decision-making and statistical analyses, they may oversimplify what is fundamentally a continuous distribution of interoceptive abilities.

The historical preference for binary classification in interoception research stems partly from clinical diagnostic traditions where categorical decisions are necessary for treatment allocation. Early interoceptive tasks, including the heartbeat counting task and heartbeat detection task, adopted binary scoring systems that classified participants based on whether they exceeded predetermined accuracy thresholds. This approach has persisted despite growing evidence that interoceptive abilities exist along multiple continuous dimensions rather than as discrete categories (Garfinkel et al., 2015).

The data presented in this thesis illustrate this continuum clearly. In Chapter 6, between 34% and 48% of participants were classified as interoceptive depending on the statistical threshold adopted, highlighting the arbitrary nature of such cutoffs. Similarly, the moderate correlations observed between the PAT and PST in Chapter 7 ($r = .252$ in the full sample, $r = .349$ when restricted to those passing control tasks) suggest dimensional variation rather than distinct categorical groups. These findings align with recent critiques of dichotomization in interoception research, which argue that binary classification reduces statistical power and

creates artificial boundaries that do not reflect underlying biological or psychological reality (Desmedt et al., 2023).

The implications of this methodological choice extend beyond statistical considerations. Binary classification may obscure important individual differences that could predict therapeutic responsiveness. For instance, individuals with intermediate interoceptive abilities —those near classification thresholds—might show different patterns of response to biometric feedback interventions than those at the extremes of the distribution. The continuous nature of interoceptive abilities is further supported by neuroimaging evidence showing graded activation patterns in interoceptive brain networks rather than discrete on/off states (Critchley & Harrison, 2013). For a more expansive argument of issues that should be taken into consideration regarding this point, see the Future Directions chapter.

8.4.3. Distinguishing Interoceptive Abilities from Task-Specific Demands

A fundamental challenge in interoception research is determining whether performance differences across tasks reflect true variation in interoceptive abilities or differences in task demands such as sustained attention, working memory, motivation, or response strategies. The moderate correlation between PAT and PST (r = .252-.349) could indicate either that they measure partially overlapping interoceptive abilities or that they share some but not all cognitive demands.

For instance, the PAT requires sustained attention across multiple trials and the ability to hold a mental representation of the heartbeat-tone relationship while adjusting the dial, whereas the PST requires comparing and ranking multiple stimuli in working memory. Individual differences in these cognitive abilities, rather than interoception per se, might drive performance differences.

Future research should employ cognitive control tasks matched to the specific demands of interoceptive tasks, use statistical approaches like structural equation modeling to partial out shared variance due to general cognitive abilities, and develop paradigms that minimize extraneous cognitive demands while preserving interoceptive assessment validity. Only by systematically addressing these confounds can we determine whether observed individual differences reflect genuine variation in the ability to perceive bodily signals.

8.4.4. Clinical Utility of Novel Interoception Measures

The novel interoception measures developed in this thesis require validation in clinical populations to establish their utility for understanding psychopathology. Specific questions include: Do individuals with anxiety disorders, depression, or other mental health conditions show distinct patterns of performance on the CARED and PAT tasks? Can these measures predict treatment response or serve as biomarkers for mental health vulnerability?

Additionally, the optimal approach for combining multiple interoception measures remains unclear. Given that different tasks appear to capture distinct aspects of interoceptive processing, future research should examine how patterns of performance across multiple measures relate to clinical presentations and therapeutic outcomes.

8.4.5. Longitudinal Development and Training Effects

All studies in this thesis employed cross-sectional designs, leaving important questions about the stability and trainability of interoceptive abilities unanswered. Can interoceptive accuracy be improved through targeted training, and would such improvements translate to enhanced effectiveness of biometric feedback interventions? The ACT study emphasized present-moment bodily awareness but did not directly measure changes in interoceptive abilities, representing a missed opportunity to examine training effects.

Longitudinal research examining how interoceptive abilities change over time, particularly in relation to mental health symptoms and treatment outcomes, would provide crucial insights for developing interoception-based therapeutic approaches.

8.5. Strengths of the Work Reported in this Thesis

This thesis makes several novel contributions to digital mental health intervention research and interoception science, advancing both methodological innovation and theoretical understanding.

8.5.1. Methodological Innovation in Interoception Measurement

The development of reliable smartphone-based interoception measures represents a significant methodological advance, addressing critical limitations of traditional laboratory-based approaches. The CARED task's use of ecological momentary assessment over four weeks provides unprecedented insights into interoceptive abilities across real-world contexts and physiological states. Unlike traditional measures that assess interoception only at rest, the

CARED captures individual differences across varying levels of cardiac arousal, enhancing ecological validity.

The PAT's elimination of physiological confounds that plague traditional measures represents another important advance. By demonstrating that PAT performance is independent of heart rate variability, body mass index, and other physiological characteristics, this research provides a cleaner measure of interoceptive processing that is not confounded by signal strength differences between individuals.

8.5.2. Bridging Applied and Basic Science

This thesis successfully bridges applied digital health intervention research with fundamental interoception science. The progression from therapeutic intervention studies to measurement development illustrates how applied research conducted in industrial settings and its resulting findings can drive basic science innovation. The observation that users struggled to interpret biometric feedback provided the scientific motivation for developing better interoception measures, demonstrating how user experience can inform theoretical development.

This bidirectional relationship between applied and basic research strengthens both domains: the therapeutic studies provide real-world evidence for digital intervention effectiveness, while the measurement studies provide tools that could enhance future intervention designs.

8.5.3. Comprehensive Validation Approach

The thesis employs rigorous validation approaches across multiple studies. The PAT validation included both convergent validity testing (correlation with another cardiac interoception measure) and discriminant validity assessment (lack of correlation with exteroceptive control tasks). The CARED task development included qualitative checks to ensure participants understood task instructions and were not using external cues.

The use of pre-registered study protocols (Chapter 7) and publication of validation studies in peer-reviewed journals enhances the reproducibility and credibility of the findings. The decision to make all data publicly available through platforms such as the Oxford University Research Archive further supports open science practices.

8.5.4. Person-Based Intervention Development

The ACT study's use of the person-based approach represents a methodological strength that is often overlooked in digital intervention research. By conducting extensive qualitative research with the target population before intervention development, the study ensured that the therapeutic content was relevant and acceptable to individuals with generalized anxiety disorder. This approach may explain the high usability ratings and engagement levels observed in this study compared to many digital mental health interventions.

The strengths of the person-based approach as deployed in Chapter 4 can be stated more specifically. First, the PBA provided a structured, evidence-grounded development pathway: the qualitative work with the target population (adults with diagnosed GAD) generated personas and guiding principles that translated user concerns into concrete intervention design choices, rather than relying on the development team's intuitions. Second, the approach was rapid enough to be compatible with the timelines of digital intervention development — a substantive consideration when working in an industry-academic context with constrained delivery windows. Third, by foregrounding the target population's psychosocial context, the approach increased the likelihood that the resulting intervention would be relevant rather than only usable.

The limitations of the person-based approach as applied across this thesis warrant equally explicit acknowledgement. First, the PBA was applied only in Chapter 4. The empirical work of Chapters 2 and 3 — and the development of the measurement tools in Chapters 5–7 — proceeded without comparable structured user-input methodology, creating internal inconsistency in the thesis's approach to participatory design. This is the asymmetry that the panel's comment about why a person-centred approach appears in Chapter 4 and not before draws attention to (see also s. 1.3.5 on the MRC framework and the chronological mismatch between the empirical sequence and the developmental progression). Second, the PBA typically draws on relatively small qualitative samples that may not capture the full diversity of the target population, particularly across demographic, cultural, and clinical dimensions. Third, the PBA, as it is generally practised, operates at the level of usability and intervention relevance rather than as a participatory framework that gives service users substantive decision-making authority over the design of the intervention. The approach is closer to evidence-based design refined by user input than to transformative co-design. For an intervention space — biometric digital therapeutics — where the design choices being made (what data is collected, how it is fed back to users, what is labelled as 'abnormal') have

significant implications for how users experience their own bodily signals, this distinction matters. Future iterations of this work would benefit from a more participatory framework, discussed in s. 9.4 below.

8.5.5. Patient and Public Involvement Across the Empirical Programme

A candid reflection on patient and public involvement (PPI) across the empirical chapters of this thesis, conducted against the GRIPP-2 reporting standards (Staniszewska et al., 2017), reveals an asymmetric pattern. Five of the six empirical studies (Chapters 2, 3, 5, 6 and 7) were conducted without formal PPI consultation: study design, intervention configuration, outcome selection, and measurement-task design were determined by the research team without structured involvement of people with lived experience of the conditions targeted (subclinical anxiety, workplace stress, generalised anxiety disorder) or of the populations whose interoceptive processing was being measured. Chapter 4 (Hemmings et al., 2021) is the partial exception, having used the Person-Based Approach (Yardley et al., 2015) to ground intervention design in qualitative work with adults with diagnosed GAD. As discussed in s. 8.5.4, the PBA is not equivalent to PPI in the GRIPP-2 sense: it generates personas and guiding principles from user qualitative work, but does not give users substantive decision-making authority over intervention design.

This asymmetry should be acknowledged honestly rather than minimised. The therapeutic studies of Chapters 2 and 3 would have benefited from structured PPI consultation with university students and working adults respectively at the design stage — particularly around the framing of biometric feedback, the timing of notifications, the data-collection footprint, and the trade-offs between data minimisation and personalisation flagged in s. 9.4. The measurement-development studies of Chapters 5–7 would similarly have benefited from PPI involvement with people with lived experience of the conditions in which atypical interoceptive processing is hypothesised, particularly given the thesis's argument that these measures could inform personalised digital therapeutics for those populations (s. 8.2.3, s. 8.4.4).

It is worth noting where informal feedback did enter the design process. Across the BioBase deployments of Chapters 2–4, user feedback emerged through customer-support interactions and post-study interviews rather than through structured PPI sessions. The observation that participants struggled to meaningfully integrate biometric data with their subjective experiences — which became central to the thesis's argument that interoceptive processing moderates response to biometric feedback (s. 1.3) — emerged in part from such informal feedback. The fact that this feedback was unstructured and unsystematic, however, means it cannot stand in for formal PPI in the way GRIPP-2 requires.

Future work in this space — particularly the experience-based co-design approach proposed in s. 9.4 — should treat PPI as a structural commitment from study inception rather than as a methodological add-on, and should report PPI activities transparently against the GRIPP-2 short-form checklist. The thesis-level lesson is that the absence of formal PPI was not a single oversight but a programmatic gap that ran through five of six empirical studies, and that closing this gap requires the upfront resourcing and time commitment that the participatory frameworks discussed in s. 9.4 explicitly demand.

A note on the learning trajectory of this research programme. The thesis as a whole

reflects a research programme that matured methodologically across its empirical chapters,

and the maturation is visible in the order in which intervention-development principles enter the work. The clearest example is the appearance of the Person-Based Approach (Yardley et al., 2015) in Chapter 4 and not earlier. The PBA framing entered the programme when Nicola Hemmings joined the team to lead the digital-ACT-for-GAD study, bringing the structured intervention-development methodology that her clinical-psychology training had given her. With hindsight, PBA-style principles — qualitative formative work with the target population, iterative cycles of refinement, explicit logic-modelling of intervention components — should have informed the design of the original BioBase platform in Chapter 2 and its workplace adaptation in Chapter 3. Those earlier studies were instead shaped by a more product-led trajectory typical of the commercial mHealth context in which the work was conducted, in which a generic platform was developed and then tested in successive populations. The shift in methodological standards is also visible in the second half of the thesis: the measurement-development work in Chapters 5–7 reflects a steadier alignment with the conventions of academic psychophysics, including pre-registration (Chapter 7), explicit psychometric framing of interoceptive accuracy and awareness as separable constructs (Chapters 5–7), and the conservative validation logic that became increasingly central as the programme moved from biometric-driven therapeutics to interoception measurement. The trajectory was not a linear plan executed from the outset; it was the path the programme took as colleagues with different methodological commitments joined it at different phases. The chapter-introduction notes threaded through Chapters 2–7 make the specific contributions of each phase visible.

9. Future Directions

The research presented in this thesis opens multiple avenues for advancing our understanding of interoception and its application in digital mental health interventions. This section outlines key priorities for future research, organized around methodological innovations, technological advances, clinical translation, and theoretical developments.

9.1. Moving Beyond Binary Classification: Toward Dimensional Approaches

Future research should consider adopting dimensional approaches that preserve the full range of individual variation in interoceptive abilities. The lack of concordance between different interoceptive measures and the absence of correlations across interoceptive domains suggests that interoception may not represent a unitary construct, as highlighted by recent critiques (Schoeller et al., 2025), where the authors argue interoception may not be a coherent unitary construct, cataloguing evidence that shows performance across different interoceptive tasks (cardiac, respiratory, gastrointestinal) rarely correlates, suggesting that interoception might be better conceptualized as multiple independent abilities rather than a single general capacity.

Dimensional approaches could include using standardized continuous scores (z-scores or percentiles), implementing Bayesian methods that provide probability estimates of interoceptive ability, or employ machine learning approaches capable of handling continuous outcomes. Such approaches align with broader movements in psychiatry toward dimensional models of psychopathology, such as the Research Domain Criteria (RDoC) framework, which emphasizes continuous distributions of function from normal to abnormal (Cuthbert & Insel, 2013).

For clinical translation, dimensional approaches could enable more precise therapeutic matching. Rather than simply categorizing individuals as able or unable to perceive their heartbeats, clinicians could develop "interoceptive profiles" that capture strengths and weaknesses across multiple interoceptive domains and tasks. Digital therapeutics could then be tailored to these profiles, providing more intensive interoceptive training for those with greater impairments while offering advanced biometric feedback to those with stronger baseline abilities.

The challenge moving forward is to balance the practical needs of clinical decision-making with the scientific imperative to accurately represent the continuous nature of interoceptive

abilities. While binary classifications may remain necessary for some clinical and research purposes, acknowledging and measuring the full spectrum of interoceptive abilities will be essential for advancing our understanding of the relationship between bodily awareness and mental health.

9.2. Ecological Validity and Real-World Assessment

The smartphone-based measures developed in this thesis (CARED task, PAT, PST) represent important steps toward ecologically valid assessment, but significant opportunities remain for enhancing real-world measurement of interoception. Future research should prioritize:

Continuous Monitoring Approaches: Development of passive assessment methods that capture interoceptive processing during daily activities without requiring active task engagement. This could involve analyzing patterns of smartphone interaction, wearable sensor data, or ambient computing signals that correlate with interoceptive awareness.

Context-Aware Assessment: Creating adaptive testing protocols that adjust to environmental contexts, stress levels, and activity patterns. The CARED task's ecological momentary assessment approach could be expanded to include contextual factors that may moderate interoceptive accuracy, such as social settings, emotional states, or cognitive load.

Cross-Domain Integration: While cardiac interoception has dominated the field, future work should develop integrated assessment batteries that capture interoceptive abilities across respiratory, gastrointestinal, and other bodily domains. Studies examining correlations in performance across tasks have found that individuals who excel in one interoceptive domain do not necessarily perform well in others (Schoeller et al., 2025), suggesting domain-specific assessment approaches are needed.

9.3. Integration with Digital Therapeutics: From Assessment to Intervention

The convergence of interoception measurement and digital therapeutics presents unprecedented opportunities for personalized mental health interventions. Digital therapeutics trials have surged since 2005, with mental health fields particularly well-positioned to benefit from technological advances in natural language processing, sentiment analysis, and virtual reality technology (Masanneck & Stern, 2024). Future research priorities should include:

Adaptive Intervention Protocols: Developing algorithms that adjust biometric feedback presentation based on real-time assessment of interoceptive abilities. This could involve starting with simplified physiological displays for individuals with lower interoceptive accuracy and progressively increasing complexity as abilities improve.

Interoceptive Training Modules: Creating standardized training protocols that systematically enhance different aspects of interoceptive processing. These could target specific deficits identified through comprehensive assessment, such as attention to bodily signals, accuracy of perception, or metacognitive awareness of interoceptive abilities.

Biomarker-Guided Therapy: Digital biomarkers obtained from real-time and continuous log data are being used to improve the objectiveness of endpoints in digital therapeutic trials (Huh et al., 2022). Future work should establish interoceptive biomarkers that predict treatment response and guide intervention selection.

9.4. Experience-Based Co-Design for Future Iterations of Interoception-Informed Digital Therapeutics

Building on the strengths and limitations of the person-based approach reflected on in s. 8.5.4, future iterations of digital interoception-informed therapeutics would benefit from a more participatory development methodology than the PBA. Experience-based co-design (EBCD; Bate & Robert, 2006) is one such framework. Originally developed for redesigning hospital services, EBCD differs from the person-based approach in three respects that are relevant to the intervention design questions raised by this thesis. First, it is iterative rather than primarily front-loaded: rather than generating personas and guiding principles before intervention build, EBCD structures repeated engagement events across the development cycle. Second, it is dual-stakeholder, bringing service users and the staff who deliver services together to identify what Bate & Robert (2006) call 'touchpoints' — specific moments in the user journey where

experience is shaped by service design — and to co-design improvements at those touchpoints. Third, EBCD is structurally transformative rather than confirmatory: where the PBA tends to refine an intervention concept already developed by the research team, EBCD treats users as participants in the design process itself, with substantive influence over what the intervention does and how it does it (Donetto et al., 2015).

Three design questions raised by the empirical chapters of this thesis would benefit particularly from an EBCD approach in future work. First, *what biometric feedback should look like for users with different interoceptive abilities*. The thesis's broader argument — that interoceptive processing may moderate response to biometric feedback (s. 1.3 and s. 8.2.3) — implies that "one-size-fits-all" feedback designs are likely to fail substantial subsets of users. EBCD with users who span the range of interoceptive abilities, including individuals with disrupted interoceptive processing, would help identify what feedback formats, intensities, and framings are acceptable and useful across that range. Second, *what acceptable trade-offs are between data minimisation and personalisation*. Passive biometric monitoring (HRV, sleep, activity) collects substantial personal data; users' tolerance for this depends on the perceived benefit-to-privacy ratio and on the trust they have in how data is held. EBCD provides a structured forum for those trade-offs to be negotiated rather than assumed. Third, *how to design feedback that does not pathologise normal variation*. The same biometric pattern can be presented as 'low HRV — stress signal' or as 'within your normal range'; the framing matters for how users interpret their bodily signals. EBCD's dual-stakeholder structure, involving both users and clinicians or coaches, would help identify framings that support rather than undermine users' interoceptive self-knowledge.

