



Hyllested Pedersen, Casper (2025) *Pursuing clarity of purpose and generalizable research practices for mental health apps and recommendations*. PhD thesis.

<https://theses.gla.ac.uk/84932/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

Pursuing clarity of purpose and generalizable research practices for mental health apps and recommendations

Casper Hyllested Pedersen

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow



University
of Glasgow

September 2024

Abstract

Mental health issues are a diverse and widespread problem. Whilst effective treatments and solutions exist, numerous obstacles make it difficult to deliver them to the people who need them. One promising way to lower access barriers is through mental health apps, a ubiquitous, often inexpensive solution. Furthermore, through the existence of recommender systems, delivery of content within the app can be personalized, allowing for the personal tailoring of a widely available resource. Yet whilst there have been some promising results in the field regarding their efficacy, commercial deployment is outpacing the supporting science. Many apps are untested, there are very few unifying methodological frameworks in place and a shallow understanding of mechanisms of change. The research that does exist is comprised mostly of exploration and testing of novel algorithms or solutions leading to little coherence between studies and generalizability of findings is largely unknown. Whilst, to some extent, this is to be expected in early-stage research, it would be far more beneficial in the long run to pursue foundational improvements now rather than later.

This thesis aims to address these issues, to shore up the foundations of research into mental health apps and recommender systems, to identify generalizable practices and pursue a deeper understanding of how we can enable positive mental health change. The current work focuses on establishing the link between engagement and mental health outcomes, as research in the field has a tendency to make assumptions regarding the beneficial effects of increased engagement. Through a systematic review of recommender systems in the mental health context and three studies, we evaluate how recommendations can be tailored towards both outcomes, how we can increase congruence in research by clear, goal-oriented definition of variables, and whether academic research translates to real world effects. The current work investigates the influence of financial incentivization on engagement and mental health outcomes, analyzes a commercial dataset gathered over several years and explores the relationship between different character traits, behaviors, facets of engagement and short-term and long-term mental health outcomes across a number of domains. In the final chapter the disparate threads will be brought together, presenting broadly applicable recommendations for how researchers can structure individual research to generate more cohesive value in the field as a whole, and suggest how future research may continue to pursue a stronger foundational understanding of mental health change.

Contents

Abstract	i
Acknowledgements	ix
Declaration	x
1 Mental health: Problems and solutions	1
1.1 Introduction	1
1.2 The scope of mental health issues	3
1.3 Treatments, techniques and their barriers	5
1.4 The ubiquitous solution	8
1.5 The role of recommender systems	10
2 Systematic review of mental health recommendations	15
2.1 Introduction	15
2.1.1 Recommender systems and how they work	15
2.1.2 Recommendations, purpose and fairness	18
2.2 Methods	20
2.2.1 Scopus literature search	20
2.2.2 Evaluation and comparison	22
2.3 Results	22
2.4 Discussion	26
3 Financial incentivization and its implications	31
3.1 Introduction	31
3.1.1 The consequences of compensation	31
3.1.2 The current study	34
3.2 Methods	36
3.2.1 Participants	36
3.2.2 Design	36

3.2.3	Procedure	37
3.2.4	Materials	38
3.2.5	Analysis	41
3.3	Results	42
3.3.1	Descriptives and completion metrics	42
3.3.2	PCA and analysis of post-activity ratings	45
3.3.3	Analysis of engagement	46
3.3.4	Analysis of change in WHO-5 and PSS-10 scores	47
3.4	Discussion	48
3.4.1	Main findings	48
3.4.2	Further exploration of variables	50
3.4.3	Future directions	52
3.4.4	Limitations	53
3.4.5	Conclusion	54
4	Exploring real world data	55
4.1	Introduction	55
4.2	Methods	57
4.2.1	Participants, assessments, activities and filters	57
4.2.2	Materials	58
4.2.3	Analysis	62
4.3	Results	65
4.3.1	Descriptives and psychometrics	65
4.3.2	Regression models	72
4.4	Discussion	75
4.4.1	Comparing mental health assessments	75
4.4.2	Mental health change	77
4.4.3	Limitations	78
4.4.4	Conclusion	79
5	Establishing predictors of improvement	80
5.1	Introduction	80
5.2	Methods	83
5.2.1	Participants	83
5.2.2	Procedure	83
5.2.3	Measures	84
5.2.4	Analysis	87
5.3	Results	89
5.3.1	Descriptives and completion metrics	89

5.3.2	Psychometric analysis of the post-activity ratings	93
5.3.3	Analysis of post-activity ratings and their potential predictors.	95
5.3.4	Analysis of offboarding data.	100
5.3.5	Analysis of follow-up data	101
5.4	Discussion	101
5.4.1	The immediate effect of mental health activities	101
5.4.2	Activity effects on short-term improvement	103
5.4.3	Predictors of mean activity ratings	104
5.4.4	Long term change	105
5.5	Proof of concept recommender system	108
5.5.1	Rationale and methods	108
5.5.2	Results and discussion	110
5.5.3	Limitations	112
5.5.4	Conclusion	112
6	General discussion	113
6.1	Summary of previous chapters	113
6.2	Working towards clarity of purpose	116
6.3	Establishing methodological guidelines	118
6.3.1	Defining and selecting engagement and mental health variables	118
6.3.2	Generalizability and ecological validity	121
6.3.3	Algorithmic considerations	123
6.4	Conclusion	124
	References	126
	Appendix Chapter 3	143
	Appendix Chapter 5	145

List of Tables

2.1	Key findings from the systematic review.	24
3.1	Final model predicting effort ratings (adjusted $R^2 = 0.094$)	46
3.2	Final model predicting <i>usefulness and satisfaction</i> component (adjusted $R^2 = 0.084$)	46
3.3	Final model predicting activity completions	47
3.4	Final model predicting change in WHO-5 scores (adjusted $R^2 = 0.202$)	48
3.5	Final model predicting change in PSS-10 scores (adjusted $R^2 = 0.272$)	48
4.1	First assessment scores for both groups	67
4.2	Change in assessment scores for both groups	69
4.3	Regression results for mental health change over time in each group	72
4.4	Regression results for group comparison of mental health change	72
4.5	Regression results for predictors of mental health change	73
4.6	Regression results for predictors of activity completions	75
5.1	Intraclass Correlation Coefficients	93
5.2	Regression results for predictors of post-activity ratings	96
5.3	Regression results for final models predicting mean post-activity ratings	99
5.4	Final model predicting activity completions	100
5.5	Regression results for final models predicting change in mental health	101
5.6	Results of algorithm evaluations	110
1	Model predicting effort ratings including all variables (adjusted $R^2 = 0.049$)	143
2	Model predicting usefulness and satisfaction component including all variables (adjusted $R^2 = 0.028$)	143
3	Model predicting activity completions including all variables	144
4	Model predicting WHO-5 change including all variables (adjusted $R^2 = 0.173$)	144
5	Model predicting PSS-10 change including all variables (adjusted $R^2 = 0.211$)	144
6	Models predicting mean post-activity ratings including all variables	145

7 Model predicting activity completions including all variables 149

8 Models predicting mean post-activity ratings including all variables 150

List of Figures

1.1	Primary focus areas of the thesis	3
2.1	Prisma Diagram of filtering procedure.	23
3.1	Data collection procedure.	37
3.2	Recommendations page.	39
3.3	Example of an activity page.	40
3.4	Ratings part 1	40
3.5	Ratings part 2	41
3.6	Total individual activity completions and their duration	43
3.7	Total activity completions by date and reminder	44
3.8	Total activity completions by group at each activity interval.	45
4.1	Home and library screens of the foundations app	61
	(a) Home screen for Foundations app.	61
	(b) Library screen for Foundations app.	61
4.1	Activity pages for the "Body Reset" activity in Foundations	62
	(c) First page for the "Body Reset" activity in Foundations.	62
	(d) Main activity page for the "Body Reset" activity in Foundations.	62
4.2	Distribution of total activity completions by user	65
4.3	Activity completions by module	66
4.4	Distribution of user mental health scores in the main group for their first assessment	67
4.5	Distribution of user mental health scores in the control group for their first as- essment	68
4.6	Distribution of user mental health score change in the main group	69
4.7	Distribution of user mental health score change in the control group	70
4.8	Gstudy variance profiles	71
5.1	Library page.	86
5.2	Ratings part 1	87
5.3	Ratings part 2	87

5.4	Distribution of activity completions	89
5.5	Total activity completions by type	90
5.6	Distribution of scores for post-activity ratings	91
5.7	Mental health scores at onboarding and exit	92
	(a) WEMWBS-14 scores at onboarding and exit.	92
	(b) PSS-10 scores at onboarding and exit.	92
5.7	Mental health scores at onboarding, exit and follow-up	92
	(c) WEMWBS-14 scores at onboarding, exit and follow-up.	92
	(d) PSS-10 scores at onboarding, exit and follow-up.	92
5.8	Mental health score change from onboarding to exit.	92
	(a) WEMWBS-14 score change from onboarding to exit.	92
	(b) PSS-10 score change from onboarding to exit.	92
5.8	Mental health score change from onboarding to exit and exit to follow-up.	93
	(c) WEMWBS-14 score change from onboarding to exit and exit to follow-up.	93
	(d) PSS-10 score change from onboarding to exit and exit to follow-up.	93
5.9	Gstudy variance profiles	95

Acknowledgements

Several individuals contributed to this thesis. My two supervisors, Christoph Scheepers and Aleksandar Matic have been involved every step of the way. They have helped shaped the research at each stage, from study design to final edits.

Another researcher has also been directly involved in the work behind multiple chapters. For the studies described in chapter 3 and chapter 5, Juliane Kloidt made considerable contributions. For both studies, Juliane Kloidt had equal parts in the study design, handled most of the data collection in Qualtrics, and managed participant recruitment in Prolific. For the study in chapter 3, Juliane Kloidt also wrote parts of the manuscript and part of the code for the analysis.

Several individuals currently or previously employed at Koa Health have also made contributions along the way. Aside from my supervisor Aleksandar Matic, special thanks go to Bartłomiej Skorulski, Ludovik Coba, Federico Lucchesi, Gabriele Sottocornola and Giovanni Maffei for helping shape the research project in the early stages, and João Guerreiro and Roger Garriga Calleja for their contributions later on, especially in the methods and analysis of the data presented in chapter 4.

Finally I want to thank Marion Roth for continuous input and for helping with editing and structuring the thesis.

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where it is stated otherwise by reference or acknowledgment, the work presented is entirely my own.

Chapter 1

Mental health: Problems and solutions

1.1 Introduction

Mental health issues are a widespread, global problem with potentially severe individual and social consequences. It is also a massively varied field of research, with a multitude of symptoms, disorders and treatments. It has also become increasingly interdisciplinary, with technological advancements becoming more and more relevant to how we design our studies and our solutions. The focus of this thesis is understanding how different disciplines are being brought together, and how we can best shape research going forward. It will look at how mental health apps have emerged as a popular solution to mental health issues due to its ubiquitous nature and far-reaching advantages. It will also investigate the personalization techniques that are being implemented, mostly through recommender systems, to tailor content to individual wants and needs.

The focus is largely methodological. The field of mental health apps, and especially the study of recommender systems within mental health apps, is young, but there is a great deal of knowledge that has been inherited from the parent fields of psychology and computer science. Each field has its own practices and traditions, and in many cases collaborative efforts between respective researchers in each area brings advantages as they support each other with complementary skill sets and expertise. However, each field also has its own baggage, well-established methodological problems, or even just practices that do not translate well into a novel focus area. This work will focus on where computer science and psychology support each other, and where they clash, or bring inherited problems to novel applications.

One of the ways this thesis sets itself apart, is that it does not aim to present a novel solution by the end. As will be demonstrated throughout, there is a tendency across parent fields to focus on innovation and new ways to tackle the problem of mental health. New treatments, new apps,

new algorithms, these constitute a large part of published papers. It is rare, especially with algorithms, that research is replicated, and studies are often designed to answer the question "Does it work?", not "How does it work?" or "Why does it work?". This has led to an over-saturation of solutions, without a strong ability to differentiate between them, and a lesser understanding of generalizability, ecological validity or mechanisms of change.

Here, we will explore how we can refine our future research to give us better clarity of purpose and coherence of findings across the literature. We will focus on three pillars of research, shown in figure 1.1, with each chapter addressing different research questions within, driving ultimately towards a stronger clarity of purpose and more robust research in the field. We will look at the relationship between engagement and mental health improvement, and how different stakeholders may have different target outcomes for mental health apps. We will explore how academic methodologies reflect real world findings and how to improve generalizability and ecological validity of research. We will explore the mechanisms of mental health change, and how to define these as concrete variables to be used in research and to develop personalized solutions. We will explore these themes throughout the current work to drive towards robust results, clarity of purpose and an understanding of mechanisms of change in mental health recommender systems.