A candid acknowledgement should close this section. At the time the empirical work of this thesis was conducted, EBCD was not standard practice in mHealth development, and the commercial timelines of the BioBeats partnership (see s. 3.7.3) would have constrained the more extended, iterative engagement that EBCD requires. These constraints are real, but they should not be normalised as inevitable. Future industry-academic work in this space, conducted under the kinds of safeguards proposed in s. 3.7.3, should explicitly plan for and resource participatory engagement of the depth that EBCD implies. The measurement work of Chapters 5–7 — particularly the development of remote, smartphone-deployable interoception measures — provides a substrate for such engagement: the ability to characterise users' interoceptive profiles before, during, and after EBCD events would allow the participatory design itself to be informed by the kind of phenotyping the thesis argues for.

9.5. Mechanistic Understanding Through Multimodal Research

To advance beyond correlational findings, future research must employ multimodal approaches that elucidate the mechanisms linking interoception to mental health outcomes:

Neuroimaging Studies: Combining interoception tasks with neuroimaging to identify neural markers of interoceptive processing differences. This could reveal whether individuals with similar behavioral performance show different patterns of neural activation, suggesting distinct compensatory mechanisms or processing strategies.

Computational Modeling: Recent work emphasizes the need for mechanistic understanding of how signals from the body are transduced, transmitted, compressed, and integrated by the brain to produce affective states (M. J. Feldman et al., 2024). Developing computational models that formalize the relationship between physiological signals, interoceptive processing, and emotional experience could provide testable predictions about intervention mechanisms.

Longitudinal Developmental Studies: Tracking interoceptive abilities across the lifespan to understand how these abilities develop, stabilize, and potentially decline. This could identify critical periods for intervention and reveal whether early interoceptive differences predict later mental health outcomes.

9.6. Clinical Translation and Implementation Science

Moving from research findings to clinical practice requires dedicated implementation research:

Healthcare Integration: Developing implementation strategies for incorporating interoceptive assessment into routine clinical care. This includes creating brief screening tools suitable for primary care settings, establishing referral pathways for comprehensive assessment, and training healthcare providers in interpreting interoceptive profiles.

Health Economics: Conducting cost-effectiveness analyses of interoception-informed interventions compared to standard care. This evidence will be crucial for securing reimbursement and ensuring equitable access to these novel therapeutic approaches.

9.7. Technological Innovations and Emerging Opportunities

Recent technological advances offer new possibilities for interoception research and intervention:

Artificial Intelligence Applications: AI-powered Digital Therapeutics hold potential for enhancing stress prevention by promoting the scalability of P5 Medicine (predictive, preventive, personalized, participatory, and precision medicine). Machine learning algorithms could identify subtle patterns in interoceptive data that predict clinical outcomes or treatment response for particular molecular medicine interventions, such as GLP1 drugs, which modulate interoceptive signaling through vagal afferent pathways and central stress-responsive neural circuits (L. A. Jones & Brierley, 2025). These medications act as interoceptive inputs to central eating control networks while simultaneously influencing hypothalamic-pituitary-adrenal axis responses to interoceptive stress. AI algorithms could potentially predict individual variations in treatment response by analyzing baseline interoceptive sensitivity patterns, vagal tone measurements, and stress-induced changes in gut-brain signaling that are mediated by the GLP-1 receptor's role in stress (Holt & Trapp, 2016). This could enable personalized dosing strategies and identify patients most likely to benefit from GLP-1 receptor agonist therapy, while also predicting and preventing stress-related side effects through real-time monitoring of interoceptive biomarkers and adaptive intervention protocols that optimize both metabolic outcomes and stress resilience. Additionally, while research strongly supports that GLP-1 drugs work through interoceptive pathways, more clinical studies are needed to directly measure their effects generally on interoceptive abilities in humans.

Virtual and Augmented Reality: Immersive technologies could create controlled environments for interoceptive training, allowing individuals to practice bodily awareness in simulated scenarios that gradually increase in complexity and ecological validity.

Digital Twins for Personalized Medicine: Digital twins enable healthcare professionals to gain a comprehensive understanding of patients' health, personalize treatment plans, and make

data-driven decisions (Vallée, 2024). Future work could develop "interoceptive digital twins" that model individual patterns of bodily signal processing to predict intervention response.

9.8. Addressing Fundamental Questions About Interoception

Several fundamental questions remain unanswered and warrant dedicated research attention:

Is interoception best conceptualized as a general ability that manifests across different bodily domains, or as a collection of independent domain-specific abilities? The moderate correlations observed between PAT and PST in this thesis, combined with evidence that performance across different interoceptive tasks rarely correlates, suggest a complex picture requiring theoretical refinement.

State vs. Trait Dynamics: The moderate test-retest reliability observed for the PAT (ICC = 0.508) suggests both stable individual differences and state-dependent variation. Future research should disentangle trait-like interoceptive abilities from state-dependent fluctuations, potentially identifying factors that temporarily enhance or impair interoceptive processing.

Cultural and Demographic Factors: Most interoception research has focused on Western, educated populations. Future work should examine how cultural beliefs about the body, linguistic differences in describing bodily sensations, and demographic factors influence interoceptive processing and its relationship to mental health.

9.9. Methodological Innovations for the Phase Adjustment Task and Related Measures

Specific improvements for the measures developed in this thesis should be prioritized:

Enhanced Scoring Algorithms: Developing machine learning approaches that can identify subtle patterns in PAT performance beyond simple consistency metrics. This could reveal different strategies individuals use to perceive their heartbeats, such as attending to specific phases of the cardiac cycle or using respiratory cues.

Multi-Signal Integration: Expanding beyond cardiac signals to create integrated tasks that assess how individuals perceive relationships between different physiological systems (e.g., respiratory-cardiac coupling, gastric-cardiac interactions).

Gamification and Engagement: Digital therapeutics often attempt to deliver engaging experiences through gamification or social support to enhance adherence (Lutz et al., 2022). Future versions of interoception tasks could incorporate game-like elements to maintain engagement during assessment and training, particularly important for clinical populations with attention or motivation difficulties.

9.10. Building an Evidence Base for Clinical Utility

To establish interoception assessment as clinically valuable, future research must demonstrate:

Predictive Validity: Large-scale studies showing that interoceptive measures predict important clinical outcomes, such as treatment response, relapse risk, or functional impairment. This requires moving beyond cross-sectional associations to prospective prediction studies.

Incremental Validity: Demonstrating that interoceptive assessment adds predictive value beyond existing clinical measures. This is particularly important for justifying the additional time and resources required for comprehensive interoceptive assessment.

Clinical Decision Support: Developing evidence-based guidelines for how interoceptive assessment results should inform treatment decisions. This could include algorithms that recommend specific interventions based on interoceptive profiles, similar to pharmacogenetic testing for medication selection.

9.11. Ethical Considerations and Responsible Innovation

As interoception-based digital therapeutics advance, several ethical considerations require attention:

Data Privacy and Security: With the robust amount of personal data digital therapeutic products collect, there is evident concern regarding security and privacy. Future systems must implement robust security measures while maintaining the accessibility and usability that make smartphone-based assessment attractive.

Equity and Access: Ensuring that interoception-based interventions do not exacerbate existing health disparities. This includes developing culturally appropriate measures, ensuring compatibility with diverse technologies, and addressing potential biases in algorithm development. As an example of the importance of this point, early version of the smartphone applications developed for early studies worked only in Apple iPhones, which are typically more expensive than Android phones, and the requirement to own a smartphone capable of running the task (which was processor and resource demanding) in the first place excluded hundreds of participants in early recruitment stages.

Informed Consent for Continuous Monitoring: As assessment moves toward passive, continuous monitoring, new models of informed consent will be needed that clearly communicate what data is collected, how it is used, and what insights it might reveal about an individual's mental state.

10. Conclusions

This thesis successfully achieved its stated aim to develop novel measures of interoception that can be deployed remotely. The research journey from biometric-driven therapeutics to interoception measurement development illuminates both the promise and limitations of technology-based approaches to mental health intervention.

The therapeutic intervention studies demonstrate that smartphone-based digital therapeutics incorporating biometric feedback can produce significant and sustained improvements in anxiety and well-being across different populations. Effect sizes observed in these studies (Cohen's d = 0.66-0.96) are comparable to or exceed those typically reported for digital mental health interventions in recent meta-analyses (Buntrock, 2024).

However, the consistent observation that participants struggled to meaningfully interpret biometric data revealed a fundamental challenge in purely data-driven approaches to mental health intervention. This insight drove the development of novel interoception measures that address critical limitations in existing approaches and provide new tools for understanding individual differences in bodily awareness.

The CARED task established proof-of-concept for smartphone-based interoception measurement, demonstrating that cardiac interoceptive accuracy can be assessed in naturalistic environments rather than only in controlled laboratory settings. This addresses a recognised limitation of laboratory-based approaches, although the present study does not constitute a formal validation of the CARED task: validation against established interoception measures, evaluation of test-retest reliability, and comparison with matched control tasks were all outside the scope of this proof-of-concept work and remain priorities for future research (see s. 8.1.5.2). The finding that approximately one-third of participants were classified as interoceptive is consistent with the base rates reported for traditional heartbeat counting tasks (Brener & Ring, 2016); this base-rate convergence is reassuring but does not by itself establish convergent validity between the CARED task and the heartbeat counting task at the level of individual-participant performance.

Most importantly, this thesis establishes the conceptual and methodological foundation for developing personalized digital mental health interventions based on interoceptive

phenotyping. Rather than applying one-size-fits-all approaches, future interventions could assess individuals' interoceptive profiles and tailor biometric feedback accordingly. Individuals with poor interoceptive accuracy might benefit from interoceptive training before engaging with biometric feedback, while those with adequate interoceptive abilities might immediately benefit from real-time physiological monitoring.

The research also highlights that the effectiveness of biometric feedback interventions may depend critically on users' ability to connect objective physiological data with subjective emotional experience. This connection appears to be a learnable skill that could be enhanced through targeted training, though the specific mechanisms and optimal training approaches remain to be established.

Looking forward, this work provides essential tools and theoretical insights for developing the next generation of personalized digital mental health interventions. The novel interoception measures enable researchers to test hypotheses about the role of bodily awareness in mental health, while the therapeutic studies demonstrate the promise of biometric feedback approaches when appropriately implemented.

The ultimate vision emerging from this research is of digital therapeutics that can assess, understand, and enhance individuals' connections with their bodily signals, potentially addressing the fundamental disconnection from bodily experience that can affect therapeutic outcomes. While significant challenges remain in translating these findings to clinical practice, this thesis establishes the scientific foundation for pursuing this important goal.

Appendix A

Multimedia Appendix 1. Content and theoretical contextualization of the three psycho-educational courses contained in BioBase.

Course title and HSE/JDR content theme covered	Aims and objectives of each course	Course-targeted HSE questions	Modules covered
Managing your Workload - Demands	AIMS: To reduce the perceived stress experienced due to overwhelming demands. Objectives: To recognise stress is an imbalance of demands and resources. To recognise and categorise current demands and resources. To reduce demands and increase resources through time management techniques.	Different groups at work demand things from me that are hard to combine I have unachievable deadlines I have to work very intensively I have to neglect some tasks because I have too much to do I am unable to take sufficient breaks I am pressured to work long hours I have to work very fast I have unrealistic time pressures	JDR model[1]; basic psychological needs (Mastery)[2]; time management techniques[3]; concept of 'Flow'[4]
Taking Back Control - Control	AIM: To increase the perception of control Objectives: Understanding burnout and its consequences on decision making, Reframing situations to feel more in control. Using recovery techniques as a way to prevent burnout and feel more in control.	I can decide when to take a break I have a say in my own work speed I have a choice in deciding how to do at work I have a choice in deciding what I do at work I have some say over the way I work My working time can be flexible	JDR model[1]; basic psychological needs (Autonomy)[2]; Conservation of Resources Model[5]; Recovery[6]; Psychological Detachment[7]; CBT (Circle of Control, and the link between thoughts, feelings and emotions)[8]; self-compassion[9].
Shaping social support - Support	AIM: To increase social connections and social support. Objectives; To understand the health implications of a lack of social support. To reconnect via small daily interactions, and build on existing social connections linked to self-identify.	If work gets difficult my colleagues will help me I get help and support I need from colleagues I receive the respect at work I deserve from my colleagues My colleagues are willing to listen to my work-related problem	Basic psychological needs (Connectedness)[2]; CBT (Transactional Analysis)[10]; Positive Psychology (acts of kindness, gratitude)[11]; Social Psychology[12]; JDR - Social resources[13].

Appendix B

Supplementary Materials For Chapter 5

End of study questions

At the end of the procedure, participants were asked the following questions:

1. To answer the questions during the study, did you ever manually check your pulse by touching your fingers against your wrist or neck? If yes, how often did you do this? Please estimate from 0% to 100%, how many of the questions you answered by manually checking your pulse.
1. Did you use any other methods to find out your heart rate?
2. The questions in this study asked you “Is your heart beating faster than usual?”. Different people can have different ideas of what “usual” heart rate is. For example, you could say that your heart rate when resting is your "usual". Or, "usual" might be what you feel your heart rate normally is at this time of the day.

   Could you please briefly explain what you thought “usual” heart rate meant when answering the questions?

Appendix C

Supplement

[S1] Instructions for the PST and PST Control Task

Baseline

1. How well can you match a sound with your heartbeat? Let's find out.
2. First, find a quiet place where you can comfortably sit upright for around 20 minutes. Please don't wear earphones. We will begin by capturing a couple of minutes of your heart rate, to get a sense for how your heart does its thing.
3. Get ready to check your heartbeat

    a. We will shortly turn on the LED Flash and camera on this phone and will use it to take your heart rate. Please place your index finger over both the camera and flash and try to turn the square red.

[Live camera display requesting covering of flash and camera with the finger to 'make the square red']

4. Please unmute your device and make sure you can hear the tone before continuing

[Beep button that plays tone when pressed and can be repeated as many times as required]

[Continue button displays when beep button pressed]

Control task

1. In this task you will be presented with a series of sounds. Your task is to decide which series of sounds happen most closely in time after the flashing of the shape.

3. You will hear three sets of three tones. One set of tones will happen just after the flash of the shape with almost no delay, one set will follow with a slightly bigger delay, and one set will follow with an even bigger delay.

4. Your job is to try and work out the order of delays. You will get them in a random order.

5. All tones are triggered by the flashing shape. This means the tones cannot come before the flashing shape.

6. Each series of tones will be presented with a coloured shape. When asked, rank the shapes in order of those that most closely follow each flash.

7. Remember, tones cannot come before the flash of the shape. When asked to rank, you should put the shape associated with the tone that was closest after the flash first, followed by the next and least close at the end.

8. Don’t worry if you find this task difficult, people are sometimes better than they think they are!

9. While you complete the task, please keep the index finger of one hand over the camera and flash, and the five fingers of the other hand touching the phone screen. There will be a blue box on the screen where you should place your five fingers at the start of each trial. You can move them when the box turns green, but do not lift them off the screen. There will be a short video to demonstrate.

10. Please click play on the video below - you must watch it to the end to continue.

[Example video]

11. During the task, you’ll often need 5 fingers on the screen. You can move them (but not lift them) once you have started. Please press the box with 5 fingers.

12. Now a few practice trials. Remember to keep your finger over the camera and flash during the task.

13. Rank the tones in order following the flash. Remember, tones cannot come before the flash.

14. The tone closest to the flash should be first, and the tone furthest after the flash should be last. Tap an icon to add or remove it to the list.

At the end of the CONTROL task

15. How much do you agree with the following statement? I am confident that I could detect differences in the delays between the flashing shape and the sounds (response options from Strongly disagree to Strongly agree on a 7-point scale).

16. How much do you agree with the following statement? I am confident that I could order the sounds in order of time after the flashing shape (response options from Strongly disagree to Strongly agree on a 7-point scale).

Heartbeat task

1. In this task you will be presented with a series of tones. Your task is to decide which series contains tones that happen most closely in time after your heartbeats.
2. You will hear three sets of three tones. One set of tones will happen just after your heartbeats with almost no delay, one set will follow with a slightly bigger delay, and one set will follow with an even bigger delay.
3. Your job is to try and work out the order of delays. You will get them in a random order.
4. All tones are triggered by your heartbeat. This means the tones cannot come before your heartbeat.

5. Each series of tones will be presented with a coloured shape. When asked, rank the shapes in order of those that happen most closely after each heartbeat.
6. Remember, tones cannot come before your heartbeat. When asked to rank, you should put the shape associated with the tone that was closest after your heartbeat first, followed by the next, and the furthest away at the end.
7. People can feel their heartbeat in different places in their body, such as their chest or fingers. Where do you currently feel your heartbeat most clearly? Please show us on the body map below. You can choose any of the highlighted body parts. If you can't feel it at all, select none. PLEASE only use heartbeats you feel in that location throughout all trials.

[Body map]

8. Don't worry if you find this task difficult, people are sometimes better than they think they are! For the duration of this task, please do not actively try to feel your pulse with your hand; we are only interested in what you feel inside! You might feel your heartbeat in various bodily locations. Just make sure you pick one and stick to using that one during the task. When you are ready to start, please sit comfortably upright and press "continue".

9. While you complete the task, please keep the index finger of one hand over the camera and flash, and the five fingers of the other hand touching the phone screen. There will be a blue box on the screen where you should place your five fingers at the start of each trial. You can move them when the box turns green, but do not lift them off the screen. There will be a short video to demonstrate.

10. Please click play on the video below - you must watch it to the end to continue.

[Example video]

11. During the task, you'll often need 5 fingers on the screen. You can move them (but not lift them) once you have started. Please press the box with 5 fingers.

12. Now a few practice trials. Remember to keep your finger over the camera and flash during the task.

13. Now rank the sounds in order following your heartbeat. Remember, tones cannot come before your heartbeat.

14. Ready to start?

15. Rank the tones in order following your heartbeat. The tone closest to your heartbeat should be first, and the tone furthest after your heartbeat should be last. Tap an icon to add or remove it to the list.

At the end of the whole HEARTBEAT task

16. How much do you agree with the following statement? I am confident I could feel my heartbeat during the heartbeat task. (response options from Strongly disagree to Strongly agree on a 7-point scale).

17. How much do you agree with the following statement? I am confident that I could detect differences in the delays between my heartbeat and the sounds (response options from Strongly disagree to Strongly agree on a 7-point scale).

18. How much do you agree with the following statement? I am confident that I could order the sounds in order of time after my actual heartbeat (response options from Strongly disagree to Strongly agree on a 7-point scale).

19. How many locations on your body did you use to feel your heartbeat during the task?

a. Please indicate where they are on the body map below.

[Body map]

Appendix D

CONSORT 2025 reporting checklist for the Chapter 2 randomised controlled trial

This appendix presents a retrospective audit of Chapter 2 (the BioBase student RCT; Ponzo et al., 2020) against the 30-item CONSORT 2025 statement (Hopewell et al., 2025). The trial was conducted in 2019–2020, predating the CONSORT 2025 update, and was reported in line with CONSORT 2010 in its peer-reviewed publication. The audit below identifies, for each CONSORT 2025 item, the location within Chapter 2 at which the item is addressed, and notes items where retrospective alignment is incomplete or impossible.
Status categories used in the table are: **Reported** (the item is addressed in the chapter, with the location specified); **Partially reported** (some, but not all, elements of the item are present); **Reported in published version** (the item is in Ponzo et al., 2020, but not in the chapter as adapted); **Not reported** (the item is absent and a retrospective addition is recommended or the gap is noted as a limitation); and **Not applicable** (the item does not apply to a digital intervention trial of this type). Where reporting reflects work-in-progress revisions documented in the response-to-examiners letter, the relevant panel-point identifier (e.g., C2-1, C2-2, C2-4) is given.

Item	CONSORT 2025 item description	Reporting status	Location in Chapter 2	Notes / retrospective limitations
Title and abstract				
1a	Identification as a randomised trial in the title	**Reported**	Chapter title: "Chapter Two: Digital Biometric Feedback for Student Anxiety: A Randomized Controlled Trial". Chapter opening page.	Compliant with CONSORT 2025.
1b	Structured summary of trial design, methods, results, and conclusions	**Partially reported**	Chapter abstract (s. 2.2).	The published version (Ponzo et al., 2020) carries a structured abstract; the chapter-as-adapted uses a paragraph-form abstract that nevertheless covers design, methods, results and conclusions. Considered acceptable in a thesis context.
Open science (new section in CONSORT 2025)				
2	Name of trial registry, identifying number (with URL) and date of registration	**Reported**	Pre-registration declaration, s. 2.4 (between s. 2.4.1 Ethics Approval and s. 2.4.2 Study Design), p. 42.	Pre-registered on the Open Science Framework prior to enrolment (https://osf.io/2zd45/); registration ID 2zd45. Note: the trial was not separately registered on a ClinicalTrials.gov-style registry; OSF is the available record.
3	Where the trial protocol and statistical analysis plan can be accessed	**Reported**	Pre-registration declaration, s. 2.4, p. 42.	OSF pre-registration sets out the design, randomisation/minimisation procedures, primary and secondary outcomes, and analysis plan

				(https://osf.io/2zd45/). Analyses follow the pre-registered plan; no deviations.
4	Where and how individual de-identified participant data (including data dictionary), statistical code, and any other materials can be accessed	**Not reported**	—	New CONSORT 2025 item. Individual de-identified participant-level data are not currently shared on OSF; the OSF page hosts the pre-registration. To strengthen alignment, a brief data-availability statement could be added to s. 2.4 (e.g., "Aggregated trial outputs are available at https://osf.io/2zd45/; individual de-identified data are available from the corresponding author on reasonable request.").
5a	Sources of funding and other support, and role of funders	**Reported in published version**	Ponzo et al. (2020), "Acknowledgments" and "Conflicts of Interest" sections of the published article (https://mhealth.jmir.org/2020/4/e17767).	Funding and role-of-funder statements are in the underlying JMIR publication. Recommend a short cross-reference in the chapter (e.g., end of Methods) so readers do not have to consult the source paper.
5b	Financial and other conflicts of interest of the manuscript authors	**Reported in published version**	Ponzo et al. (2020), "Conflicts of Interest" section.	New CONSORT 2025 item. The published version declares author affiliations with Huma Therapeutics Ltd (the commercial developer of BioBase). Recommend explicit cross-reference in the chapter.
Introduction				
6	Scientific background and rationale	**Reported**	Chapter 2 Introduction (s. 2.3).	Background on student mental health, the rationale for digital therapeutics, and the rationale for biometric feedback are presented.
7	Specific objectives related to benefits and harms	**Partially reported**	End of Chapter 2 Introduction (s. 2.3).	Benefits objectives explicitly stated (reductions in anxiety, improvements in wellbeing, reduction in depressive symptoms). Harms objectives are not pre-specified — see item 15.
Methods				
8	Patient and public involvement in design, conduct and reporting	**Not reported**	—	New CONSORT 2025 item. PPI was not formally documented for the Chapter 2 trial. This is being addressed across the thesis via panel point G1 (PPI/GRIPP-2 sections) and will be added to s. 2.4 once the GRIPP-2 audit is completed.
9	Trial design (type, allocation ratio,	**Reported**	s. 2.4.2 (Study Design).	Parallel-group, waitlist-controlled RCT; 1:1 allocation

	framework)			ratio; assessments at baseline, 2, 4 and 6 weeks. Framework implicit (superiority); could be made explicit in a brief revision.
10	Important changes to the trial after it commenced, including any outcomes or analyses not pre-specified, with reason	**Partially reported**	s. 2.4.4 (Randomization and Blinding).	The single change documented is the move from single-blind to open-label design (with logistical-communication rationale). No outcome or analysis changes from the pre-registration are reported.
11	Trial setting (e.g., community, hospital) and locations (countries, sites)	**Reported**	s. 2.4.3 (Recruitment).	Setting: UK higher-education community. Recruitment via institutional participant pools at multiple UK universities, social media, mailing lists, flyers, and university staff.
12a	Eligibility criteria for participants	**Reported**	s. 2.4.3 (Recruitment).	Six inclusion criteria and two exclusion criteria explicitly listed (age 18–25, full-time UK student, DASS-21 thresholds, iPhone 6+, no prior psychiatric/ neurological history, not pregnant, English-speaking; exclusions for current therapy/counselling).
12b	Eligibility criteria for sites and individuals delivering the interventions	**Not applicable**	—	New CONSORT 2025 item. Not applicable: the intervention is digitally self-administered (smartphone app and wearable) with no human deliverers requiring specific training or certification.
13	Intervention and comparator with sufficient detail to allow replication; access to additional materials (e.g., intervention manual)	**Reported**	s. 2.4.5 (Intervention Group) and corresponding Waitlist Control description.	BioBase app + BioBeam wearable, 42 modules grouped under the Job Demands–Resources model (demands, control, support). Module-level detail is provided in Multimedia Appendix 1 of the published version (Ponzo et al., 2020); chapter could add an explicit cross-reference.
14	Prespecified primary and secondary outcomes, with measurement variable, analysis metric, method of aggregation, and time point	**Reported**	s. 2.4.6 (Primary Outcome), s. 2.4.7 (Secondary Outcome), s. 2.4.8 (Additional Measures).	Primary: STAI-S-6 at 4 weeks. Secondary: WEMWBS (wellbeing). Additional: DASS-21 depression and PHQ-9. Measurement variables, time points and aggregation are specified.
15	How harms were defined and assessed (e.g., systematically, non-systematically)	**Not reported**	—	New CONSORT 2025 item. The trial did not implement a formal harms-surveillance procedure (no adverse-event reporting framework was specified in the protocol). This is a retrospective reporting limitation; a brief acknowledgement in s. 2.7

				(Limitations) is recommended to state explicitly that no systematic harms assessment was conducted, and that future trials of biometric digital therapeutics should pre-specify adverse-event monitoring.
16a	How sample size was determined, including all assumptions	**Reported**	s. 2.4.4.1 (Power), p. 65 (post-C2-1 revision).	C2-1 revision (29 May 2026) expanded this section to ~175 words: anchored to the Kawadler et al. (2020) feasibility pilot (d ≈ 0.96); acknowledges the optimism of this estimate relative to meta-analytic norms for digital mental-health interventions (Pauley et al., 2023, g = 0.80 [95% CI 0.68–0.93]); states assumed and observed dropout rates.
16b	Explanation of any interim analyses and stopping guidelines	**Not applicable**	—	No interim analyses or stopping rules were pre-specified for this trial.
17a	Who generated the random allocation sequence and the method used	**Partially reported**	s. 2.4.4 (Randomization and Blinding).	Method: the R Minirand package was used to generate the allocation list. The chapter does not specify which member of the research team generated the sequence. A short clarification could be added.
17b	Type of randomisation and details of any restrictions (e.g., stratification, blocking and block size)	**Reported**	s. 2.4.4 (Randomization and Blinding).	Minimisation was used (not block randomisation), with four minimisation factors: gender (2 levels), age (7 levels), DASS-21 anxiety severity (5 levels), and DASS-21 stress severity (5 levels). The first participant was allocated at random; subsequent allocations minimised between-group imbalance on these factors.
18	Mechanism used to implement the random allocation sequence (e.g., central computer; sequentially numbered, opaque, sealed containers), including any steps to conceal the sequence	**Not reported**	—	The chapter does not explicitly describe the allocation-concealment mechanism (e.g., whether allocation was concealed from the researcher who confirmed eligibility before assignment). A brief clarification is recommended.
19	Whether the personnel who enrolled and those who assigned participants to interventions had access to the random allocation sequence	**Not reported**	—	Not explicitly addressed in the chapter. A brief clarification of who enrolled and who assigned is recommended.

20a	Who was blinded after assignment to interventions (e.g., participants, providers, outcome assessors, data analysts)	**Reported**	s. 2.4.4 (Randomization and Blinding).	Originally designed as single-blind; subsequently unblinded due to communication-clarity considerations between the research team and participants. The trial was therefore open-label in its final form.
20b	If blinded, how blinding was achieved and description of the similarity of interventions	**Not applicable**	—	Not applicable: the trial was unblinded. The waitlist control did not receive a sham app, so participant blinding was not feasible.
21a	Statistical methods used to compare groups for primary and secondary outcomes, including harms	**Reported (benefits); N/A (harms)**	s. 2.4.4.3 (Data Analysis).	Linear mixed models with group, time, and group × time interaction as fixed effects; subjects (and time within subjects) as nested random effects. Planned comparisons via paired-samples t-tests; Cohen's d with pooled SD. Harms statistical methods: N/A (see item 15).
21b	Definition of who is included in each analysis (e.g., all randomised participants) and in which group	**Partially reported / under revision**	s. 2.4.4.2 (Data Exclusion); to be expanded via panel point C2-2.	The chapter currently describes a per-protocol analytic strategy (participants who did not download/open the app and those who scored in the normal range at baseline were excluded). This departs from CONSORT 2025's emphasis on intention-to-treat. C2-2 is in progress and will (a) make the per-protocol nature explicit, (b) discuss the direction of bias introduced, and (c) — if the dataset permits — report a sensitivity ITT analysis as a sub-appendix.
21c	How missing data were handled in the analysis	**Reported**	s. 2.4.4.3 (Data Analysis).	New CONSORT 2025 item. The chapter states that linear mixed models were chosen specifically for their handling of incomplete data across time points (Hesselmann, 2018).
21d	Methods for any additional analyses (e.g., subgroup and sensitivity), distinguishing prespecified from post hoc	**Not applicable / planned**	—	No subgroup or sensitivity analyses were originally reported. The sensitivity ITT analysis being added under C2-2 will be reported as a post-hoc sensitivity analysis with this distinction stated explicitly.
Results				
22a	For each group, the numbers of participants who were randomly assigned, received intended intervention, and were analysed for the	**Reported**	Figure 2 (CONSORT-style flow diagram) and accompanying Results text (s. 2.5).	Figure 2 documents: 805 screened, 543 excluded (ineligible/duplicate), 262 randomised (130 intervention / 132 control), with completers at T0, T1, T2 and T3 broken out by group.

	primary outcome			
22b	For each group, losses and exclusions after randomisation, together with reasons	**Reported**	Figure 2 and Results text (s. 2.5).	Reasons documented at each time point (e.g., "not present at T0", "did not download the app", "excluded from T0", "excluded from T0/T2").
23a	Dates defining the periods of recruitment and follow-up for outcomes of benefits and harms	**Partially reported**	s. 2.4.3 (Recruitment).	Recruitment dates explicitly reported (October–November 2019). End-of-follow-up date is implicit (recruitment + 6 weeks ≈ January 2020) but not explicitly stated. A short clarification is recommended.
23b	If relevant, why the trial ended or was stopped	**Not applicable**	—	The trial ran to its planned protocol completion; no early stopping. A one-sentence statement to this effect could be added to the Results section.
24a	Intervention and comparator as they were actually administered (e.g., who delivered, adherence, fidelity)	**Partially reported**	s. 2.5 (Results) — usage statistics and exclusions; s. 2.6 (Discussion).	New CONSORT 2025 item. Adherence indicators are present (module engagement, exclusions for non-openers). Fidelity in the strict sense is difficult to assess for self-administered digital interventions; a brief acknowledgement could be added that the app delivered the intended content (technical fidelity) but participant-level engagement varied.
24b	Concomitant care received during the trial for each group	**Not reported**	—	New CONSORT 2025 item. The chapter notes that participants currently in therapy or counselling at screening were excluded, but does not report whether participants accessed any concomitant mental-health care during the 4–6-week trial period. This is a retrospective reporting limitation.
25	A table showing baseline demographic and clinical characteristics for each group	**Reported**	Table 1 in the published version (Ponzo et al., 2020), reproduced in the chapter Results.	Baseline demographics (sex, age) and clinical characteristics (DASS-21 anxiety and stress) reported by group with p-values for between-group baseline comparisons.
26	For each primary and secondary outcome, by group: number analysed, number with available data, result, estimated effect size and precision; for binary outcomes, both	**Reported**	Tables in Results (s. 2.5), reproducing Tables from Ponzo et al. (2020).	Revised CONSORT 2025 item. Effect estimates, 95% CIs, p-values, and Cohen's d reported for each outcome at each time point. n per analysis is shown in LMM table footers. No binary outcomes (so absolute/relative effect size distinction does not apply).

	absolute and relative effect size			
27	All harms or unintended events in each group	**Not reported**	—	See item 15. No systematic harms-surveillance procedure was implemented and no harms are reported. To be acknowledged as a retrospective reporting limitation in s. 2.7 (Limitations).
28	Any other analyses performed, including subgroup and sensitivity analyses, distinguishing prespecified from post hoc	**Not applicable / planned**	—	No ancillary analyses were originally reported. The sensitivity ITT analysis under C2-2 will, if produced, be reported under this item as post hoc.
Discussion				
29	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	**Partially reported / under revision**	s. 2.6 (Discussion); revisions in progress under panel point C2-4.	Current interpretation language is overstated (panel concern); C2-4 will reframe outcomes as preliminary indicative evidence, acknowledge candidate non-specific effects of the CBT/mindfulness/biofeedback combination, and reposition BioBase as a candidate treatment adjunct rather than a standalone efficacious intervention.
30	Trial limitations, addressing sources of potential bias, imprecision, generalisability, and, if relevant, multiplicity of analyses	**Reported / under expansion**	s. 2.7 (Limitations), p. 49 (post-C2-1 revision); further expansion via C2-2 and C2-4.	C2-1 revision (29 May 2026) added explicit acknowledgement that the trial was powered for an optimistic effect, and reported the observed ~53% dropout. C2-2 will add limitations text on the per-protocol vs ITT issue. C2-4 will further calibrate the Discussion conclusions. Generalisability discussion (formerly a separate CONSORT 2010 item) is now incorporated under this item, per CONSORT 2025.

Note on the CONSORT flow diagram

Figure 2 in Chapter 2 documents participant flow from screening through randomisation to analysis at each time point (T0, T1, T2, T3), with reasons for exclusion shown at each stage. The structure is consistent with the CONSORT 2025 four-phase flow diagram (enrolment, intervention allocation, follow-up, data analysis), although the labelling does not explicitly use those four CONSORT phase headings. The diagram is therefore broadly compliant with CONSORT 2025; an updated version using the CONSORT 2025 phase labels could be substituted to make the alignment explicit, but the substantive information required by item 22a–b is present.

Summary of retrospective limitations

Seven CONSORT 2025 items are not reported in the chapter or in the published version: item 4 (individual de-identified data sharing — new CONSORT 2025 item); item 8 (PPI documentation — new item, being added across the thesis under panel point G1); item 15 (how harms were defined and assessed — new item); item 18 (allocation concealment mechanism); item 19 (who enrolled and who assigned); item 24b (concomitant care — new item); and item 27 (all harms or unintended events). Of these, items 4, 8, 18, 19 and 24b can be added retrospectively as brief textual clarifications in the chapter Methods, with appropriate honesty about what was and was not documented at the time of conduct. Items 15 and 27 reflect the absence of a pre-specified harms-surveillance procedure in the original trial; these are acknowledged as retrospective reporting limitations in s. 2.7 (Limitations) and as a methodological priority for future trials of biometric digital therapeutics.

Reference

Hopewell, S., Chan, A.-W., Collins, G. S., Hróbjartsson, A., Moher, D., Schulz, K. F., Tunn, R., Aggarwal, R., Berkwits, M., Berlin, J. A., Bhandari, N., Butcher, N. J., Campbell, M. K., Chidebe, R. C. W., Elbourne, D., Farmer, A., Fergusson, D. A., Golub, R. M., Goodman, S. N., … Boutron, I. (2025). CONSORT 2025 statement: Updated guideline for reporting randomised trials. *BMJ, 389*, e081123. https://doi.org/10.1136/bmj-2024-081123

Appendix E

Intention-to-treat sensitivity analysis for the Chapter 2 randomised controlled trial

Preamble. This appendix presents a sensitivity analysis addressing the panel concern that the analyses reported in Chapter 2 (Ponzo et al., 2020) followed a per-protocol logic rather than the intention-to-treat (ITT) principle now emphasised in CONSORT 2025 (Hopewell et al., 2025). A formal re-analysis of the raw trial data was not feasible: the dataset is held by Huma Therapeutics Ltd, which acquired BioBeats Ltd in 2021, and is no longer accessible to the candidate following the decommissioning of the company's research infrastructure (see s. 8.3 for the thesis-wide transparency reflection on this constraint). The analysis below is therefore constructed from the published summary statistics in Ponzo et al. (2020, Tables 2 and 3, and Figure 2) together with the meta-analytic per-protocol-versus-ITT bias estimates of Linardon and Fuller-Tyszkiewicz (2020). Three converging lines of evidence — a Linardon-derived bias correction and three scenario-based imputations for participants lost after baseline measurement — yield an ITT-equivalent range that is reported in §5 and summarised in s. 2.7 of the main thesis.

1. What the published analysis actually estimated

Before computing an ITT correction, it is important to be precise about what the published analysis estimated. The chapter describes the analytic sample as the 146 randomised participants who met the post-randomisation baseline DASS-21 thresholds and provided a T0 measurement (Ponzo et al., 2020, Figure 2). Within that sample, the published linear mixed model used maximum-likelihood estimation, which provides unbiased estimates under the missing-at-random (MAR) assumption when participants drop out after baseline. The reported LMM included 123 participants and 491 observations (Ponzo et al., 2020, Table 2), meaning that 23 of the 146 baseline-measured participants provided only a T0 measurement and no follow-up data and were not included in the LMM.