This thesis represents a four-year effort to merge parallel fields and parallel approaches. It was an attempt to bring together academia and industry, psychology and computer science. An attempt to take reflective psychological methodologies and ground them in reality, investigate generalizability and see what happens when the tight leash of controlled variables is loosened just a little bit to reflect what occurs in the real world. It tried to tame data gathered in the wild, impose a modicum of control upon the chaos of commercial information and find out what happens underneath the hood rather than just replacing the engine to see if the car drives faster.

To begin with, the rest of this chapter will establish the context for the rest of the thesis. It will not jump straight into mental health apps, activities and recommendations, but rather start by acknowledging the issues that defined a *need* for these solutions, and keep that in mind throughout. Mental health apps are meant to improve or maintain user mental health, first and foremost. This, as we will see, is sometimes forgotten, or at least not acknowledged. Everything will be framed by this, but it will also go beyond a binary mindset. It will look not just at whether a solution *helps* or *does not help*, but *who* it can potentially help and reflect on why that might be, and how we can better facilitate getting the right solutions to the right users. It will, in over a hundred pages, barely scratch the surface of what we can do, and show how much more needs to be done.

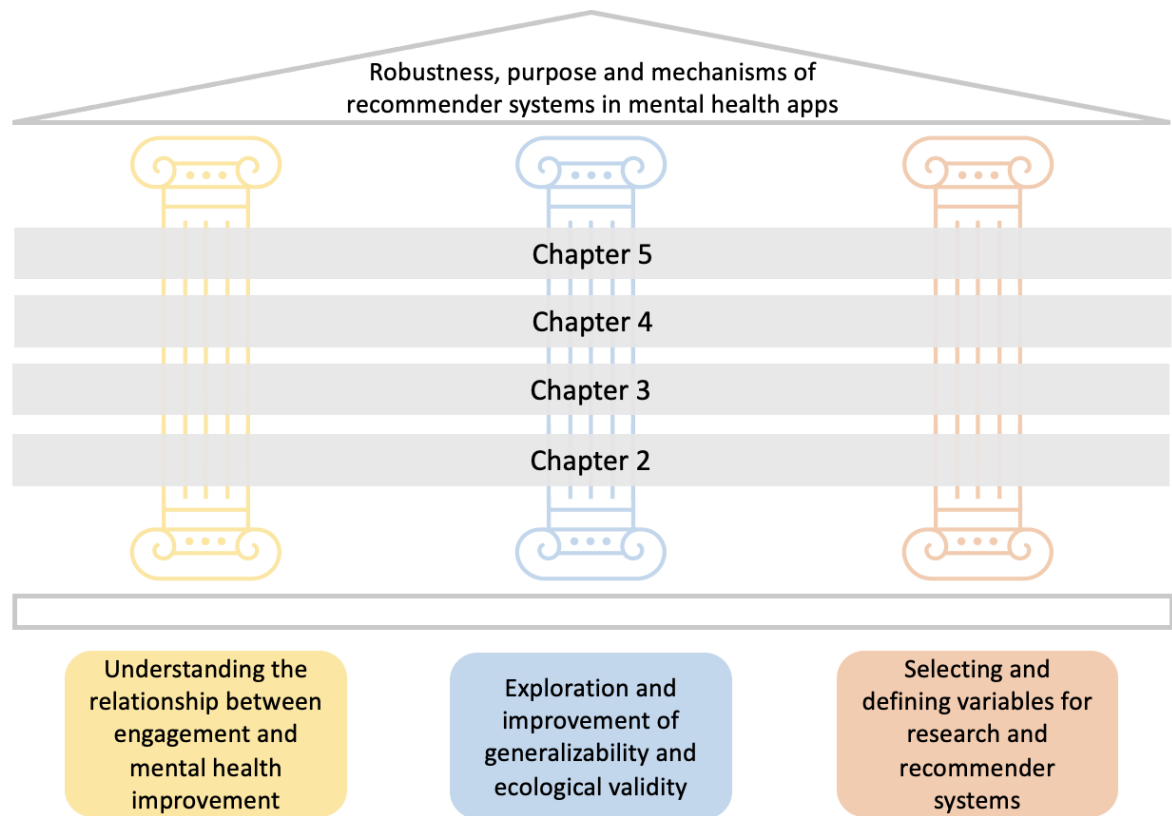


Figure 1.1: Primary focus areas of the thesis

1.2 The scope of mental health issues

In later chapters this thesis will primarily investigate a constrained set of online mental health activities based on mindfulness and meditation, positive psychology and relaxation techniques. It will investigate the effect of these largely on stress and mental wellbeing in application users and study participants. However, much as the specific findings are valuable in their own right, the aim of the research was much wider, more methodological. This chapter, and to some extent Chapter 2, will shed light on the broader context which every subsequent chapter will relate back to.

The first part of this comes down to mental health. There are hundreds of classified mental health disorders in the Diagnostics and Statistics Manual of mental disorders (DSM-5; American Psychiatric Association (2013)). On an overarching scale the problem is severe, and has affected individuals, groups and politics for decades on a global level (The WHO World Mental Health Survey Consortium, 2004; World Health Organization, 2022), at a staggering cost to national economies. In a large scale mental health survey, 31% of 13,984 students screened positive for a 12-month mental disorder (Auerbach et al., 2018). In the US, mental illness affects 1 in 5 adults annually (Substance Abuse and Mental Health Services Administration, 2020). In

Denmark, the percentage of adults reporting high stress scores increased from 20.8% to 29.1% from 2010 to 2021 (Rosendahl, 2022). In China, mental illness accounts for over 20% of all diseases (Fan, 2013), and 4.1% of the population suffer from clinical levels of depression alone (Qin, Wang, & Hsieh, 2018). During the Covid-19 pandemic, numbers skyrocketed, and the prevalence of anxiety in the general population across continents was 27.3% (Pashazadeh Kan et al., 2021).

That is as far as I will go in my description of the general problem. I could fill a book simply listing each mental health issue and the damage it brings with it. I am not interested in saturating my introduction with information that will not be referred to again. Throughout the thesis we will occasionally explore specific fields of mental health as they become relevant to the research at hand, but for now, all I am trying to make clear is that there is a mental health problem. And with this problem brings the demand, the *need* for a solution.