The published analysis is therefore best characterised as a **modified intention-to-treat (mITT) analysis** of the 146 randomised participants with baseline measurement, restricted to the 123 with at least one post-baseline observation. It departs from a strict CONSORT-2025-compliant ITT analysis in two respects, addressed separately below: (a) the 23 baseline-measured-but-never-followed participants are excluded from the LMM (§3), and (b) the 116 randomised participants who never reached baseline measurement (44% of the randomised sample) are excluded altogether (§4).

Table E1. Trial flow and the participants entering each analytic frame.

Stage	Intervention	Control	Total	Comment
Randomised	130	132	262	Full randomised sample.
Measured at T0 (baseline)	72	74	146	116 randomised participants (44%) excluded before baseline measurement — see §4.
Entered LMM analysis	≈ 60	≈ 63	123	Per Table 2 (Ponzo et al., 2020): N ID = 123, 491 observations. Approximate within-group split inferred.
Completed T2 (4 weeks)	55	61	116	Per Table 3 t-test df.
Completed T3 (6	59	64	123	Per Table 3 t-test df.

weeks)				

2. Bias-correction synthesis using the Linardon and Fuller-Tyszkiewicz (2020) meta-analytic benchmark

Linardon and Fuller-Tyszkiewicz (2020) meta-analysed 70 randomised controlled trials of smartphone-delivered mental-health interventions. Across the subset of trials that reported both per-protocol and intention-to-treat estimates of the primary outcome, the mean difference between the two estimates was Δd = 0.18, with per-protocol estimates systematically higher than ITT estimates. This Δd = 0.18 provides an empirical anchor for estimating the expected magnitude of the per-protocol-to-ITT correction in the present trial.

Applied to the present trial's between-groups effect sizes derived from the LMM Group × Time interactions (T2: 7.09 STAI-S-6 points, d ≈ 0.65 using pooled SD at T2; T3: 4.56 STAI-S-6 points, d ≈ 0.35 using pooled SD at T3), the Linardon-corrected ITT-equivalent estimates are:

- At T2 (4 weeks, post-intervention): d ≈ 0.65 − 0.18 = 0.47, equivalent to approximately 5.14 STAI-S-6 points.
- At T3 (6 weeks, follow-up): d ≈ 0.35 − 0.18 = 0.17, equivalent to approximately 2.24 STAI-S-6 points.

The Linardon-derived correction places the T2 ITT-equivalent estimate in the small-to-medium range, consistent with the recent meta-analytic mean for digital mental-health interventions overall (g = 0.80, 95% CI [0.68, 0.93]; Pauley et al., 2023). The T3 estimate falls below the small range under this correction, which is consistent with the broader observation that intervention effects in digital mental-health trials tend to attenuate during post-intervention follow-up. A caveat is that Δd = 0.18 is the meta-analytic mean across 70 trials; the variability of the per-protocol-to-ITT gap across individual trials is not characterised here, so the Linardon-corrected estimate should be read as a single point on the distribution of plausible ITT estimates rather than as a precision bound.

3. Scenario-based sensitivity analysis using published summary statistics

A complementary approach is to construct hypothetical ITT estimates by assigning outcome values to the participants who were measured at baseline but did not reach the follow-up assessments. Under this approach we expand the analytic sample to the full 146 baseline-measured participants (i.e., we extend the published analysis from n = 123 to n = 146 by imputing T2 and T3 values for the 23 baseline-only participants and, more conservatively, for participants lost between T0 and each follow-up point). Three scenarios are considered.

Scenario A — Worst case (no-change imputation)

Participants in either arm who were measured at T0 but not at the follow-up time point are assigned their group's baseline mean at the follow-up. This implements the assumption that drop-out is associated with no benefit from being assigned to the intervention. At T2, with 17 intervention participants and 13 control participants imputed at baseline values, the new group means are 47.59 (intervention) and 51.06 (control), giving an intervention-versus-control between-groups change of 5.38 STAI-S-6 points and d ≈ 0.50. At T3, with 13 intervention participants and 10 control participants imputed at baseline values, the equivalent between-groups effect is 3.64 STAI-S-6 points and d ≈ 0.28.

Scenario B — Best case (as-completers imputation)

Imputed participants are assigned their group's completer mean at the follow-up time point. This implements the assumption that drop-out is unrelated to outcome, and reproduces the published per-protocol estimates exactly: T2 effect = 6.92 STAI-S-6 points, d ≈ 0.64; T3 effect = 4.56 STAI-S-6 points, d ≈ 0.35.

Scenario C — Jump-to-reference (MNAR imputation favouring control)

Intervention participants who dropped out are assigned the control completer mean at the follow-up time point; control dropouts are assigned their group baseline mean. This is the standard 'jump-to-reference' approach used to test missing-not-at-random sensitivity in pharmaceutical trials (Carpenter et al., 2013), and it implements the assumption that intervention drop-outs ceased to benefit from the intervention and converged on the control trajectory. T2 effect = 5.47 STAI-S-6 points, d ≈ 0.50; T3 effect = 4.38 STAI-S-6 points, d ≈ 0.34.

Across the three scenarios, the T2 between-groups effect ranges from 5.38 to 6.92 STAI-S-6 points (d ≈ 0.50 to 0.64), and the T3 between-groups effect ranges from 3.64 to 4.56 STAI-S-6 points (d ≈ 0.28 to 0.35). These ranges sit close to the Linardon-derived corrections in §2, providing convergent evidence that the published mITT-like LMM estimate is biased upward by something on the order of 1–2 STAI-S-6 points at T2 and ~0–1 STAI-S-6 points at T3 relative to a hypothetical ITT estimate on the full baseline-measured sample. A caveat applicable to all three scenarios is that the imputations do not propagate uncertainty about the imputed values; the resulting effect sizes are point estimates rather than full inferential statistics, and the published 95% confidence intervals (LMM T2: [2.94, 11.24]; LMM T3: [0.49, 8.63]) cannot be straightforwardly re-derived from summary statistics alone.

Table E2. Summary of ITT-equivalent estimates from all methods.

Estimate	T2 (STAI pts)	T2 d	T3 (STAI pts)	T3 d
Published LMM (mITT-like, n = 123 via ML; Ponzo et al., 2020)	7.09 [2.94, 11.24]	0.65	4.56 [0.49, 8.63]	0.35
Linardon-corrected (Δd = 0.18; Linardon & Fuller-Tyszkiewicz, 2020)	≈ 5.14	0.47	≈ 2.24	0.17
Scenario A — worst case (lost-after-T0 had no change)	5.38	0.50	3.64	0.28
Scenario B — best case (lost-after-T0 as completers)	6.92	0.64	4.56	0.35
Scenario C — jump-to-reference (int dropouts as ctrl completers)	5.47	0.50	4.38	0.34
Sensitivity range across all methods	**5.14 – 7.09 STAI pts**	**0.47 – 0.65**	**2.24 – 4.56 STAI pts**	**0.17 – 0.35**

4. The structural limitation: pre-baseline exclusion of 116 randomised participants

A separate and more consequential departure from strict ITT is that 116 of the 262 randomised participants (44%) did not reach baseline measurement — they were excluded either because they did not attend the T0 assessment or because their T0 DASS-21 scores fell below the inclusion thresholds (Ponzo et al., 2020, Figure 2). Both forms of attrition occurred between the random allocation and the first measurement and therefore cannot be addressed by maximum-likelihood handling of within-trial missingness, by the Linardon-derived bias correction, or by the scenario-based imputations in §3. A strict CONSORT-2025-compliant ITT analysis would require assigning hypothetical baseline and follow-up values to these 116 participants, which would in turn require access either to their screening-stage DASS-21 scores (to anchor a baseline imputation) or to the raw trial data (to apply more sophisticated multiple-imputation methods). Neither is feasible in the absence of the original dataset.

This pre-baseline attrition rate is high but not unusual for digitally recruited workplace and student mental-health trials: the mean meta-analytic short-term attrition rate in 70 smartphone mental-health RCTs is 24.1% (95% CI [19.3, 29.6]; Linardon & Fuller-Tyszkiewicz, 2020), and pre-baseline attrition (the additional loss between consent and first measurement) typically inflates that figure further. The 44% loss observed here is at the upper end of what is reported in the smartphone-RCT literature, and it represents a structural feature of the trial's recruitment-retention pipeline rather than a flaw in the analytic strategy. The implication is that the present trial's ITT-equivalent estimates, even after the corrections in §§2–3, should be interpreted as estimating the effect of being assigned to BioBase in a population that successfully reaches baseline measurement, rather than the effect of being randomised to BioBase in the originally targeted student-anxiety population.

5. Summary and reported ITT-equivalent estimates

Combining the bias-correction synthesis (§2) and the scenario-based imputations (§3), the trial's ITT-equivalent between-groups effect on the primary outcome (STAI-S-6) is estimated to lie in the following ranges (Table E2):

- At T2 (4 weeks, post-intervention): 5.14 – 7.09 STAI-S-6 points, equivalent to d ≈ 0.47 – 0.65. The trial's published mITT-like estimate (7.09 STAI-S-6 points, d ≈ 0.65) sits at the upper bound of this range; the Linardon-corrected and worst-case sensitivity estimates sit at the lower bound.

- At T3 (6 weeks, follow-up): 2.24 – 4.56 STAI-S-6 points, equivalent to d ≈ 0.17 – 0.35. The published mITT-like estimate (4.56 STAI-S-6 points, d ≈ 0.35) sits at the upper bound; the Linardon-corrected estimate (d ≈ 0.17) sits at the lower bound, which is below the conventional threshold for a small effect.

Both ranges are within the small-to-medium range reported in recent meta-analyses of digital mental-health interventions for anxiety (e.g., Pauley et al., 2023, g = 0.80 [0.68, 0.93]). The qualitative conclusion that the BioBase intervention reduced state anxiety at the post-intervention time point therefore holds across all imputation strategies considered here. The post-intervention follow-up effect at T3, however, is markedly less robust: under the Linardon-corrected estimate, the implied d ≈ 0.17 falls below conventional thresholds for a small effect and the originally reported confidence interval (LMM T3: [0.49, 8.63]) would, under a comparable shift, no longer exclude zero. This sensitivity-analytic observation strengthens the recalibration of the chapter's conclusions argued under panel point C2-4: the follow-up findings should be reported as preliminary indicative evidence of sustained effect rather than as a confirmed efficacy claim.

The estimates in this appendix should be read alongside the structural limitation discussed in §4: the 44% pre-baseline attrition rate means that any ITT-equivalent estimate computed from the available data describes the population of participants who successfully reached baseline measurement, not the population of all randomised participants. A strict CONSORT-

2025-compliant ITT analysis on all 262 randomised participants would require access to the original dataset and, plausibly, multiple imputation from screening-stage DASS-21 scores, neither of which is feasible. This is acknowledged as a transparency limitation in s. 8.3 of the main thesis.

References

Carpenter, J. R., Roger, J. H., & Kenward, M. G. (2013). Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics, 23*(6), 1352–1371. https://doi.org/10.1080/10543406.2013.834911

Hopewell, S., Chan, A.-W., Collins, G. S., Hróbjartsson, A., Moher, D., Schulz, K. F., … Boutron, I. (2025). CONSORT 2025 statement: Updated guideline for reporting randomised trials. *BMJ, 389*, e081123. https://doi.org/10.1136/bmj-2024-081123

Linardon, J., & Fuller-Tyszkiewicz, M. (2020). Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. *Journal of Consulting and Clinical Psychology, 88*(1), 1–13. https://doi.org/10.1037/ccp0000459

Pauley, D., Cuijpers, P., Papola, D., Miguel, C., & Karyotaki, E. (2023). Two decades of digital interventions for anxiety disorders: A systematic review and meta-analysis of treatment effectiveness. *Psychological Medicine, 53*(2), 567–579. https://doi.org/10.1017/S0033291721001999

Ponzo, S., Morelli, D., Kawadler, J. M., Hemmings, N. R., Bird, G., & Plans, D. (2020). Efficacy of the digital therapeutic mobile app BioBase to reduce stress and improve mental well-being among university students: Randomized controlled trial. *JMIR mHealth and uHealth, 8*(4), e17767. https://doi.org/10.2196/17767

References

Abraham, C., & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *27*(3), 379–387. https://doi.org/10.1037/0278-6133.27.3.379

Adams, K. L., Edwards, A., Peart, C., Ellett, L., Mendes, I., Bird, G., & Murphy, J. (2022). The association between anxiety and cardiac interoceptive accuracy: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, *140*(104754), 104754. https://doi.org/10.1016/j.neubiorev.2022.104754

Ahtinen, A., Mattila, E., Välkkynen, P., Kaipainen, K., Vanhala, T., Ermes, M., Sairanen, E., Myllymäki, T., & Lappalainen, R. (2013). Mobile mental wellness training for stress management: feasibility and design implications based on a one-month field study. *JMIR mHealth and uHealth*, *1*(2), e11. https://doi.org/10.2196/mhealth.2596

Aikens, K. A., Astin, J., Pelletier, K. R., Levanovich, K., Baase, C. M., Park, Y. Y., & Bodnar, C. M. (2014). Mindfulness goes to work: impact of an online workplace intervention. *Journal of Occupational and Environmental Medicine*, *56*(7), 721–731. https://doi.org/10.1097/JOM.0000000000000209

Ainley, V., Maister, L., & Tsakiris, M. (2015). Heartfelt empathy? No association between interoceptive awareness, questionnaire measures of empathy, reading the mind in the eyes task or the director task. *Frontiers in Psychology*, *6*, 554. https://doi.org/10.3389/fpsyg.2015.00554

Anderson, E. R., & Hope, D. A. (2009). The relationship among social phobia, objective and perceived physiological reactivity, and anxiety sensitivity in an adolescent population. *Journal of Anxiety Disorders*, *23*(1), 18–26. https://doi.org/10.1016/j.janxdis.2008.03.011

Andersson, G. (2009). Using the Internet to provide cognitive behaviour therapy. *Behaviour Research and Therapy*, *47*(3), 175–180. https://doi.org/10.1016/j.brat.2009.01.010

Andrade, L. H., Alonso, J., Mneimneh, Z., Wells, J. E., Al-Hamzawi, A., Borges, G., Bromet, E.,

Bruffaerts, R., de Girolamo, G., de Graaf, R., Florescu, S., Gureje, O., Hinkov, H. R., Hu, C., Huang, Y., Hwang, I., Jin, R., Karam, E. G., Kovess-Masfety, V., … Kessler, R. C. (2014). Barriers to mental health treatment: results from the WHO World Mental Health surveys. *Psychological Medicine*, *44*(6), 1303–1317. https://doi.org/10.1017/s0033291713001943

Anton, M. T., & Jones, D. J. (2017). Adoption of technology-enhanced treatments: Conceptual and practical considerations. *Clinical Psychology: A Publication of the Division of Clinical Psychology of the American Psychological Association*, *24*(3), 223–240. https://doi.org/10.1111/cpsp.12197

Arntz, A. (2002). Cognitive therapy versus interoceptive exposure as treatment of panic disorder without agoraphobia. *Behaviour Research and Therapy*, *40*(3), 325–341. https://doi.org/10.1016/s0005-7967(01)00014-6

Bacciu, D., Colombo, M., Morelli, D., & Plans, D. (2018). Randomized neural networks for preference learning with physiological data. *Neurocomputing*, *298*, 9–20.

Bakker, A. B., & Demerouti, E. (2007). The Job Demands-Resources model: state of the art. *Journal of Managerial Psychology*, *22*(3), 309–328. https://doi.org/10.1108/02683940710733115

Bakker, D., Kazantzis, N., Rickwood, D., & Rickard, N. (2016). Mental health smartphone apps: Review and evidence-based recommendations for future developments. *JMIR Mental Health*, *3*(1), e7. https://doi.org/10.2196/mental.4984

Baldwin, D. S., Allgulander, C., Bandelow, B., Ferre, F., & Pallanti, S. (2012). An international survey of reported prescribing practice in the treatment of patients with generalised anxiety disorder. *The World Journal of Biological Psychiatry: The Official Journal of the World Federation of Societies of Biological Psychiatry*, *13*(7), 510–516. https://doi.org/10.3109/15622975.2011.624548

Bandelow, B., Michaelis, S., & Wedekind, D. (2017). Treatment of anxiety disorders. *Dialogues in Clinical Neuroscience*, *19*(2), 93–107. https://doi.org/10.31887/dcns.2017.19.2/bbandelow

Barlow, D. H. (2017). *Anxiety and its disorders, second edition* (2nd ed.). Guilford Publications.