This, however is easier said than done, because it is a multifaceted problem. The statistics mentioned are snapshots. Broad, general, terrifying snapshots. They are also, in their own disheartening way, straight forward. They are big numbers, easy to comprehend, and shocking enough to scare all but the most numbed researchers, not to mention laypeople unfamiliar with the extent of mental health problems. But beyond their purpose in fear mongering, they are not often actually *useful*. Much of this thesis concerns itself with generalizability and individual differences, and very rarely is it safe to use central tendencies and generic statistics when dealing with different subsamples of the population, and certainly not when dealing with unique human beings. To demonstrate this, let us sample the literature to more clearly show how muddy the waters of mental health research can get.

Within groups, prevalence and manifestation of mental health issues vary greatly. For example, gender differences are marked in anxiety. The same systematic review that reported continental prevalence of anxiety reported that prevalence of anxiety in women was at 32.4%, and 24.9% for men (Pashazadeh Kan et al., 2021). The prevalence of anxiety *symptoms* was at a 20% difference (47.8% for women, 27.8% for men). Externalizing disorders however (such as attention-deficit disorders or substance use disorders) present more commonly in men (Boyd et al., 2015). Location is another factor which leads to heterogeneous results. Across continents, the the prevalence of anxiety varied massively, ranging from 24.5% in Asia to 61.8% in Africa (Pashazadeh Kan et al., 2021). Beyond this, even when looking at individuals who share common traits, individual differences still influence mental health. One study with a relatively homogeneous sample (all male, aged 18-30, participating at a single location) still showed considerable differences in stress sensitivity, linked to neurological and physiological differences and personality (Henckens et al., 2016). Beyond this there exists an abundance of situational or circumstantial risk and protective factors which can influence whether (and how) individuals may experience any mental health disorder (Arango et al., 2021).

Beyond prevalence and risk factors, it is far from universal how mental health issues present. Each person's disorder may manifest differently. Diagnoses guided by the DSM-5 are often determined by a number of symptoms from a larger potential pool (American Psychiatric Association, 2013). A major depressive episode, for example, is diagnosed by a minimum of 5 symptoms out of a possible 9. A minor depressive episode is reflected in only 2-4 symptoms. Depression can look very different for individuals depending on their combination of symptoms, and some solutions will fit better to some than others. Comorbidity of mental disorders can then further compound research difficulties. Individuals suffering from one mental health disorder are more likely to develop another (McGrath et al., 2020), explaining why it is often common to see people reporting multiple lifelong mental health disorders (McLafferty et al., 2017). Each new mental health problem can bring with it a whole host of new symptoms and interactions.

This barrage of contrasting statistics is enough to set anyone's head spinning, and it is still only a very slim section of mental health research. But this is not an indictment of the field or the literature, simply a reflection of the chaotic nature of the subject material. Even anecdotally, if we consider the uniqueness of the individuals we know, the differences across cultures and generations that we experience, it is not surprising that the human experience and perspective is wide and varied. Considering this diversity is simultaneously important and frustrating when researching predictors of mental health, or designing a mental health app.

However, despite these differences, we *do* have solutions. Some treatments, some activities, some techniques *work*. So why does it remain such a widespread problem? Certainly, individual differences play a part. Some people are more susceptible to mental health problems than others, and focusing on individual solutions could remedy that. Some things are also harder to treat than others. Sometimes it comes down to availability, sometimes awareness, sometimes money. The next section will explore what solutions have emerged over the last few decades to deal with mental health problems and highlight some factors that can get in the way of progress.

1.3 Treatments, techniques and their barriers

Solutions come in many forms. In many cases, protective or remedial measures are manageable by individuals, without need for external help. Exercising, for example, has been reported as the most consistent preventative measure across a broad spectrum of disorders (Arango et al., 2021). Mindfulness and meditation has been shown to be effective in improving mental wellbeing and perceived stress levels (Zollars, Poirier, & Pailden, 2019). Maintaining social relationships and engaging with social activities can both help prevent depression and improve depressive symptoms (Cruwys, Haslam, Dingle, Haslam, & Jetten, 2014).

When it comes to therapeutic options it can get crowded. So much so that I have to be careful

not to stray too far from the purpose of the thesis and descend into a labyrinth of acronyms. There are countless reviews detailing the merits of Cognitive Behavioral Therapy (CBT) across several mental health issues such as depression, anxiety and insomnia (López-López et al., 2019; M. D. Mitchell, Gehrman, Perlis, & Umscheid, 2012; Twomey, O'Reilly, & Byrne, 2015), each comprised of tens of separate studies. Therapists have a plethora of choices fit to purpose, such as dialectical behavioral therapy for borderline personality disorder (May, Richardi, & Barth, 2016) or emotion focused therapy for generalized anxiety disorder (Timulak et al., 2022). The list could go on, but this work is not dedicated to comparative analysis of different therapies. To those interested, there exists an abundance of literature debating the merits of specific solutions for specific disorders. As to those that hold direct relevance to the work in later chapters, they will be described in more detail in the appropriate sections.

However, looking at the literature from a bird's eye view serves an important function for the broader context the later chapters reside within. One thing that is immediately clear when researching how to combat mental health issues, is that there is no shortage of options available. Options that are well researched, and on group levels at least, often show great efficacy. Some are options that have been available for decades. So why does the problem persist? Why does the problem in certain cases seem to be *growing*?

There are several possible reasons, likely working in concert. In part, it is simply that mental health issues are largely treatable, not curable. No therapy guarantees success, and especially not when the generalizability of research results to a wider population is uncertain (Stuart, Bradshaw, & Leaf, 2015). Furthermore, many mental health issues are recurrent, not chronic, and even in therapies that show positive results, there are often considerable relapse rates (Ali et al., 2017; Levy, O'Bryan, & Tolin, 2021). So whilst new therapies are continuously developed and old ones are continuously improved, mental health services are bailing water from a leaking boat. This is not a negative reflection of clinicians and mental health professionals, only an acknowledgement of the limitations inherent from the nature of the mental health problem and the tools we currently have at hand.