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews. Neuroscience*, *16*(7), 419–429. https://doi.org/10.1038/nrn3950

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. In *arXiv [stat.CO]*. arXiv. http://arxiv.org/abs/1406.5823

Bate, P., & Robert, G. (2006). Experience-based design: From redesigning the system around the patient to co-designing services with the patient. Quality and Safety in Health Care, 15(5), 307–310. https://doi.org/10.1136/qshc.2005.016527

Batterham, P. J., Calear, A. L., Sunderland, M., Kay-Lambkin, F., Farrer, L. M., Christensen, H., & Gulliver, A. (2021). A brief intervention to increase uptake and adherence of an internet-based program for depression and anxiety (enhancing engagement with psychosocial interventions): Randomized controlled trial. *Journal of Medical Internet Research*, *23*(7), e23029. https://doi.org/10.2196/23029

Beatty, L., & Binnion, C. (2016). A systematic review of predictors of, and reasons for, adherence to online psychological interventions. *International Journal of Behavioral Medicine*, *23*(6), 776–794. https://doi.org/10.1007/s12529-016-9556-9

Beck, J. G., Shipherd, J. C., & Zebb, B. J. (1997). How does interoceptive exposure for panic disorder work? An uncontrolled case study. *Journal of Anxiety Disorders*, *11*(5), 541–556. https://doi.org/10.1016/s0887-6185(97)00030-3

Behar, E., DiMarco, I. D., Hekler, E. B., Mohlman, J., & Staples, A. M. (2009). Current theoretical models of generalized anxiety disorder (GAD): conceptual review and treatment implications. *Journal of Anxiety Disorders*, *23*(8), 1011–1023. https://doi.org/10.1016/j.janxdis.2009.07.006

Bhui, K. S., Dinos, S., Stansfeld, S. A., & White, P. D. (2012). A synthesis of the evidence for managing stress at work: a review of the reviews reporting on anxiety, depression, and absenteeism. *Journal of Environmental and Public Health*, *2012*, 515874. https://doi.org/10.1155/2012/515874

Bibi, A., Lin, M., Zhang, X. C., & Margraf, J. (2020). Psychometric properties and measurement invariance of Depression, Anxiety and Stress Scales (DASS-21) across cultures: MEASUREMENT INVARIANCE OF DASS-21. *International Journal of Psychology: Journal International de Psychologie*, *55*(6), 916–925. https://doi.org/10.1002/ijop.12671

Bird, G., & Cook, R. (2013). Mixed emotions: the contribution of alexithymia to the emotional

symptoms of autism. *Translational Psychiatry*, *3*(7), e285. https://doi.org/10.1038/tp.2013.61

Blanck, P., Perleth, S., Heidenreich, T., Kröger, P., Ditzen, B., Bents, H., & Mander, J. (2018). Effects of mindfulness exercises as stand-alone intervention on symptoms of anxiety and depression: Systematic review and meta-analysis. *Behaviour Research and Therapy*, *102*, 25–35. https://doi.org/10.1016/j.brat.2017.12.002

Blanco, C., Okuda, M., Wright, C., Hasin, D. S., Grant, B. F., Liu, S.-M., & Olfson, M. (2008). Mental health of college students and their non-college-attending peers: results from the National Epidemiologic Study on Alcohol and Related Conditions. *Archives of General Psychiatry*, *65*(12), 1429–1437. https://doi.org/10.1001/archpsyc.65.12.1429

Boettcher, H., Brake, C. A., & Barlow, D. H. (2016). Origins and outlook of interoceptive exposure. *Journal of Behavior Therapy and Experimental Psychiatry*, *53*, 41–51. https://doi.org/10.1016/j.jbtep.2015.10.009

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., Waltz, T., & Zettle, R. D. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire-II: a revised measure of psychological inflexibility and experiential avoidance. *Behavior Therapy*, *42*(4), 676–688. https://doi.org/10.1016/j.beth.2011.03.007

Borgogna, N. C., McDermott, R. C., Berry, A., Lathan, E. C., & Gonzales, J. (2020). A multicultural examination of experiential avoidance: AAQ – II measurement comparisons across Asian American, Black, Latinx, Middle Eastern, and White college students. *Journal of Contextual Behavioral Science*, *16*, 1–8. https://doi.org/10.1016/j.jcbs.2020.01.011

Bornemann, B., Herbert, B. M., Mehling, W. E., & Singer, T. (2014). Differential changes in self-reported aspects of interoceptive awareness through 3 months of contemplative training. *Frontiers in Psychology*, *5*, 1504. https://doi.org/10.3389/fpsyg.2014.01504

Bostock, S., Crosswell, A. D., Prather, A. A., & Steptoe, A. (2019). Mindfulness on-the-go: Effects of a mindfulness meditation app on work stress and well-being. *Journal of Occupational Health Psychology*, *24*(1), 127–138. https://doi.org/10.1037/ocp0000118

Bradbury, K., Steele, M., Corbett, T., Geraghty, A. W. A., Krusche, A., Heber, E., Easton, S., Cheetham-Blake, T., Slodkowska-Barabasz, J., Müller, A. M., Smith, K., Wilde, L. J., Payne, L.,

Singh, K., Bacon, R., Burford, T., Summers, K., Turner, L., Richardson, A., … Yardley, L. (2019). Developing a digital intervention for cancer survivors: an evidence-, theory- and person-based approach. *Npj Digital Medicine*, *2*(1), 85. https://doi.org/10.1038/s41746-019-0163-4

Brener, J., & Kluvitse, C. (1988). Heartbeat detection: judgments of the simultaneity of external stimuli and heartbeats. *Psychophysiology*, *25*(5), 554–561. https://doi.org/10.1111/j.1469-8986.1988.tb01891.x

Brener, J., Liu, X., & Ring, C. (1993). A method of constant stimuli for examining heartbeat detection: comparison with the Brener-Kluvitse and Whitehead methods. *Psychophysiology*, *30*(6), 657–665. https://doi.org/10.1111/j.1469-8986.1993.tb02091.x

Brener, J., & Ring, C. (2016). Towards a psychophysics of interoceptive processes: the measurement of heartbeat detection. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1708). https://doi.org/10.1098/rstb.2016.0015

Brener, J., Ring, C., & Liu, X. (1994). Effects of data limitations on heartbeat detection in the method of constant stimuli. *Psychophysiology*, *31*(3), 309–312. https://doi.org/10.1111/j.1469-8986.1994.tb02219.x

Brewer, R., Cook, R., & Bird, G. (2016). Alexithymia: a general deficit of interoception. *Royal Society Open Science*, *3*(10), 150664. https://doi.org/10.1098/rsos.150664

Brewer, R., Murphy, J., & Bird, G. (2021). Atypical interoception as a common risk factor for psychopathology: A review. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2021.07.036

Brown, M., Glendenning, A., Hoon, A. E., & John, A. (2016). Effectiveness of Web-delivered acceptance and commitment therapy in relation to mental health and well-being: A systematic review and meta-analysis. *Journal of Medical Internet Research*, *18*(8), e221. https://doi.org/10.2196/jmir.6200

Brown, P. (2016). The Higher Education Policy Institute. In *The Invisible Problem? Improving Students' Mental Health*.

Buntrock, C. (2024). Cost-effectiveness of digital interventions for mental health: current evidence, common misconceptions, and future directions. *Frontiers in Digital Health*, *6*, 1486728.

https://doi.org/10.3389/fdgth.2024.1486728

Cabrera, A., Kolacz, J., Pailhez, G., Bulbena-Cabre, A., Bulbena, A., & Porges, S. W. (2018). Assessing body awareness and autonomic reactivity: Factor structure and psychometric properties of the Body Perception Questionnaire-Short Form (BPQ-SF). *International Journal of Methods in Psychiatric Research*, *27*(2), e1596. https://doi.org/10.1002/mpr.1596

Cameron, O. G. (2001). Interoception: the inside story--a model for psychosomatic processes. *Psychosomatic Medicine*, *63*(5), 697–710. https://doi.org/10.1097/00006842-200109000-00001

Cardwell, J. M., Lewis, E. G., Smith, K. C., Holt, E. R., Baillie, S., Allister, R., & Adams, V. J. (2013). A cross-sectional study of mental health in UK veterinary undergraduates. *The Veterinary Record*, *173*(11), 266. https://doi.org/10.1136/vr.101390

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., & Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cognitive Behaviour Therapy*, *47*(1), 1–18. https://doi.org/10.1080/16506073.2017.1401115

Carolan, S., & de Visser, R. O. (2018). Employees' perspectives on the facilitators and barriers to engaging with digital mental health interventions in the workplace: Qualitative study. *JMIR Mental Health*, *5*(1), e8. https://doi.org/10.2196/mental.9146

Carpenter, J. R., Roger, J. H., & Kenward, M. G. (2013). Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, *23*(6), 1352–1371. https://doi.org/10.1080/10543406.2013.834911

Carroll, D., & Whellock, J. (1980). Heart rate perception and the voluntary control of heart rate. *Biological Psychology*, *11*(3-4), 169–180. https://doi.org/10.1016/0301-0511(80)90053-8

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: One General Psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Ceunen, E., Van Diest, I., & Vlaeyen, J. W. S. (2013). Accuracy and awareness of perception: related,

yet distinct (commentary on Herbert et al., 2012). *Biological Psychology*, *92*(2), 426–427. https://doi.org/10.1016/j.biopsycho.2012.09.012

Ceunen, E., Vlaeyen, J. W. S., & Van Diest, I. (2016). On the origin of interoception. *Frontiers in Psychology*, *7*, 743. https://doi.org/10.3389/fpsyg.2016.00743

Chandola, T., Brunner, E., & Marmot, M. (2006). Chronic stress at work and the metabolic syndrome: prospective study. *BMJ (Clinical Research Ed.)*, *332*(7540), 521–525. https://doi.org/10.1136/bmj.38693.435301.80

Chou, T.-L., Chang, L.-I., & Chung, M.-H. (2015). The mediating and moderating effects of sleep hygiene practice on anxiety and insomnia in hospital nurses. *International Journal of Nursing Practice*, *21 Suppl 2*, 9–18. https://doi.org/10.1111/ijn.12164

Clemens, W. J. (1984). Temporal arrangement of signals in heartbeat discrimination procedures. *Psychophysiology*, *21*(2), 187–190. https://doi.org/10.1111/j.1469-8986.1984.tb00202.x

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Coll, M.-P., Viding, E., Rütgen, M., Silani, G., Lamm, C., Catmur, C., & Bird, G. (2017). Are we really measuring empathy? Proposal for a new measurement framework. *Neuroscience and Biobehavioral Reviews*, *83*, 132–139. https://doi.org/10.1016/j.neubiorev.2017.10.009

Cook, R., Brewer, R., Shah, P., & Bird, G. (2013). Alexithymia, not autism, predicts poor recognition of emotional facial expressions. *Psychological Science*, *24*(5), 723–732. https://doi.org/10.1177/0956797612463582

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews. Neuroscience*, *3*(8), 655–666. https://doi.org/10.1038/nrn894

Craske, M. G., Meadows, E. A., & Barlow, D. H. (1994). *Therapist's guide for the Mastery of Your Anxiety and Panic II (MAP II) program and agoraphobia supplement*. https://scholar.google.com/citations?user=RwaSybEAAAAJ&hl=en&oi=sra

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: an inhibitory learning approach. *Behaviour Research and Therapy*, *58*, 10–23. https://doi.org/10.1016/j.brat.2014.04.006

Critchley, H. D., & Harrison, N. A. (2013). Visceral influences on brain and behavior. *Neuron*, *77*(4), 624–638. https://doi.org/10.1016/j.neuron.2013.02.008

Critchley, H. D., & Nagai, Y. (2012). How emotions are shaped by bodily states. *Emotion Review: Journal of the International Society for Research on Emotion*, *4*(2), 163–168. https://doi.org/10.1177/1754073911430132

Cropley, M., Plans, D., Morelli, D., Sütterlin, S., Inceoglu, I., Thomas, G., & Chu, C. (2017). The association between work-related rumination and heart rate variability: A field study. *Frontiers in Human Neuroscience*, *11*, 27. https://doi.org/10.3389/fnhum.2017.00027

Cuijpers, P., Beekman, A. T. F., & Reynolds, C. F., 3rd. (2012). Preventing depression: a global priority. *JAMA: The Journal of the American Medical Association*, *307*(10), 1033–1034. https://doi.org/10.1001/jama.2012.271

Cuijpers, P., Donker, T., van Straten, A., Li, J., & Andersson, G. (2010). Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. *Psychological Medicine*, *40*(12), 1943–1957. https://doi.org/10.1017/S0033291710000772

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*, *11*(1), 126. https://doi.org/10.1186/1741-7015-11-126

Czyz, E. K., Horwitz, A. G., Eisenberg, D., Kramer, A., & King, C. A. (2013). Self-reported barriers to professional help seeking among college students at elevated risk for suicide. *Journal of American College Health: J of ACH*, *61*(7), 398–406. https://doi.org/10.1080/07448481.2013.820731

Dahlin, M., Andersson, G., Magnusson, K., Johansson, T., Sjögren, J., Håkansson, A., Pettersson, M., Kadowaki, Å., Cuijpers, P., & Carlbring, P. (2016). Internet-delivered acceptance-based behaviour therapy for generalized anxiety disorder: A randomized controlled trial. *Behaviour Research and Therapy*, *77*, 86–95. https://doi.org/10.1016/j.brat.2015.12.007

Dale, A., & Anderson, D. (1978). Information variables in voluntary control and classical conditioning of heart rate: field dependence and heart-rate perception. *Perceptual and Motor Skills*, *47*(1), 79–85. https://doi.org/10.2466/pms.1978.47.1.79

Damasio, A. (2008). *Descartes' error*. Vintage Digital.

Daubenmier, J., Sze, J., Kerr, C. E., Kemeny, M. E., & Mehling, W. (2013). Follow your breath: respiratory interoceptive accuracy in experienced meditators: Respiratory interoceptive accuracy in meditators. *Psychophysiology*, *50*(8), 777–789. https://doi.org/10.1111/psyp.12057

Deady, M., Johnston, D., Milne, D., Glozier, N., Peters, D., Calvo, R., & Harvey, S. (2018). Preliminary effectiveness of a smartphone app to reduce depressive symptoms in the workplace: Feasibility and acceptability study. *JMIR mHealth and uHealth*, *6*(12), e11661. https://doi.org/10.2196/11661

Dear, B. F., Staples, L. G., Terides, M. D., Karin, E., Zou, J., Johnston, L., Gandy, M., Fogliati, V. J., Wootton, B. M., McEvoy, P. M., & Titov, N. (2015). Transdiagnostic versus disorder-specific and clinician-guided versus self-guided internet-delivered treatment for generalized anxiety disorder and comorbid disorders: A randomized controlled trial. *Journal of Anxiety Disorders*, *36*, 63–77. https://doi.org/10.1016/j.janxdis.2015.09.003

De Ridder, B., Van Rompaey, B., Kampen, J. K., Haine, S., & Dilles, T. (2018). Smartphone apps using photoplethysmography for heart rate monitoring: Meta-analysis. *JMIR Cardio*, *2*(1), e4. https://doi.org/10.2196/cardio.8802

Desmedt, O., Corneille, O., Luminet, O., Murphy, J., Bird, G., & Maurage, P. (2020). Contribution of time estimation and knowledge to Heartbeat Counting Task performance under original and adapted instructions. *Biological Psychology*, *154*(107904), 107904. https://doi.org/10.1016/j.biopsycho.2020.107904

Desmedt, O., Heeren, A., Corneille, O., & Luminet, O. (2022). What do measures of self-report interoception measure? Insights from a systematic review, latent factor analysis, and network approach. *Biological Psychology*, *169*(108289), 108289. https://doi.org/10.1016/j.biopsycho.2022.108289

Desmedt, O., Luminet, O., & Corneille, O. (2018). The heartbeat counting task largely involves non-interoceptive processes: Evidence from both the original and an adapted counting task. *Biological Psychology*, *138*, 185–188. https://doi.org/10.1016/j.biopsycho.2018.09.004

Desmedt, O., Luminet, O., Maurage, P., & Corneille, O. (2025). Discrepancies in the definition and

measurement of human interoception: A comprehensive discussion and suggested ways forward. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *20*(1), 76–98. https://doi.org/10.1177/17456916231191537

Desmedt, O., Luminet, O., Walentynowicz, M., & Corneille, O. (2023). The new measures of interoceptive accuracy: A systematic review and assessment. *Neuroscience and Biobehavioral Reviews*, *153*(105388), 105388. https://doi.org/10.1016/j.neubiorev.2023.105388

Desmedt, O., Van Den Houte, M., Walentynowicz, M., Dekeyser, S., Luminet, O., & Corneille, O. (2020). *A systematic review and meta-analysis on the association between Heartbeat Counting Task performance and mental disorders and their risk factors among adults*. https://doi.org/10.31219/osf.io/h3by9

Diel, A., Schröter, I. C., Frewer, A.-L., Jansen, C., Robitzsch, A., Gradl-Dietsch, G., Teufel, M., & Bäuerle, A. (2024). A systematic review and meta analysis on digital mental health interventions in inpatient settings. *Npj Digital Medicine*, *7*(1), 253. https://doi.org/10.1038/s41746-024-01252-z

Domhardt, M., Geßlein, H., von Rezori, R. E., & Baumeister, H. (2019). Internet- and mobile-based interventions for anxiety disorders: A meta-analytic review of intervention components. *Depression and Anxiety*, *36*(3), 213–224. https://doi.org/10.1002/da.22860

Donker, T., Batterham, P. J., Warmerdam, L., Bennett, K., Bennett, A., Cuijpers, P., Griffiths, K. M., & Christensen, H. (2013). Predictors and moderators of response to internet-delivered Interpersonal Psychotherapy and Cognitive Behavior Therapy for depression. *Journal of Affective Disorders*, *151*(1), 343–351. https://doi.org/10.1016/j.jad.2013.06.020

Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M.-R., & Christensen, H. (2013). Smartphones for smarter delivery of mental health programs: a systematic review. *Journal of Medical Internet Research*, *15*(11), e247. https://doi.org/10.2196/jmir.2791

Donkin, L., Christensen, H., Naismith, S. L., Neal, B., Hickie, I. B., & Glozier, N. (2011). A systematic review of the impact of adherence on the effectiveness of e-therapies. *Journal of Medical Internet Research*, *13*(3), e52. https://doi.org/10.2196/jmir.1772

Donetto, S., Pierri, P., Tsianakas, V., & Robert, G. (2015). Experience-based co-design and healthcare

improvement: Realising participatory design in the public sector. *The Design Journal*, *18*(2), 227–248. https://doi.org/10.2752/175630615X14212498964312

Dunn, B. D., Dalgleish, T., Ogilvie, A. D., & Lawrence, A. D. (2007). Heartbeat perception in depression. *Behaviour Research and Therapy*, *45*(8), 1921–1930. https://doi.org/10.1016/j.brat.2006.09.008

Dunn, B. D., Stefanovitch, I., Evans, D., Oliver, C., Hawkins, A., & Dalgleish, T. (2010). Can you feel the beat? Interoceptive awareness is an interactive function of anxiety- and depression-specific symptom dimensions. *Behaviour Research and Therapy*, *48*(11), 1133–1138. https://doi.org/10.1016/j.brat.2010.07.006

Ebert, D. D., Cuijpers, P., Muñoz, R. F., & Baumeister, H. (2017). Prevention of mental health disorders using Internet- and mobile-based interventions: A narrative review and recommendations for future research. *Frontiers in Psychiatry*, *8*, 116. https://doi.org/10.3389/fpsyt.2017.00116

Ebert, D. D., Heber, E., Berking, M., Riper, H., Cuijpers, P., Funk, B., & Lehr, D. (2016). Self-guided internet-based and mobile-based stress management for employees: results of a randomised controlled trial. *Occupational and Environmental Medicine*, *73*(5), 315–323. https://doi.org/10.1136/oemed-2015-103269

Eggart, M., Lange, A., Binser, M. J., Queri, S., & Müller-Oerlinghausen, B. (2019). Major Depressive Disorder Is Associated with Impaired Interoceptive Accuracy: A Systematic Review. *Brain Sciences*, *9*(6). https://doi.org/10.3390/brainsci9060131

Eisenberg, D., Downs, M. F., Golberstein, E., & Zivin, K. (2009). Stigma and help seeking for mental health among college students. *Medical Care Research and Review: MCRR*, *66*(5), 522–541. https://doi.org/10.1177/1077558709335173

Eisenberg, D., Hunt, J., Speer, N., & Zivin, K. (2011). Mental health service utilization among college students in the United States. *The Journal of Nervous and Mental Disease*, *199*(5), 301–308. https://doi.org/10.1097/NMD.0b013e3182175123

England, N. H. S. (2020). *NHS Performance Statistics*. https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2020/01/Combined-Performance-Summary-January-November-

December-data-2020-c7d3g.pdf

Ernst, J., Northoff, G., Böker, H., Seifritz, E., & Grimm, S. (2013). Interoceptive awareness enhances neural activity during empathy. *Human Brain Mapping*, *34*(7), 1615–1624. https://doi.org/10.1002/hbm.22014

Eysenbach, G. (2005). The law of attrition. *Journal of Medical Internet Research*, *7*(1), e11. https://doi.org/10.2196/jmir.7.1.e11

Feldman, J. M., Kutner, H., Matte, L., Lupkin, M., Steinberg, D., Sidora-Arcoleo, K., Serebrisky, D., & Warman, K. (2012). Prediction of peak flow values followed by feedback improves perception of lung function and adherence to inhaled corticosteroids in children with asthma. *Thorax*, *67*(12), 1040–1045. https://doi.org/10.1136/thoraxjnl-2012-201789

Feldman, M. J., Bliss-Moreau, E., & Lindquist, K. A. (2024). The neurobiology of interoception and affect. *Trends in Cognitive Sciences*, *28*(7), 643–661. https://doi.org/10.1016/j.tics.2024.01.009

Ferentzi, E., Bogdány, T., Szabolcs, Z., Csala, B., Horváth, Á., & Köteles, F. (2018). Multichannel investigation of interoception: Sensitivity is not a generalizable feature. *Frontiers in Human Neuroscience*, *12*, 223. https://doi.org/10.3389/fnhum.2018.00223

Field, A. (2024). *Discovering statistics using IBM SPSS statistics* (6th ed.). SAGE Publications.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2), e19. https://doi.org/10.2196/mental.7785

Fleming, T., Bavin, L., Lucassen, M., Stasiak, K., Hopkins, S., & Merry, S. (2020). *Beyond the trial: Systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety*. https://doi.org/10.26686/wgtn.12331259.v4

Fritz, G. K., Adams, S. K., McQuaid, E. L., Klein, R., Kopel, S., Nassau, J., & Mansell, A. (2007). Symptom perception in pediatric asthma: resistive loading and in vivo assessment compared. *Chest*, *132*(3), 884–889. https://doi.org/10.1378/chest.06-2140

Fukushima, H., Terasawa, Y., & Umeda, S. (2011). Association between interoception and empathy: evidence from heartbeat-evoked brain potential. *International Journal of Psychophysiology:*

Official Journal of the International Organization of Psychophysiology, *79*(2), 259–265. https://doi.org/10.1016/j.ijpsycho.2010.10.015

Furman, D. J., Waugh, C. E., Bhattacharjee, K., Thompson, R. J., & Gotlib, I. H. (2013). Interoceptive awareness, positive affect, and decision making in major depressive disorder. *Journal of Affective Disorders*, *151*(2), 780–785. https://doi.org/10.1016/j.jad.2013.06.044

Gabriele, E., Spooner, R., Brewer, R., & Murphy, J. (2022). Dissociations between self-reported interoceptive accuracy and attention: Evidence from the Interoceptive Attention Scale. *Biological Psychology*, *168*(108243), 108243. https://doi.org/10.1016/j.biopsycho.2021.108243

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *Coefficients of Interrater Reliability and Agreement for quantitative, ordinal and nominal data: ICC, Finn-Coefficient, Robinson's A, Kendall's W, and Cohen's Kappa*.

Gannon, L. R. (1980). Cardiac perception and the voluntary control of heart rate. *Physiological Psychology*, *8*(4), 509–514. https://doi.org/10.3758/bf03326485

Garfinkel, S. N., Manassei, M. F., Hamilton-Fletcher, G., In den Bosch, Y., Critchley, H. D., & Engels, M. (2016). Interoceptive dimensions across cardiac and respiratory axes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1708). https://doi.org/10.1098/rstb.2016.0014

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology*, *104*, 65–74. https://doi.org/10.1016/j.biopsycho.2014.11.004

Garfinkel, S. N., Tiley, C., O'Keeffe, S., Harrison, N. A., Seth, A. K., & Critchley, H. D. (2016). Discrepancies between dimensions of interoception in autism: Implications for emotion and anxiety. *Biological Psychology*, *114*, 117–126. https://doi.org/10.1016/j.biopsycho.2015.12.003

Gendron, M., & Barrett, L. F. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion Review: Journal of the International Society for Research on Emotion*, *1*(4), 316–339. https://doi.org/10.1177/1754073909338877

Goetter, E. M., Frumkin, M. R., Palitz, S. A., Swee, M. B., Baker, A. W., Bui, E., & Simon, N. M. (2020). Barriers to mental health treatment among individuals with social anxiety disorder and

generalized anxiety disorder. *Psychological Services*, *17*(1), 5–12. https://doi.org/10.1037/ser0000254

Griez, E., & van den Hout, M. A. (1983). Treatment of phobophobia by exposure to CO2-induced anxiety symptoms. *The Journal of Nervous and Mental Disease*, *171*(8), 506–508. https://doi.org/10.1097/00005053-198308000-00009

Gruwez, H., Ezzat, D., Van Puyvelde, T., Dhont, S., Meekers, E., Bruckers, L., Wouters, F., Kellens, M., Van Herendael, H., Rivero-Ayerza, M., Nuyens, D., Haemers, P., & Pison, L. (2024). Real-world validation of smartphone-based photoplethysmography for rate and rhythm monitoring in atrial fibrillation. *Europace: European Pacing, Arrhythmias, and Cardiac Electrophysiology: Journal of the Working Groups on Cardiac Pacing, Arrhythmias, and Cardiac Cellular Electrophysiology of the European Society of Cardiology*, *26*(4), euae065. https://doi.org/10.1093/europace/euae065

Hammer, L. B., & Sauter, S. (2013). Total worker health and work-life stress. *Journal of Occupational and Environmental Medicine*, *55*(12 Suppl), S25–S29. https://doi.org/10.1097/JOM.0000000000000043

Harshaw, C. (2015). Interoceptive dysfunction: toward an integrated framework for understanding somatic and affective disturbance in depression. *Psychological Bulletin*, *141*(2), 311–363. https://doi.org/10.1037/a0038101

Harver, A. (1994). Effects of feedback on the ability of asthmatic subjects to detect increases in the flow-resistive component to breathing. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *13*(1), 52–62. https://doi.org/10.1037//0278-6133.13.1.52

Haslam, C., Jetten, J., Cruwys, T., Dingle, G., & Haslam, S. A. (2018). *The new psychology of health*. Routledge.

Häusser, J. A., Mojzisch, A., Niesel, M., & Schulz-Hardt, S. (2010). Ten years on: A review of recent research on the Job Demand–Control (-Support) model and psychological well-being. *Work and Stress*, *24*(1), 1–35. https://doi.org/10.1080/02678371003683747

Hayes, S. C. (2016). Acceptance and Commitment Therapy, Relational Frame Theory, and the third

wave of behavioral and cognitive therapies - republished article. *Behavior Therapy*, *47*(6), 869–885. https://doi.org/10.1016/j.beth.2016.11.006

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: model, processes and outcomes. *Behaviour Research and Therapy*, *44*(1), 1–25. https://doi.org/10.1016/j.brat.2005.06.006

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (2016). *Acceptance and commitment therapy, second edition* (2nd ed.). Guilford Publications.

Heaton, P., Reichenbacher, L., Sauter, D., Allen, R., Scott, S., & Hill, E. (2012). Measuring the effects of alexithymia on perception of emotional vocalizations in autistic spectrum disorder and typical development. *Psychological Medicine*, *42*(11), 2453–2459. https://doi.org/10.1017/S0033291712000621

Heber, E., Ebert, D. D., Lehr, D., Nobis, S., Berking, M., & Riper, H. (2013). Efficacy and cost-effectiveness of a web-based and mobile stress-management intervention for employees: design of a randomized controlled trial. *BMC Public Health*, *13*(1), 655. https://doi.org/10.1186/1471-2458-13-655

Heber, E., Lehr, D., Ebert, D. D., Berking, M., & Riper, H. (2016). Web-based and mobile stress management intervention for employees: A randomized controlled trial. *Journal of Medical Internet Research*, *18*(1), e21. https://doi.org/10.2196/jmir.5112

Hedman, E., Ljótsson, B., & Lindefors, N. (2012). Cognitive behavior therapy via the Internet: a systematic review of applications, clinical efficacy and cost-effectiveness. *Expert Review of Pharmacoeconomics & Outcomes Research*, *12*(6), 745–764. https://doi.org/10.1586/erp.12.67

Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. *The British Journal of Clinical Psychology*, *44*(Pt 2), 227–239. https://doi.org/10.1348/014466505X29657

Herbert, B. M., Herbert, C., & Pollatos, O. (2011). On the relationship between interoceptive awareness and alexithymia: is interoceptive awareness related to emotional awareness? *Journal of Personality*, *79*(5), 1149–1175. https://doi.org/10.1111/j.1467-6494.2011.00717.x

Herbert, B. M., Muth, E. R., Pollatos, O., & Herbert, C. (2012). Interoception across modalities: on the relationship between cardiac awareness and the sensitivity for gastric functions. *PloS One*, *7*(5), e36646. https://doi.org/10.1371/journal.pone.0036646

Herman, A. M., Rae, C. L., Critchley, H. D., & Duka, T. (2019). Interoceptive accuracy predicts nonplanning trait impulsivity. *Psychophysiology*, *56*(6), e13339. https://doi.org/10.1111/psyp.13339

Herrman, H., Saxena, S., & Moodie, R. (2004). *Promoting Mental Health: Concepts, Emerging Evidence, Practice. Geneva: World Health Organisation*.

Hesselmann, G. (2018). Applying linear mixed effects models (LMMs) in within-participant designs with subjective trial-based assessments of awareness—a caveat. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00788

He, Y., & Li, Y. (2013). Physical activity recognition utilizing the built-in kinematic sensors of a smartphone. *International Journal of Distributed Sensor Networks*, *9*(4), 481580. https://doi.org/10.1155/2013/481580

Hickman, L., Seyedsalehi, A., Cook, J. L., Bird, G., & Murphy, J. (2020). The relationship between heartbeat counting and heartbeat discrimination: A meta-analysis. *Biological Psychology*, *156*, 107949. https://doi.org/10.1016/j.biopsycho.2020.107949

Hoffmann, D., Rask, C. U., Hedman-Lagerlöf, E., Ljótsson, B., & Frostholm, L. (2018). Development and feasibility testing of internet-delivered acceptance and commitment therapy for severe health anxiety: Pilot study. *JMIR Mental Health*, *5*(2), e28. https://doi.org/10.2196/mental.9198

Hopewell, S., Chan, A.-W., Collins, G. S., Hróbjartsson, A., Moher, D., Schulz, K. F., Tunn, R., Aggarwal, R., Berkwits, M., Berlin, J. A., Bhandari, N., Butcher, N. J., Campbell, M. K., Chidebe, R. C. W., Elbourne, D., Farmer, A., Fergusson, D. A., Golub, R. M., Goodman, S. N., … Boutron, I. (2025). CONSORT 2025 statement: Updated guideline for reporting randomised trials. *BMJ*, *389*, e081123. https://doi.org/10.1136/bmj-2024-081123

Höller, I., Stenzel, J.-S., Rath, D., & Forkmann, T. (2021). Listen to your heart-ecological momentary assessment of interoceptive accuracy, awareness and sensibility: A pilot study. *International Journal of Environmental Research and Public Health*, *18*(9), 4893.

https://doi.org/10.3390/ijerph18094893

Holt, M. K., & Trapp, S. (2016). The physiological role of the brain GLP-1 system in stress. *Cogent Biology*, *2*(1), 1229086. https://doi.org/10.1080/23312025.2016.1229086

Howarth, A., Quesada, J., Donnelly, T., & Mills, P. R. (2019). The development of "Make One Small Change": an e-health intervention for the workplace developed using the Person-Based Approach. *Digital Health*, *5*, 2055207619852856. https://doi.org/10.1177/2055207619852856

Howarth, A., Quesada, J., Silva, J., Judycki, S., & Mills, P. R. (2018). The impact of digital health interventions on health-related outcomes in the workplace: A systematic review. *Digital Health*, *4*, 2055207618770861. https://doi.org/10.1177/2055207618770861

Huberty, J., Green, J., Glissmann, C., Larkey, L., Puzia, M., & Lee, C. (n.d.). *Efficacy of the mindfulness meditation mobile app "Calm" to reduce stress among college students: randomized controlled trial*.

Huberty, J., Green, J., Glissmann, C., Larkey, L., Puzia, M., & Lee, C. (2019). Efficacy of the mindfulness meditation mobile app "Calm" to reduce stress among college students: Randomized controlled trial. *JMIR mHealth and uHealth*, *7*(6), e14273. https://doi.org/10.2196/14273

Huh, K. Y., Oh, J., Lee, S., & Yu, K.-S. (2022). Clinical evaluation of digital therapeutics: Present and future. *Healthcare Informatics Research*, *28*(3), 188–197. https://doi.org/10.4258/hir.2022.28.3.188

Ivanova, E., Lindner, P., Ly, K. H., Dahlin, M., Vernmark, K., Andersson, G., & Carlbring, P. (2016). Guided and unguided Acceptance and Commitment Therapy for social anxiety disorder and/or panic disorder provided via the Internet and a smartphone application: A randomized controlled trial. *Journal of Anxiety Disorders*, *44*, 27–35. https://doi.org/10.1016/j.janxdis.2016.09.012

James, W. (1894). Discussion: The physical basis of emotion. *Psychological Review*, *1*(5), 516–529. https://doi.org/10.1037/h0065078

Janssens, T., & Harver, A. (2015). Effects of symptom perception interventions on trigger identification and quality of life in children with asthma. *Pulmonary Medicine*, *2015*, 825137. https://doi.org/10.1155/2015/825137

Jones, G. E., & Hollandsworth, J. G. (1981). Heart rate discrimination before and after exercise-

induced augmented cardiac activity. *Psychophysiology*, *18*(3), 252–257. https://doi.org/10.1111/j.1469-8986.1981.tb03029.x

Jones, L. A., & Brierley, D. I. (2025). GLP-1 and the neurobiology of eating control: Recent advances. *Endocrinology*, *166*(2), bqae167. https://doi.org/10.1210/endocr/bqae167

Joyce, S., Modini, M., Christensen, H., Mykletun, A., Bryant, R., Mitchell, P. B., & Harvey, S. B. (2016). Workplace interventions for common mental disorders: a systematic meta-review. *Psychological Medicine*, *46*(4), 683–697. https://doi.org/10.1017/S0033291715002408

Kabat-Zinn, J. (1982). An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: theoretical considerations and preliminary results. *General Hospital Psychiatry*, *4*(1), 33–47. https://doi.org/10.1016/0163-8343(82)90026-3

Kählke, F., Berger, T., Schulz, A., Baumeister, H., Berking, M., Cuijpers, P., Bruffaerts, R., Auerbach, R. P., Kessler, R. C., & Ebert, D. D. (2019). Efficacy and cost-effectiveness of an unguided, internet-based self-help intervention for social anxiety disorder in university students: protocol of a randomized controlled trial. *BMC Psychiatry*, *19*(1), 197. https://doi.org/10.1186/s12888-019-2125-4

Karasek, R. A., join(’ '. (1979). Job demands, job decision latitude, and mental strain: Implications for job redesign. *Administrative Science Quarterly*, *24*(2), 285. https://doi.org/10.2307/2392498

Karekla, M., Constantinou, P., Ioannou, M., Gloster, A. T., & Kareklas, I. (2019). The phenomenon of treatment dropout, reasons and moderators in Acceptance and Commitment Therapy and other active treatments: A meta-analytic review. *Clinical Psychology in Europe*, *1*(3). https://doi.org/10.32872/cpe.v1i3.33058

Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review*, *30*(7), 865–878. https://doi.org/10.1016/j.cpr.2010.03.001

Kawadler, J. M., Hemmings, N. R., Ponzo, S., Morelli, D., Bird, G., & Plans, D. (2020). *Effectiveness of a smart phone application BioBase for reducing anxiety and increasing mental wellbeing: a pilot feasibility and acceptability study*.

Kelders, S. M., Kok, R. N., Ossebaard, H. C., & Van Gemert-Pijnen, J. E. W. C. (2012). Persuasive system design does matter: a systematic review of adherence to web-based interventions. *Journal*

of Medical Internet Research, *14*(6), e152. https://doi.org/10.2196/jmir.2104

Kelson, J. N., Lam, M. K., Keep, M., & Campbell, A. J. (2017). Development and evaluation of an online acceptance and commitment therapy program for anxiety: Phase I iterative design. *Journal of Technology in Human Services*, *35*(2), 135–151. https://doi.org/10.1080/15228835.2017.1309311

Kelson, J., Rollin, A., Ridout, B., & Campbell, A. (2019). Internet-delivered acceptance and Commitment Therapy for anxiety treatment: Systematic review. *Journal of Medical Internet Research*, *21*(1), e12530. https://doi.org/10.2196/12530

Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 617–627. https://doi.org/10.1001/archpsyc.62.6.617

Khalsa, S. S., Adolphs, R., Cameron, O. G., Critchley, H. D., Davenport, P. W., Feinstein, J. S., Feusner, J. D., Garfinkel, S. N., Lane, R. D., Mehling, W. E., Meuret, A. E., Nemeroff, C. B., Oppenheimer, S., Petzschner, F. H., Pollatos, O., Rhudy, J. L., Schramm, L. P., Simmons, W. K., Stein, M. B., … Interoception Summit 2016 participants. (2018). Interoception and Mental Health: A Roadmap. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *3*(6), 501–513. https://doi.org/10.1016/j.bpsc.2017.12.004

Khalsa, S. S., & Lapidus, R. C. (2016). Can Interoception Improve the Pragmatic Search for Biomarkers in Psychiatry? *Frontiers in Psychiatry / Frontiers Research Foundation*, *7*, 121. https://doi.org/10.3389/fpsyt.2016.00121

Khalsa, S. S., Rudrauf, D., Damasio, A. R., Davidson, R. J., Lutz, A., & Tranel, D. (2008). Interoceptive awareness in experienced meditators. *Psychophysiology*, *45*(4), 671–677. https://doi.org/10.1111/j.1469-8986.2008.00666.x

Khalsa, S. S., Rudrauf, D., Feinstein, J. S., & Tranel, D. (2009). The pathways of interoceptive awareness. *Nature Neuroscience*, *12*(12), 1494–1496. https://doi.org/10.1038/nn.2411

Khalsa, S. S., Rudrauf, D., Hassanpour, M. S., Davidson, R. J., & Tranel, D. (2020). The practice of meditation is not associated with improved interoceptive awareness of the heartbeat.