That being said, there are potentially inefficiencies in how we research and develop these therapies. In the development of so many treatments, it is possible that researchers and clinicians are inadvertently and indirectly working against their own interests. Wampold (2019) describes the 'smorgasboard' of Post-Traumatic Stress Disorder (PTSD) treatments. He notes the benefits of having multiple options, allowing clinicians to tailor solutions to individual clients, and having backup plans in case one therapeutic tool does not work. However, he also notes that there are underlying issues in the way these solutions come about. He details that there is good evidence for the effectiveness of many of the treatments he describes. However, when looking at them side-by-side, there appears to be little difference between how well they work, and that the supporting studies provide, at best, weak evidence for underlying mechanisms of change. Earlier

in this chapter I alluded to this being a generic problem in solution-oriented work, stemming from the fact that research sometimes tends towards over-enthusiasm for novelty. Yet if a novel solution simply provides a new way to solve the same problem *with equal ability*, it is a lateral move, not an improvement.

I do not mean to disparage this type of research. Variety is good, as is having multiple options. Neither does Wampold (2019) discourage the search for new solutions, but he argues that it should come with an equal measure of understanding the mechanisms for change which could provide insights that would allow for improvement across the breadth of therapeutic options. This is a critical distinction, and one that will resonate throughout the rest of the present work. And as will become clear, though this tendency toward novelty may be pronounced in the field of psychotherapy for PTSD, it is far more pronounced in the field of mental health apps.

However, even if we managed to streamline and perfect the process of developing solutions for mental health issues, it matters little if the professionals using them do not even get a chance to attempt to help people. Unfortunately, this is often the case. This is particularly well researched in student samples, where mental health risks tend to be higher than the general population (Stallman, 2010) but support is often more readily available (Eisenberg, Golberstein, & Gollust, 2007; Osborn, Li, Saunders, & Fonagy, 2022). In the UK only 5.1% of students who suffer from depression seek treatment (Macaskill, 2013). In Northern Ireland 10% of students received support, but 22% reported needing it (McLafferty et al., 2017). A larger scale WHO survey reported that only 16.4% of students suffering from various disorders received help for it. One systematic review which explored the literature extensively suggested that although use of services is increasing in certain areas, most students who need help are still not getting it (Osborn et al., 2022).

There are many reasons for this. Numerous barriers to help seeking behavior exist, cultural, structural and individual. Sometimes people do not perceive the need, or have a desire to handle problems on their own (Andrade et al., 2014). Many, even if they perceive or desire the need, avoid seeking help due to stigmatization, or the perception of stigma (Clement et al., 2015; Yamaguchi, Ling, Kim, & Mino, 2014). But even if there is a perceived need and a desire for help, structural barriers can often get in the way, such as the inability to afford mental health care, or a lack of availability (Andrade et al., 2014). These latter barriers become more prominent as severity of issues increase. This makes sense, as higher severity of mental health issues can make the perceived need greater, and bring with it an acknowledgement of the difficulty of tackling problems alone. Yet even though it may be unsurprising, it still compounds the problem. For the people who have more severe issues, people who have perhaps struggled to recognize the need for help, people who have had to overcome stigma, to push through these challenges to then be faced with very tangible barriers like waiting lists or high costs can be devastating, potentially worsening a bad situation (Punton, Dodd, & McNeill, 2022).

And these are very tangible barriers. Waiting-lists for mental health services have a reputation for being long. Research shows that this isn't necessarily overall true, but the variance in waiting times is very high (Andrade et al., 2014; Edbrooke-Childs & Deighton, 2020), potentially stretching to months or even years depending on circumstances. There are simply not enough mental health professionals to help everyone who needs it (Butryn, Bryant, Marchionni, & Sholevar, 2017; Kuehn, 2022), particularly in rural areas. Supply does not meet demand. The disparity of circumstances is also clear financially. Whilst some countries or institutions offer free mental health care, there are still many who report difficulties in affording help (Andrade et al., 2014).

Finally, the big picture begins to come together. We face a mental health crisis, a *global* mental health crisis. And despite excellent efforts by researchers, clinicians, politicians and a host of other individuals, the problem remains. Mental health research and practices have made huge strides in the past decades, yet we still face an uphill struggle. There are methodological obstacles to overcome, logistic obstacles to overcome, individual and political obstacles to overcome.

So, now it is time for the answer to all of our problems. And indeed, we do have one that fits many of the criteria. Mental health apps solve many of the structural issues presented by Andrade et al. (2014), which was especially critical during Covid-19 where some of these seemed insurmountable (Kopelovich et al., 2021). They're cheap, sometimes free, they eliminate the distance issue and they don't have waiting-lists. And although it would be by far preferable to remove the stigma surrounding mental health problems entirely, they do offer a discrete way for people to test out mental health care. They are flexible in their time requirements and they can theoretically let people take control of their own mental health.

No surprise, then, that the market exploded. New apps have been flooding into app stores at an increasing pace for years (Chandrashekar, 2018). However, unfortunately more mental health apps does not necessarily equate to better mental health apps, or even better knowledge surrounding them. The following section will take a look at the literature concerning mental health apps to more thoroughly examine their current value, their potential value and some of the current issues we face in understanding them.

1.4 The ubiquitous solution

The range of potential offered by mental health apps is hard to overstate. It is the main reason the introduction of this thesis had to be so broad. Even though the data collected here largely came from only a sub-sample of activities from a single mental health app, the questions asked and the methodological considerations are relevant to the field as a whole. And the field is large and still expanding.

Wherever there is a mental health issue, there can be a mental health app. Again, we face a near endless list. We can sample the literature and find research on apps for mental wellbeing (Schulze et al., 2024), depression and anxiety (Marshall, Dunstan, & Bartik, 2020) and PTSD (Rodriguez-Paras et al., 2017). A host more is a Google Scholar search away. And those are *researched*, included in a scientific publication in some form or other. Most mental health apps are not (Anthes, 2016; Neary & Schueller, 2018). They are so versatile, because they can deliver almost any kind of solution. They can employ CBT based techniques such as cognitive restructuring and behavioral activation (Denecke, Schmid, & Nüssli, 2022) or they can be based on mindfulness and meditation (Bostock, Crosswell, Prather, & Steptoe, 2019). In some disorders where therapeutic techniques are less effective such as schizophrenia (Jauhar, Laws, & McKenna, 2019), apps still exist which focus on things such as symptom monitoring, motivational support or lifestyle management (Almeida et al., 2019; Schlosser et al., 2018). In short, mental health apps have the potential to cover almost the full range of mental health solutions, exempting perhaps only medicinal options.

So, we have established what mental health apps probably *could* do. We have established that attempts are being made. Best of all, however, is that in the cases where they are put to trial, there is consistent evidence showing that they actually work. There have been numerous reviews attesting to the benefits and efficacy of researched apps (Chandrashekar, 2018; Donker et al., 2013; Khademian, Aslani, & Bastani, 2021; Lecomte et al., 2020; Neary & Schueller, 2018; Wang, Varma, & Prosperi, 2018). However, this is where the story starts to turn. These same reviews that consistently found positive results stemming from the use of mental health apps, heavily emphasise that there are numerous issues in the field, and the critiques are largely consistent between them. First, most apps are not subjected to scientific scrutiny. Those apps that are properly tested generally perform quite well, but they are not representative of all the apps that are not tested. Not only are the benefits of untested apps questionable, they are also potentially harmful (Anthes, 2016; Torous, 2018). Worse, the over-saturation of the app market makes it harder to navigate, and with few existing guidelines it can be very difficult for consumers and clinicians to determine which apps are good (Neary & Schueller, 2018).

It is also a common thread between the reviews that commercial growth and technological advancement far outpaces scientific understanding of mental health apps. Often the solutions delivered by the apps are developed by, or in collaboration with, clinicians or mental health professionals. However, apps are an entirely different medium, and there is very little research comparing the efficacy of mental health apps to face-to-face therapy, and the sparse literature is not congruent on the matter (Cuijpers et al., 2009; Holtz, Kanthawala, Martin, Nelson, & Parrott, 2023). Most of the apps tested in trials are tested against control groups who receive no interventions. Whilst they do show positive results, we cannot make the assumption that it is for the same reasons as therapy, or that they share similar mechanisms for change. This understand-

ing is critical, not only to shore up our understanding of how current mental health apps work, but also to improve our generic understanding of what might matter when designing any mental health app. This is especially critical when there are so many untested apps, which target such a broad spectrum of mental health issues using such a wide range of techniques. The variability of the solutions means that we must push to understand what change, and what aspects of change, are generalizable.

It is also in understanding the underlying mechanisms that we begin to be able to reach more people. Both because new app developers can then understand how well existing research translates to a different field or application, but also because it is the first necessary step to understanding what predicts *individual* change. And this is even more important for mental health app research because, if neglected, it represents a huge drawback for the ubiquitous solution. In traditional therapy there are guidelines, a set of techniques, often a semi-structured process to follow. However, depending on the therapeutic process and the individual, the actual treatment can vary significantly between individuals and occasions. It is up to the clinician to decide which techniques best suit the individual client, up to the clinician to determine the appropriate number of sessions based on individual needs, and if they realize a different skill set is required they can refer clients to different practitioners. In short, they can adapt generic solutions to specific requirements.

Mental health apps, as we have discussed them so far, do not offer this. At their most basic, even when they target a specific disorder, they are generic solutions where the 'treatment plan' is left to the individual, the client. However, the technological capability to offer tailored solutions exists in recommender systems. These algorithms can take information about users, user behaviors or the content of the apps themselves to recommend the next step. This potentially covers for one of the biggest weaknesses of mental health apps, and some are already implementing them.

1.5 The role of recommender systems

Chapter 2 is entirely devoted to the literature on recommender systems, particularly in mental health, yet for now their place in this thesis must be addressed. Recommender systems are in many ways the end goal of this thesis, and originally were planned to have a central role throughout all the chapters. This work was intended to be the marriage of two fields, psychology and computer science, combining technological developments with theoretical knowledge. Mental health apps are, in theory, a perfect fit for the problems the field and industry of mental health faces. However, it quickly became clear in the early stages of the PhD process that they are being built on shaky foundations. Recommendations are more effective when they are built on appropriate knowledge. This is true whether considering algorithms or people. The better a

therapist understands their therapeutic tools, their clients and situational factors, the better their recommendation can be.

Whilst there are some qualitative differences in how algorithms operate, they are still built on from a set of input parameters chosen by people. Typically these are based on knowledge about the users or knowledge about the content recommended. And these parameters are currently not very well understood for either facet. As has already been highlighted, neither individual or group level mechanisms of change are very well researched for mental health apps. This becomes even more important for interdisciplinary research. Most specialists are specialists within one field, and interdisciplinary research is usually collaborative. So whilst experts in computer science could build the best algorithms with the data they have available, if psychologists are not clear on what that data should be, the algorithms will be limited by their input. This is critical, because we are limited in what we can collect, especially in the wild. We cannot get people to fill out scores of questionnaires, track their moods and record their eating patterns. We can, maybe, do some of them, and that means we have to choose. If we choose the wrong variables, generating useful recommendations is like catching air with a butterfly net. No matter how many times you swing, you won't get good results.

Our understanding of the content parameters are, if anything, even more poorly understood in mental health apps. Randomized Controlled Trials (RCTs) have for decades been the gold standard for evaluating interventions in medicine and psychology (Backmann, 2017; Meldrum, 2000) and have maintained that status in research on mental health apps. Whether it lives up to this status is heavily debated (Backmann, 2017; Bondemark & Ruf, 2015; Stuart et al., 2015), perhaps especially in mental health apps (Tønning, Kessing, Bardram, & Faurholt-Jepsen, 2019), but what is not debatable is their commonality. Most studies and most reviews of mental health apps are focused around RCTs. And regardless of their merits in determining efficacy, one drawback of this trend is that there is usually an exclusive focus on the intervention as a whole. The differences between individual solutions or activities within the app are almost never reported. This has similar consequences to a diminished understanding of person based parameters, in our reduced ability to select and develop optimal content for mental health apps. However, despite these limitations it is possible for good software developers to build useful algorithms. Whilst no recommender system will fix broken content, if at least some of the content works, even a simple algorithm could rank and recommend by usefulness.

Here we must stop and address an assumption made throughout the previous paragraphs. Recommender systems identify and recommend the best content for a user. The implicit assumption was that this would be the content most likely to improve their mental health. That, after all, is what a mental health professional does when selecting techniques for their clients, and is often the main reason people start using mental health apps in the first place. However, this is not the traditional use-case for recommender systems. They were traditionally developed for the

entertainment industry, often for commercial use. Whilst it had its roots in cognitive science and information retrieval, the first known recommender system, 'Grundy', recommended books users might like based on stereotypes (Rich, 1979). One of the most successful pioneering attempts was the Netflix movie recommender system Cinematch. Still today, published research papers on recommender systems are by far more common in the domain of entertainment than any other area (Deldjoo, Jannach, Bellogin, Difonzo, & Zanzonelli, 2024). These are settings where it is easy to make the theoretical assumption that user interests and commercial interests are the same. Better book or movie recommendations mean more customer satisfaction, and more sales. A win-win.