Psychophysiology, *57*(2), e13479. https://doi.org/10.1111/psyp.13479

Kifle, Y., Seng, V., & Davenport, P. W. (1997). Magnitude estimation of inspiratory resistive loads in children with life-threatening asthma. *American Journal of Respiratory and Critical Care Medicine*, *156*(5), 1530–1535. https://doi.org/10.1164/ajrccm.156.5.9703011

Knapp-Kline, K., & Kline, J. P. (2005). Heart rate, heart rate variability, and heartbeat detection with the method of constant stimuli: slow and steady wins the race. *Biological Psychology*, *69*(3), 387–396. https://doi.org/10.1016/j.biopsycho.2004.09.002

Knapp, K., Ring, C., & Brener, J. (1997). Sensitivity to mechanical stimuli and the role of general sensory and perceptual processes in heartbeat detection. *Psychophysiology*, *34*(4), 467–473. https://doi.org/10.1111/j.1469-8986.1997.tb02391.x

Knoll, J. F., & Hodapp, V. (1992). A comparison between two methods for assessing heartbeat perception. *Psychophysiology*, *29*(2), 218–222. https://www.ncbi.nlm.nih.gov/pubmed/1635964

Körmendi, J., Ferentzi, E., & Köteles, F. (2022). A heartbeat away from a valid tracking task. An empirical comparison of the mental and the motor tracking task. *Biological Psychology*, *171*(108328), 108328. https://doi.org/10.1016/j.biopsycho.2022.108328

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lachaux, J. P., Rodriguez, E., Martinerie, J., & Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping*, *8*(4), 194–208. https://doi.org/10.1002/(sici)1097-0193(1999)8:4<194::aid-hbm4>3.0.co;2-c

Lackner, R. J., & Fresco, D. M. (2016). Interaction effect of brooding rumination and interoceptive awareness on depression and anxiety symptoms. *Behaviour Research and Therapy*, *85*, 43–52. https://doi.org/10.1016/j.brat.2016.08.007

Larsson, D. E. O., Esposito, G., Critchley, H. D., Dienes, Z., & Garfinkel, S. N. (2021). Sensitivity to

changes in rate of heartbeats as a measure of interoceptive ability. *Journal of Neurophysiology*, *126*(5), 1799–1813. https://doi.org/10.1152/jn.00059.2021

Lattie, E. G., Adkins, E. C., Winquist, N., Stiles-Shields, C., Wafford, Q. E., & Graham, A. K. (2019). Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: Systematic review. *Journal of Medical Internet Research*, *21*(7), e12869. https://doi.org/10.2196/12869

Levin, M. E., Hildebrandt, M. J., Lillis, J., & Hayes, S. C. (2012). The impact of treatment components suggested by the psychological flexibility model: a meta-analysis of laboratory-based component studies. *Behavior Therapy*, *43*(4), 741–756. https://doi.org/10.1016/j.beth.2012.05.003

Levin, M. E., Pistorello, J., Seeley, J. R., & Hayes, S. C. (2014). Feasibility of a prototype web-based acceptance and commitment therapy prevention program for college students. *Journal of American College Health: J of ACH*, *62*(1), 20–30. https://doi.org/10.1080/07448481.2013.843533

Lewis, J. R. (2018). The system usability scale: Past, present, and future. *International Journal of Human-Computer Interaction*, *34*(7), 577–590. https://doi.org/10.1080/10447318.2018.1455307

Linardon, J., & Fuller-Tyszkiewicz, M. (2020). Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. *Journal of Consulting and Clinical Psychology*, *88*(1), 1–13. https://doi.org/10.1037/ccp0000459

Lovibond, P. F., & Lovibond, S. H. (1995). Depression Anxiety and Stress Scales. In *PsycTESTS Dataset*. https://doi.org/10.1037/t39835-000

Lovibond, S. H., & Lovibond, P. F. (1996). *Manual for the Depression Anxiety Stress Scales. Australia: Psychology Foundation of Australia*.

Lungu, A., & Sun, M. (2016). Time for a change: College students' preference for technology-mediated versus face-to-face help for emotional distress. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, *22*(12), 991–1000. https://doi.org/10.1089/tmj.2015.0214

Lutz, J., Offidani, E., Taraboanta, L., Lakhan, S. E., & Campellone, T. R. (2022). Appropriate controls

for digital therapeutic clinical trials: A narrative review of control conditions in clinical trials of digital therapeutics (DTx) deploying psychosocial, cognitive, or behavioral content. *Frontiers in Digital Health*, *4*, 823977. https://doi.org/10.3389/fdgth.2022.823977

Ly, K. H., Asplund, K., & Andersson, G. (2014). Stress management for middle managers via an acceptance and commitment-based smartphone application: A randomized controlled trial. *Internet Interventions*, *1*(3), 95–101. https://doi.org/10.1016/j.invent.2014.06.003

Mak, W. W. S., Chan, A. T. Y., Cheung, E. Y. L., Lin, C. L. Y., & Ngai, K. C. S. (2015). Enhancing Web-based mindfulness training for mental health promotion with the health action process approach: randomized controlled trial. *Journal of Medical Internet Research*, *17*(1), e8. https://doi.org/10.2196/jmir.3746

Marques, L., LeBlanc, N. J., Weingarden, H. M., Timpano, K. R., Jenike, M., & Wilhelm, S. (2010). Barriers to treatment and service utilization in an internet sample of individuals with obsessive-compulsive symptoms. *Depression and Anxiety*, *27*(5), 470–475. https://doi.org/10.1002/da.20694

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). *The British Journal of Clinical Psychology*, *31*(3), 301–306. https://doi.org/10.1111/j.2044-8260.1992.tb00997.x

Masanneck, L., & Stern, A. D. (2024). Tracing digital therapeutics research across medical specialties: Evidence from ClinicalTrials.Gov. *Clinical Pharmacology and Therapeutics*, *116*(1), 177–185. https://doi.org/10.1002/cpt.3260

Mass, R., Dahme, B., & Richter, R. (1992). Comparison of two different methods for measuring interoception of obstructions in asthma patients. *Pneumologie (Stuttgart, Germany)*, *46*(5), 183–189. https://www.ncbi.nlm.nih.gov/pubmed/1615003

Mattila, E., Lappalainen, R., Välkkynen, P., Sairanen, E., Lappalainen, P., Karhunen, L., Peuhkuri, K., Korpela, R., Kolehmainen, M., & Ermes, M. (2016). Usage and dose response of a mobile Acceptance and Commitment Therapy app: Secondary analysis of the intervention arm of a randomized controlled trial. *JMIR mHealth and uHealth*, *4*(3), e90. https://doi.org/10.2196/mhealth.5241

McFarland, R. A. (1975). Heart rate perception and heart rate control. *Psychophysiology*, *12*(4), 402–405. https://doi.org/10.1111/j.1469-8986.1975.tb00011.x

McKay, F. H., Cheng, C., Wright, A., Shill, J., Stephens, H., & Uccellini, M. (2018). Evaluating mobile phone applications for health behaviour change: A systematic review. *Journal of Telemedicine and Telecare*, *24*(1), 22–30. https://doi.org/10.1177/1357633X16673538

Melville, K. M., Casey, L. M., & Kavanagh, D. J. (2010). Dropout from Internet-based treatment for psychological disorders. *The British Journal of Clinical Psychology*, *49*(Pt 4), 455–471. https://doi.org/10.1348/014466509X472138

Morelli, D., Bartoloni, L., Colombo, M., Plans, D., & Clifton, D. A. (2018). Profiling the propagation of error from PPG to HRV features in a wearable physiological-monitoring device. *Healthcare Technology Letters*, *5*(2), 59–64. https://doi.org/10.1049/htl.2017.0039

Mowbray, C. T., Megivern, D., Mandiberg, J. M., Strauss, S., Stein, C. H., Collins, K., Kopels, S., Curlin, C., & Lett, R. (2006). Campus mental health services: recommendations for change. *The American Journal of Orthopsychiatry*, *76*(2), 226–237. https://doi.org/10.1037/0002-9432.76.2.226

Mul, C.-L., Stagg, S. D., Herbelin, B., & Aspell, J. E. (2018). The feeling of me feeling for you: Interoception, alexithymia and empathy in autism. *Journal of Autism and Developmental Disorders*, *48*(9), 2953–2967. https://doi.org/10.1007/s10803-018-3564-3

Murphy, J. (2024). Interoception: Where do we go from here? *Quarterly Journal of Experimental Psychology (2006)*, *77*(2), 223–229. https://doi.org/10.1177/17470218231172725

Murphy, J., Brewer, R., Catmur, C., & Bird, G. (2017). Interoception and psychopathology: A developmental neuroscience perspective. *Developmental Cognitive Neuroscience*, *23*, 45–56. https://doi.org/10.1016/j.dcn.2016.12.006

Murphy, J., Brewer, R., Coll, M.-P., Plans, D., Hall, M., Shiu, S. S., Catmur, C., & Bird, G. (2019). I feel it in my finger: Measurement device affects cardiac interoceptive accuracy. *Biological Psychology*, *148*, 107765.

Murphy, J., Brewer, R., Hobson, H., Catmur, C., & Bird, G. (2018). Is alexithymia characterised by impaired interoception? Further evidence, the importance of control variables, and the problems

with the Heartbeat Counting Task. *Biological Psychology*, *136*, 189–197. https://doi.org/10.1016/j.biopsycho.2018.05.010

Murphy, J., Brewer, R., Plans, D., Khalsa, S. S., Catmur, C., & Bird, G. (2020). Testing the independence of self-reported interoceptive accuracy and attention. *The Quarterly Journal of Experimental Psychology*, *73*(1), 115–133.

Murphy, J., Catmur, C., & Bird, G. (2017). Alexithymia Is Associated With a Multidomain, Multidimensional Failure of Interoception: Evidence From Novel Tests. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0000366

Murphy, J., Catmur, C., & Bird, G. (2019). Classifying individual differences in interoception: Implications for the measurement of interoceptive awareness. *Psychonomic Bulletin & Review*, *26*(5), 1467–1471. https://doi.org/10.3758/s13423-019-01632-7

Murphy, J., Millgate, E., Geary, H., Ichijo, E., Coll, M.-P., Brewer, R., Catmur, C., & Bird, G. (2018). Knowledge of resting heart rate mediates the relationship between intelligence and the heartbeat counting task. *Biological Psychology*, *133*, 1–3. https://doi.org/10.1016/j.biopsycho.2018.01.012

Mussgay, L., Klinkenberg, N., & Rüddel, H. (1999). Heart beat perception in patients with depressive, somatoform, and personality disorders. *Journal of Psychophysiology*, *13*(1), 27–36. https://doi.org/10.1027//0269-8803.13.1.27

Muuraiskangas, S., Harjumaa, M., Kaipainen, K., & Ermes, M. (2016). Process and effects evaluation of a digital mental health intervention targeted at improving occupational well-being: Lessons from an intervention study with failed adoption. *JMIR Mental Health*, *3*(2), e13. https://doi.org/10.2196/mental.4465

NEMIAH, & JC. (1976). Alexithymia : a view of the psychosomatic process. *Modern Trends in Psychosoamtic Medicine*, *3*, 430–439. http://ci.nii.ac.jp/naid/10029519287/

Newby, J. M., McKinnon, A., Kuyken, W., Gilbody, S., & Dalgleish, T. (2015). Systematic review and meta-analysis of transdiagnostic psychological treatments for anxiety and depressive disorders in adulthood. *Clinical Psychology Review*, *40*, 91–110. https://doi.org/10.1016/j.cpr.2015.06.002

NICE. (2011). *Generalised anxiety disorder and panic disorder in adults: management*. NICE. https://

www.nice.org.uk/guidance/cg113/chapter/Recommendations

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

O'Brien, W. H., Reid, G. J., & Jones, K. R. (1998). Differences in heartbeat awareness among males with higher and lower levels of systolic blood pressure. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *29*(1), 53–63. https://doi.org/10.1016/s0167-8760(98)00004-x

O'Connor, S., Hanlon, P., O'Donnell, C. A., Garcia, S., Glanville, J., & Mair, F. S. (2016). Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. *BMC Medical Informatics and Decision Making*, *16*(1), 120. https://doi.org/10.1186/s12911-016-0359-3

Ong, C. W., Lee, E. B., & Twohig, M. P. (2018). A meta-analysis of dropout rates in acceptance and commitment therapy. *Behaviour Research and Therapy*, *104*, 14–33. https://doi.org/10.1016/j.brat.2018.02.004

Palmer, R., Morelli, D., Plans, D., Todd, J., Aspell, J. E., Leganes, M., Bird, G., & Murphy, J. (n.d.). Evidence for Cardiac Phase-Linked Perception of Heartbeats. *(under Review)*.

Palmer, R., Murphy, J., Bird, J., Piercy, T., & Adams, K.. (n.d.). Refinements of the Phase Adjustment Task (PAT 2.0). *(under Review)*.

Palser, E. R., Fotopoulou, A., Pellicano, E., & Kilner, J. M. (2018). The link between interoceptive processing and anxiety in children diagnosed with autism spectrum disorder: Extending adult findings into a developmental sample. *Biological Psychology*, *136*, 13–21. https://doi.org/10.1016/j.biopsycho.2018.05.003

Pauley, D., Cuijpers, P., Papola, D., Miguel, C., & Karyotaki, E. (2023). Two decades of digital interventions for anxiety disorders: A systematic review and meta-analysis of treatment effectiveness. Psychological Medicine, 53(2), 567–579. https://doi.org/10.1017/S0033291721001999

Paulus, M. P. (2013). The breathing conundrum-interoceptive sensitivity and anxiety. *Depression and*

Anxiety, *30*(4), 315–320. https://doi.org/10.1002/da.22076

Paulus, M. P., & Stein, M. B. (2010). Interoception in anxiety and depression. *Brain Structure & Function*, *214*(5-6), 451–463. https://doi.org/10.1007/s00429-010-0258-9

Pennebaker, J. W., & Roberts, T.-A. (1992). Toward a his and hers theory of emotion: Gender differences in visceral perception. *Journal of Social and Clinical Psychology*, *11*(3), 199–212. https://doi.org/10.1521/jscp.1992.11.3.199

Perrin, A. (2021, June 3). *Mobile Technology and Home Broadband 2021*. Pew Research Center. https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021/

Phillips, G. C., Jones, G. E., Rieger, E. J., & Snell, J. B. (1999). Effects of the presentation of false heart-rate feedback on the performance of two common heartbeat-detection tasks. *Psychophysiology*, *36*(4), 504–510. https://doi.org/10.1017/s0048577299980071

Plans, D., Ponzo, S., Bird, G., Ring, C., Morelli, D., Keating, C. T., Murphy, J., & Catmur, C. (2020). *Measuring Interoception: The Phase Adjustment Task*.

Pollatos, O., Traut-Mattausch, E., & Schandry, R. (2009). Differential effects of anxiety and depression on interoceptive accuracy. *Depression and Anxiety*, *26*(2), 167–173. https://doi.org/10.1002/da.20504

Porges, S. (1993). Body perception questionnaire. *Laboratory of Developmental Assessment, University of*.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734. https://doi.org/10.1017/S0954579405050340

Prentice, F., Hobson, H., Spooner, R., & Murphy, J. (2022). Gender differences in interoceptive accuracy and emotional ability: An explanation for incompatible findings. *Neuroscience and Biobehavioral Reviews*, *141*(104808), 104808. https://doi.org/10.1016/j.neubiorev.2022.104808

Prentice, J. L., & Dobson, K. S. (2014). A review of the risks and benefits associated with mobile phone applications for psychological interventions. *Psychologie Canadienne [Canadian Psychology]*, *55*(4), 282–290. https://doi.org/10.1037/a0038113

Price, M., Yuen, E. K., Goetter, E. M., Herbert, J. D., Forman, E. M., Acierno, R., & Ruggiero, K. J. (2014). mHealth: a mechanism to deliver more accessible, more effective mental health care. *Clinical Psychology & Psychotherapy*, *21*(5), 427–436. https://doi.org/10.1002/cpp.1855

Proudfoot, J., Clarke, J., Birch, M.-R., Whitton, A. E., Parker, G., Manicavasagar, V., Harrison, V., Christensen, H., & Hadzi-Pavlovic, D. (2013). Impact of a mobile phone and web program on symptom and functional outcomes for people with mild-to-moderate depression, anxiety and stress: a randomised controlled trial. *BMC Psychiatry*, *13*(1), 312. https://doi.org/10.1186/1471-244X-13-312

Proudfoot, J., Ryden, C., Everitt, B., Shapiro, D. A., Goldberg, D., Mann, A., Tylee, A., Marks, I., & Gray, J. A. (2004). Clinical efficacy of computerised cognitive-behavioural therapy for anxiety and depression in primary care: randomised controlled trial. *The British Journal of Psychiatry: The Journal of Mental Science*, *185*, 46–54. https://doi.org/10.1192/bjp.185.1.46

Quadt, L., Garfinkel, S. N., Mulcahy, J. S., Larsson, D. E., Silva, M., Jones, A.-M., Strauss, C., & Critchley, H. D. (2021). Interoceptive training to target anxiety in autistic adults (ADIE): A single-center, superiority randomized controlled trial. *EClinicalMedicine*, *39*(101042), 101042. https://doi.org/10.1016/j.eclinm.2021.101042

Quattrocki, E., & Friston, K. (2014). Autism, oxytocin and interoception. *Neuroscience and Biobehavioral Reviews*, *47*, 410–430. https://doi.org/10.1016/j.neubiorev.2014.09.012

Randall, E. M., & Bewick, B. M. (2016). Exploration of counsellors' perceptions of the redesigned service pathways: a qualitative study of a UK university student counselling service. *British Journal of Guidance & Counselling*, *44*(1), 86–98. https://doi.org/10.1080/03069885.2015.1017801

Ring, C., & Brener, J. (2018). Heartbeat counting is unrelated to heartbeat detection: A comparison of methods to quantify interoception. *Psychophysiology*, e13084. https://doi.org/10.1111/psyp.13084

Ring, C., Brener, J., Knapp, K., & Mailloux, J. (2015). Effects of heartbeat feedback on beliefs about heart rate and heartbeat counting: a cautionary tale about interoceptive awareness. *Biological Psychology*, *104*, 193–198. https://doi.org/10.1016/j.biopsycho.2014.12.010

Rodríguez-Liñares, L., Méndez, A. J., Lado, M. J., Olivieri, D. N., Vila, X. A., & Gómez-Conde, I. (2011). An open source tool for heart rate variability spectral analysis. *Computer Methods and Programs in Biomedicine*, *103*(1), 39–50. https://doi.org/10.1016/j.cmpb.2010.05.012

Roemer, L., Salters, K., Raffa, S. D., & Orsillo, S. M. (2005). Fear and avoidance of internal experiences in GAD: Preliminary tests of a conceptual model. *Cognitive Therapy and Research*, *29*(1), 71–88. https://doi.org/10.1007/s10608-005-1650-2

Rouse, C. H., Jones, G. E., & Jones, K. R. (1988). The effect of body composition and gender on cardiac awareness. *Psychophysiology*, *25*(4), 400–407. https://doi.org/10.1111/j.1469-8986.1988.tb01876.x

Rowsell, A., Muller, I., Murray, E., Little, P., Byrne, C. D., Ganahl, K., Müller, G., Gibney, S., Lyles, C. R., Lucas, A., Nutbeam, D., & Yardley, L. (2015). Views of people with high and low levels of health literacy about a digital intervention to promote physical activity for diabetes: A qualitative study in five countries. *Journal of Medical Internet Research*, *17*(10), e230. https://doi.org/10.2196/jmir.4999

Ruscio, A. M., Hallion, L. S., Lim, C. C. W., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Andrade, L. H., Borges, G., Bromet, E. J., Bunting, B., Caldas de Almeida, J. M., Demyttenaere, K., Florescu, S., de Girolamo, G., Gureje, O., Haro, J. M., He, Y., Hinkov, H., Hu, C., … Scott, K. M. (2017). Cross-sectional comparison of the epidemiology of DSM-5 generalized anxiety disorder across the globe. *JAMA Psychiatry (Chicago, Ill.)*, *74*(5), 465–475. https://doi.org/10.1001/jamapsychiatry.2017.0056

Ruwaard, J., Lange, A., Bouwman, M., Broeksteeg, J., & Schrieken, B. (2007). E-mailed standardized cognitive behavioural treatment of work-related stress: a randomized controlled trial. *Cognitive Behaviour Therapy*, *36*(3), 179–192. https://doi.org/10.1080/16506070701381863

Schaan, V. K., Schulz, A., Rubel, J. A., Bernstein, M., Domes, G., Schächinger, H., & Vögele, C. (2019). Childhood trauma affects stress-related interoceptive accuracy. *Frontiers in Psychiatry*, *10*, 750. https://doi.org/10.3389/fpsyt.2019.00750

Schachter, S., & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399. https://doi.org/10.1037/h0046234

Schandry, R. (1981). Heart beat perception and emotional experience. *Psychophysiology*, *18*(4), 483–488. https://doi.org/10.1111/j.1469-8986.1981.tb02486.x

Schandry, R., Bestler, M., & Montoya, P. (1993). On the relation between cardiodynamics and heartbeat perception. *Psychophysiology*, *30*(5), 467–474. https://doi.org/10.1111/j.1469-8986.1993.tb02070.x

Schandry, R., Leopold, C., & Vogt, M. (1996). Symptom reporting in asthma patients and insulin-dependent diabetics. *Biological Psychology*, *42*(1-2), 231–244. https://doi.org/10.1016/0301-0511(95)05157-0

Schillings, C., Schultchen, D., & Pollatos, O. (2021). Effects of a single yoga session on cardiac interoceptive accuracy and emotional experience. *Brain Sciences*, *11*(12), 1572. https://doi.org/10.3390/brainsci11121572

Schoeller, F. A., Zhang, B., Garcia, T., & Reggente, N. (2025). There is no such thing as interoception. *Frontiers in Psychology*, *16*, 1488415. https://doi.org/10.3389/fpsyg.2025.1488415

Schoenberg, P. L. A., & David, A. S. (2014). Biofeedback for psychiatric disorders: a systematic review. *Applied Psychophysiology and Biofeedback*, *39*(2), 109–135. https://doi.org/10.1007/s10484-014-9246-9

Schueller, S. M., Aguilera, A., & Mohr, D. C. (2017). Ecological momentary interventions for depression and anxiety. *Depression and Anxiety*, *34*(6), 540–545. https://doi.org/10.1002/da.22649

Schulz, A., Back, S. N., Schaan, V. K., Bertsch, K., & Vögele, C. (2021). On the construct validity of interoceptive accuracy based on heartbeat counting: Cardiovascular determinants of absolute and tilt-induced change scores. *Biological Psychology*, *164*(108168), 108168. https://doi.org/10.1016/j.biopsycho.2021.108168

Schulz, A., Lass-Hennemann, J., Sütterlin, S., Schächinger, H., & Vögele, C. (2013). Cold pressor stress induces opposite effects on cardioceptive accuracy dependent on assessment paradigm. *Biological Psychology*, *93*(1), 167–174. https://doi.org/10.1016/j.biopsycho.2013.01.007

Schulz, A., & Vögele, C. (2015). Interoception and stress. *Frontiers in Psychology*, *6*, 993. https://doi.org/10.3389/fpsyg.2015.00993

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Shah, P., Hall, R., Catmur, C., & Bird, G. (2016). Alexithymia, not autism, is associated with impaired interoception. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *81*, 215–220. https://doi.org/10.1016/j.cortex.2016.03.021

Sherman, A. R. (1972). Real-life exposure as a primary therapeutic factor in the desensitization treatment of fear. *Journal of Abnormal Psychology*, *79*(1), 19–28. https://doi.org/10.1037/h0032335

Sherrington, C. (1907). The integrative action of the nervous system. *The Journal of Nervous and Mental Disease*, *34*(12), 801–802. https://doi.org/10.1097/00005053-190734120-00021

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Sieverink, F., Kelders, S. M., & van Gemert-Pijnen, J. E. (2017). Clarifying the concept of adherence to eHealth technology: Systematic review on when usage becomes adherence. *Journal of Medical Internet Research*, *19*(12), e402. https://doi.org/10.2196/jmir.8578

Silverman, B., James, C., Misra, S., Schneider, A. T., & Chiaramonte, L. T. (1990). Training perception of acute airflow obstruction. *Annals of Allergy*, *64*(4), 373–375. https://www.ncbi.nlm.nih.gov/pubmed/2321814

Sinclair, S. J., Siefert, C. J., Slavin-Mulford, J. M., Stein, M. B., Renna, M., & Blais, M. A. (2012). Psychometric evaluation and normative data for the depression, anxiety, and stress scales-21 (DASS-21) in a nonclinical sample of U.S. adults. *Evaluation & the Health Professions*, *35*(3), 259–279. https://doi.org/10.1177/0163278711424282

Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., Blazeby, J. M., Boyd, K. A., Craig, N., French, D. P., McIntosh, E., Petticrew, M., Rycroft-Malone, J., White, M., & Moore, L. (2021). A new framework for developing and evaluating complex interventions: Update of Medical Research Council guidance. *BMJ*, *374*, n2061. https://doi.org/10.1136/bmj.n2061

Smith, R., Feinstein, J. S., Kuplicki, R., Forthman, K. L., Stewart, J. L., Paulus, M. P., Tulsa 1000

Investigators, & Khalsa, S. S. (2021). Perceptual insensitivity to the modulation of interoceptive signals in depression, anxiety, and substance use disorders. *Scientific Reports*, *11*(1), 2108. https://doi.org/10.1038/s41598-021-81307-3

Smith, R., Kuplicki, R., Feinstein, J., Forthman, K. L., Stewart, J. L., Paulus, M. P., Tulsa 1000 investigators, & Khalsa, S. S. (2020). A Bayesian computational model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. *PLoS Computational Biology*, *16*(12), e1008484. https://doi.org/10.1371/journal.pcbi.1008484

Spooner, R., Bird, J. M., Clemente, R., Izagirre, N. I., Fueyo, E. F., Watling, D., Plans, D., Brewer, R., Bird, G., & Murphy, J. (2024). No differences between remote and laboratory-based testing of cardiac interoceptive accuracy using the phase adjustment task. *Scientific Reports*, *14*(1), 28524. https://doi.org/10.1038/s41598-024-79125-4

Spooner, R., Bird, J. M., Irigoras Izagirre, N., Clemente, R., Fernandez Fueyo, E., Budworth, G., Cocirla, D., Todd, J., Aspell, J., Leganes, M., Watling, D., Plans, D., Brewer, R., & Murphy, J. (2024). Exploring sex differences in cardiac interoceptive accuracy using the phase adjustment task. *Psychophysiology*, *61*(12), e14689. https://doi.org/10.1111/psyp.14689

Spreng, R. N., McKinnon, M. C., Mar, R. A., & Levine, B. (2009). The Toronto Empathy Questionnaire: scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment*, *91*(1), 62–71. https://doi.org/10.1080/00223890802484381

Staniszewska, S., Brett, J., Simera, I., Seers, K., Mockford, C., Goodlad, S., Altman, D. G., Moher, D., Barber, R., Denegri, S., Entwistle, A., Littlejohns, P., Morris, C., Suleman, R., Thomas, V., & Tysall, C. (2017). GRIPP2 reporting checklists: Tools to improve reporting of patient and public involvement in research. *BMJ*, *358*, j3453. https://doi.org/10.1136/bmj.j3453

Stubbs, B., Koyanagi, A., Hallgren, M., Firth, J., Richards, J., Schuch, F., Rosenbaum, S., Mugisha, J., Veronese, N., Lahti, J., & Vancampfort, D. (2017). Physical activity and anxiety: A perspective from the World Health Survey. *Journal of Affective Disorders*, *208*, 545–552. https://doi.org/10.1016/j.jad.2016.10.028

Sugawara, A., Terasawa, Y., Katsunuma, R., & Sekiguchi, A. (2020). Effects of interoceptive training

on decision making, anxiety, and somatic symptoms. *BioPsychoSocial Medicine*, *14*(1), 7. https://doi.org/10.1186/s13030-020-00179-7

Suksasilp, C., & Garfinkel, S. N. (2022). Towards a comprehensive assessment of interoception in a multi-dimensional framework. *Biological Psychology*, *168*(108262), 108262. https://doi.org/10.1016/j.biopsycho.2022.108262

Swain, J., Hancock, K., Hainsworth, C., & Bowman, J. (2013). Acceptance and commitment therapy in the treatment of anxiety: a systematic review. *Clinical Psychology Review*, *33*(8), 965–978. https://doi.org/10.1016/j.cpr.2013.07.002

Tan, L., Wang, M.-J., Modini, M., Joyce, S., Mykletun, A., Christensen, H., & Harvey, S. B. (2014). Preventing the development of depression at work: a systematic review and meta-analysis of universal interventions in the workplace. *BMC Medicine*, *12*(1), 74. https://doi.org/10.1186/1741-7015-12-74

Taylor, S., Koch, W. J., & Crockett, D. J. (1991). Anxiety sensitivity, trait anxiety, and the anxiety disorders. *Journal of Anxiety Disorders*, *5*(4), 293–311. https://doi.org/10.1016/0887-6185(91)90030-w

Telch, M. J., Lucas, J. A., Schmidt, N. B., Hanna, H. H., LaNae Jaimez, T., & Lucas, R. A. (1993). Group cognitive-behavioral treatment of panic disorder. *Behaviour Research and Therapy*, *31*(3), 279–287. https://doi.org/10.1016/0005-7967(93)90026-q

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J., & Stewart-Brown, S. (2007). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health and Quality of Life Outcomes*, *5*(1), 63. https://doi.org/10.1186/1477-7525-5-63

Terasawa, Y., Moriguchi, Y., Tochizawa, S., & Umeda, S. (2014). Interoceptive sensitivity predicts sensitivity to the emotions of others. *Cognition & Emotion*, *28*(8), 1435–1448. https://doi.org/10.1080/02699931.2014.888988

Terhaar, J., Viola, F. C., Bär, K.-J., & Debener, S. (2012). Heartbeat evoked potentials mirror altered body perception in depressed patients. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *123*(10), 1950–1957.

https://doi.org/10.1016/j.clinph.2012.02.086

The Stationery Office. (2019). *Health and Safety Executive (HSE). Tackling Work-Related Stress Using the Management Standards Approach: A Step-by-Step Workbook*.

Thomas, G., Hartley, R. D., & Kincaid, J. P. (1975). Test-retest and inter-analyst reliability of the automated readability index, Flesch reading ease score, and the fog count. *Journal of Reading Behavior*, *7*(2), 149–154. https://doi.org/10.1080/10862967509547131

Thorley, C. (2017). *Not By Degrees: Improving Student Mental Health in the UK's Universities*.

Thrun, M. C., Hansen-Goos, O., Griese, R., & Ultsch, A. (2019). R Package Gaussian Mixture Models (GMM). *ResearchGate*. http://dx.doi.org/

Todd, J., Plans, D., Lee, M. C., Bird, J. M., Morelli, D., Cunningham, A., Ponzo, S., Murphy, J., Bird, G., & Aspell, J. E. (2024). Heightened interoception in adults with fibromyalgia. *Biological Psychology*, *186*(108761), 108761. https://doi.org/10.1016/j.biopsycho.2024.108761

Tonneau, F. (2004). Relational frame theory: A post-Skinnerian account of human language and cognition Edited by Steven C. Hayes, Dermot Barnes-Holmes & Bryan Roche. *British Journal of Psychology (London, England: 1953)*, *95*(2), 265–268. https://doi.org/10.1348/000712604773952467

Trevisan, D. A., Altschuler, M. R., Bagdasarov, A., Carlos, C., Duan, S., Hamo, E., Kala, S., McNair, M. L., Parker, T., Stahl, D., Winkelman, T., Zhou, M., & McPartland, J. C. (2019). A meta-analysis on the relationship between interoceptive awareness and alexithymia: Distinguishing interoceptive accuracy and sensibility. *Journal of Abnormal Psychology*, *128*(8), 765–776. https://doi.org/10.1037/abn0000454

Vaitl, D. (2004). Psychophysiologie der Interozeption. *Entspannungsverfahren. Das Praxishandbuch*, *3*, 35–51.

Vallée, A. (2024). Envisioning the future of personalized medicine: Role and realities of digital twins. *Journal of Medical Internet Research*, *26*, e50204. https://doi.org/10.2196/50204

Van Dam, N. T., Sheppard, S. C., Forsyth, J. P., & Earleywine, M. (2011). Self-compassion is a better predictor than mindfulness of symptom severity and quality of life in mixed anxiety and depression. *Journal of Anxiety Disorders*, *25*(1), 123–130.

https://doi.org/10.1016/j.janxdis.2010.08.011

Van der Does, A. J., Van Dyck, R., & Spinhoven, P. (1997). Accurate heartbeat perception in panic disorder: fact and artefact. *Journal of Affective Disorders*, *43*(2), 121–130. https://doi.org/10.1016/s0165-0327(96)01414-0

Viskovich, S., & Pakenham, K. I. (2018). Pilot evaluation of a web-based acceptance and commitment therapy program to promote mental health skills in university students. *Journal of Clinical Psychology*, *74*(12), 2047–2069. https://doi.org/10.1002/jclp.22656

Waltz, E. (2018). Pear approval signals FDA readiness for digital treatments. *Nature Biotechnology*, *36*(6), 481–482. https://doi.org/10.1038/nbt0618-481

Weber, S., Lorenz, C., & Hemmings, N. (2019). Improving stress and positive mental health at work via an app-based intervention: A large-scale multi-center randomized control trial. *Frontiers in Psychology*, *10*, 2745. https://doi.org/10.3389/fpsyg.2019.02745

Weisel, K. K., Lehr, D., Heber, E., Zarski, A.-C., Berking, M., Riper, H., & Ebert, D. D. (2018). Severely burdened individuals do not need to be excluded from internet-based and mobile-based stress management: Effect modifiers of treatment outcomes from three randomized controlled trials. *Journal of Medical Internet Research*, *20*(6), e211. https://doi.org/10.2196/jmir.9387

Werntz, A., Amado, S., Jasman, M., Ervin, A., & Rhodes, J. E. (2023). Providing human support for the use of digital mental health interventions: Systematic meta-review. *Journal of Medical Internet Research*, *25*(1), e42864. https://doi.org/10.2196/42864

Whitehead, W. E., Drescher - Psychophysiology, V. M., & 1980. (1980). Perception of gastric contractions and self-control of gastric motility. *Wiley Online Library*. http://onlinelibrary.wiley.com/doi/10.1111/j.1469-8986.1980.tb02296.x/full

Whitehead, W. E., Drescher, V. M., Heiman, P., & Blackwell, B. (1977). Relation of heart rate control to heartbeat perception. *Biofeedback and Self-Regulation*, *2*(4), 371–392. https://doi.org/10.1007/bf00998623

Whitton, A. E., Proudfoot, J., Clarke, J., Birch, M.-R., Parker, G., Manicavasagar, V., & Hadzi-Pavlovic, D. (2015). Breaking open the black box: Isolating the most potent features of a Web and mobile phone-based intervention for depression, anxiety, and stress. *JMIR Mental Health*,

2(1), e3. https://doi.org/10.2196/mental.3573

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer.

Wiens, S., & Palmer, S. N. (2001). Quadratic trend analysis and heartbeat detection. *Biological Psychology*, *58*(2), 159–175. https://doi.org/10.1016/s0301-0511(01)00110-7

Windmann, S., Schonecke, O. W., Fröhlig, G., & Maldener, G. (1999). Dissociating beliefs about heart rates and actual heart rates in patients with cardiac pacemakers. *Psychophysiology*, *36*(3), 339–342. https://doi.org/10.1017/s0048577299980381

Wittkamp, M. F., Bertsch, K., Vögele, C., & Schulz, A. (2018). A latent state-trait analysis of interoceptive accuracy. *Psychophysiology*, *55*(6), e13055. https://doi.org/10.1111/psyp.13055

Wolitzky-Taylor, K. B., Arch, J. J., Rosenfield, D., & Craske, M. G. (2012). Moderators and non-specific predictors of treatment outcome for anxiety disorders: a comparison of cognitive behavioral therapy to acceptance and commitment therapy. *Journal of Consulting and Clinical Psychology*, *80*(5), 786–799. https://doi.org/10.1037/a0029418

Yardley, L., Morrison, L., Bradbury, K., & Muller, I. (2015). The person-based approach to intervention development: application to digital health-related behavior change interventions. *Journal of Medical Internet Research*, *17*(1), e30. https://doi.org/10.2196/jmir.4055

Yardley, L., Spring, B. J., Riper, H., Morrison, L. G., Crane, D. H., Curtis, K., Merchant, G. C., Naughton, F., & Blandford, A. (2016). Understanding and promoting effective engagement with digital behavior change interventions. *American Journal of Preventive Medicine*, *51*(5), 833–842. https://doi.org/10.1016/j.amepre.2016.06.015

Yates, A. J., Jones, K. E., Marie, G. V., & Hogben, J. H. (1985). Detection of the heartbeat and events in the cardiac cycle. *Psychophysiology*, *22*(5), 561–567. https://doi.org/10.1111/j.1469-8986.1985.tb01651.x

Zamariola, G., Maurage, P., Luminet, O., & Corneille, O. (2018). Interoceptive accuracy scores from the heartbeat counting task are problematic: Evidence from simple bivariate correlations. *Biological Psychology*, *137*, 12–17. https://doi.org/10.1016/j.biopsycho.2018.06.006