But why does this matter? Are recommender systems impossible to adapt to mental health services? Is it impossible to reconcile the occasionally clashing purposes of academia and industry? Chapter 2 will consider these questions in more detail. But it matters because most experts in recommender systems will have developed their expertise from the context of computer science where normative considerations of recommendation purposes are uncommon (Deldjoo et al., 2024; Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019). And it matters because so many mental health apps are developed for commercial purposes. And unfortunately, an app's capacity to improve mental health is not necessarily the driving force in sales. Oftentimes it is things like user ratings that determine which app new users will choose, which does not by default indicate its beneficial properties (Neary & Schueller, 2018). And this is what recommender systems have traditionally improved. User enjoyment and user engagement have been the desired outcomes, recommending the content that is most likely to appeal to the individual. In an industry that is already lacking in its research foundations, where apps are continuously being developed without empirical support for their efficacy, it is a very risky assumption to make that any recommender systems implemented would target improvement and not engagement outcomes. Especially when algorithms focusing on engagement outcomes are the template for most of those who tend to build them, and it is common to ignore the idiosyncracies of an application domain when designing or researching recommender systems (Deldjoo et al., 2024). Yet this should not be taken as a damning indictment of recommender systems that focus on engagement outcomes. No app will improve mental health if users do not engage with it. And we must also consider the value for companies, fairness across stakeholders to further the field (Deldjoo et al., 2024). Engagement is important for commercial viability, and it is common in academic research to neglect these considerations in favor of a focus on user-experience, which can limit practical applications (Abdollahpouri et al., 2020; Jannach & Jugovac, 2019). This thesis does not mean to promote one approach over the other, merely to highlight that the relationship is poorly understood, and that assumption of a relationship without strong evidence can impede our understanding.

That defined this thesis. The end goal is to be able to build recommender systems within mental

health apps that can help users select the most engaging content that will lead to mental health improvement. That end goal stretches well beyond the bounds of this PhD project, because the foundations need to be better established. We need to better understand the underlying mechanics of improvement (or lack thereof) in mental health apps. We need to explore these as inputs in machine learning models. We need to consider algorithmic and technological progress through the lens of psychology when the field and the outcomes they are targeted towards are psychological in nature. To do this, we need to establish generalizable research methods, or at the very least understand how generalizable our research is.

We also cannot ignore the different goals and methods of research in industry and research in academia. It is of no help to anyone if we develop recommender systems to help people in research apps if those are discarded in favor of more commercially beneficial options. Deldjoo et al. (2024) considers this in terms of fairness, where a single-sided perspective of optimization is not necessarily optimal. A tunnel-visioned approach to improving user experience neglects the needs of the provider (for example the need to turn a profit to stay in business), and vice versa. But, as we will see in Chapter 2, the questions regarding engagement and improvement in mental health apps are rarely considered alongside each other, so we have little knowledge of whether these are synergistic or competing. There is some recent research that suggests a positive self-reinforcing link (Ruiz de Villa et al., 2023), however not enough to properly address the balance between user and provider fairness. It is possible that algorithms focusing on improvement as an outcome would lead to more commercial success as users perceive higher efficacy. Conversely it is possible that engagement is beneficial to mental health in and of itself, and pushing for higher engagement metrics will lead to more user improvement. Or it could be these are entirely uncorrelated and developers would have to choose one or the other.

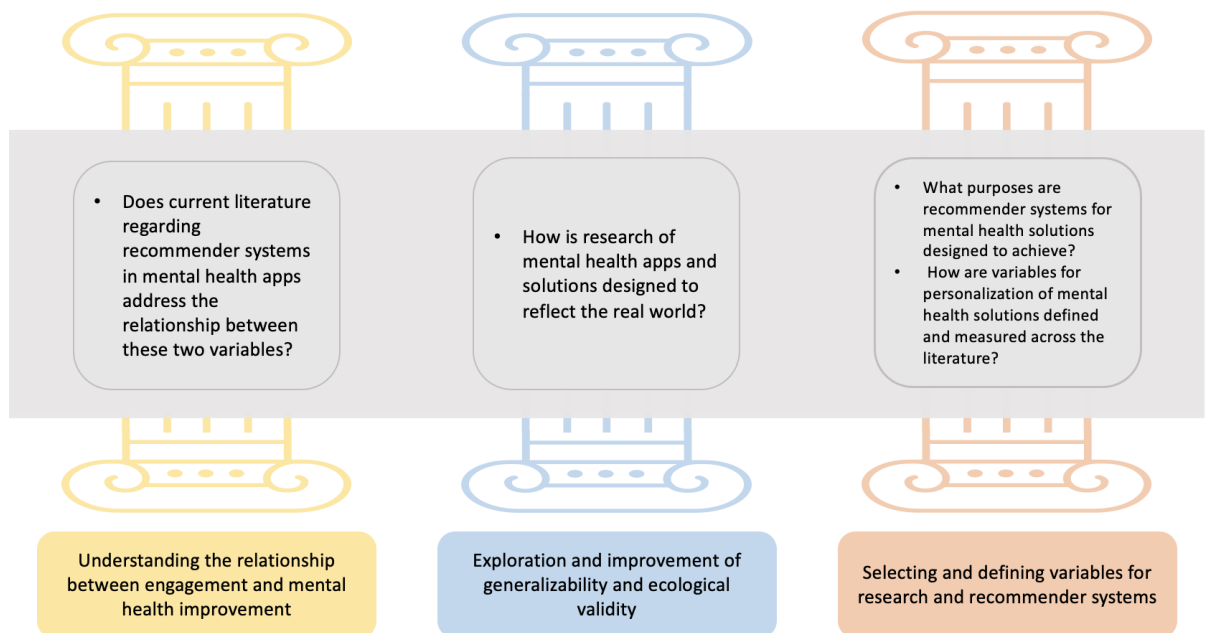
This work goes down several paths. It will not attempt to offer conclusive results in any single of these research areas. It will instead try to bring together scattered research, trying to reconcile academic and commercial goals, to attempt and further encourage *useful* research, that both improves our theoretical knowledge and has applied value. It will consider psychology and computer science as a merged path rather than parallel ones, to promote technological progress that is better suited to the context it exists in, and to drive theoretical research towards real-world purpose.

Chapter 2 will dive into the current applications of recommender systems in mental health apps, with a focus on investigating how engagement and improvement are evaluated. Chapter 3 describes an investigation what effects financial incentivization in research has on engagement and improvement metrics, and consider the methodological implications in academia. Chapter 4 will look at data gathered in the wild through a commercial mental health app, to begin to investigate what might drive mental health improvement. Chapter 5 looks once more at data collected through controlled parameters, diving deeper into what influences mental health improvement

and how it relates to engagement. Finally, Chapter 6 provides a general discussion of findings, limitations and future directions for research.

Chapter 2

Systematic review of mental health recommendations



2.1 Introduction

2.1.1 Recommender systems and how they work

Recommender Systems are currently ripe for studying. They are useful and have a broad range of applications. However, as is becoming increasingly common in such fields of research, the literature is also highly saturated with choices for anyone who wants to implement their own. There is an abundance of information and an excess of methods. Therefore, much as with the descriptions of mental health issues and treatments in Chapter 1, this thesis will not go

into the details of each possible algorithm and its permutations. There are dedicated technical works which can do this more completely (Ricci, Rokach, & Shapira, 2022). This chapter details a systematic review of how mental health apps are studied in the mental health context, specifically as regards considerations of research methodologies, engagement and mental health improvement. However, first it will provide a basic overview for any readers unfamiliar with how recommender systems work and what we can learn from parallel fields so the rest of the chapter, and the rest of the thesis is better understood.

The classic types of recommender systems have traditionally been condensed into two types: *content-based* and *collaborative filtering* (Madadipouya & Sivananthan Chelliah, 2017). These primarily distinguish between leveraging user information or content information to generate recommendations. It is hardwired into recommender system culture to explain basics through movie recommendation, and this proud tradition will be kept in this thesis. A *collaborative filter* operates based on user feedback, either explicit or implicit. Explicit feedback is often comprised of things like ratings, or a like/dislike system. Implicit feedback metrics are passively collected. These could be, for example, whether a user clicks a movie link they have seen on Netflix to read the description, whether they watch the trailer, whether they start the movie and whether they finish it. In a collaborative filter, this information is leveraged to calculate similarities between users. So, at its most simple, when generating a movie recommendation for user A, the algorithm would identify the most similar user, user B. It would then identify which movies user B enjoyed most, potentially from ratings, potentially from how many times they have re-watched a certain movie, potentially a combination of all the data available to the algorithm. It would then select, for example, the top three ranked movies for user B that user A has not watched and recommend these to user A. The number of recommendations may vary, and sometimes algorithms recommend movies user A has already watched, but the basic mechanics follow these rules.

In contrast, a *content-based* recommender system leverages information about the items. In a movie context the most important factor is often genre, but it could also include which actors feature, who the director is, the length of the movie or anything else that pertains to the movie itself. So, if a user only watches movies from the action genre, the algorithm would probably recommend an action movie for their next watch. If they have a favorite actor, movies featuring that actor would be next in line.

There are benefits and drawbacks to both. Content-based recommenders are conceptually simpler. Whilst they can be complex, it is often easier to pin down a movie description than the description of an individual. The overlaps are also often greater. Horror movies share certain features that will likely appeal to a person who enjoys horror movies, but individual preferences are typically more independent. One person may like horror and romance movies, whilst the next likes horror movies and hates romance movies. A collaborative filter, especially one built

on too little information, might recommend romance movies to the second person because they overlapped on their behaviors regarding horror movies with the first person. This is one of the reasons collaborative filtering algorithms tend to have larger data requirements.

However, once they overcome the so-called cold-start problem (performing poorly at the beginning, when relevant user data is still relatively sparse), collaborative filter systems can also be more versatile than content-based recommenders. They tend to recommend a more diverse selection of content, and usually require less prior definition of variables. They need more data from the individual user, but that user does not need to be classified by genre or personality, they can be defined purely by their behavior. Of course, this has the drawback of reducing explainability, making the process more oblique, however this is a well-researched field in its own right with continuous improvements being made (Chen, Zhang, & Wen, 2022; Vultureanu-Albisi & Badica, 2021).

Each method has further advantages and disadvantages, often linked to their applied field. It is also common to develop hybrid methods, leveraging the potential of both (De Croon et al., 2021; Patel, Desai, & Panchal, 2018; Prasad & Kumari, 2012). And whilst collaborative filters and content based recommender systems are commonly considered the base templates, many other distinct types exist, such as recommender systems based on geographical information (Suzuki et al., 2021), or non-personalized popularity based recommenders (Prasad & Kumari, 2012). In principle even using a randomized sampling technique to recommend content could be considered a recommender system.

The complexity of the field only increases from there. So far only simple classifications of recommender systems have been described, and even that only superficially. For each subtype there is an infinite choice of specific algorithms and algorithmic permutations, and once more in this thesis we can see that this huge potential for innovation and diversity has its drawbacks. One review of 121 studies with relatively stringent inclusion criteria identified 205 different machine learning algorithms for generating recommendations (Portugal, Alencar, & Cowan, 2018). This is *good*, but there is an opportunity cost. We can only speculate on the potential quality of the algorithms if those 121 studies had only focused on improving ten of them.

The diversity also comes with increased needs for transparency. Without consistent guidelines or set research standards, generalizability and reliability can be very difficult to maintain in the field. This is an established issue within recommender system research (Dacrema, Cremonesi, & Jannach, 2019) where code isn't always shared and results are sometimes inflated or non-reproducible. This is a problem that is very hard to remedy, and a clear parallel can be seen in social sciences such as psychology, where research in recent years has shown the methodological flaws of earlier research, causing a field-wide replication, theory and generalizability crisis (Oberauer & Lewandowsky, 2019; Yarkoni, 2022). Whilst the issues continue to exist in

psychology, the methodological guidelines are better established, and many translate well into the field of computer science.

One thing that is critical is an awareness of the context of research. The context of this thesis is mental health. In this chapter we will take a look at what research is done specifically in the field of recommender systems for mental health apps through a systematic review, but first it will dive a little deeper into the literature in parallel fields, considering how it might be relevant and where it might differ. Chapter 1 already described the setting from which recommender systems emerged (namely the entertainment industry), yet other contexts exist where the purpose of recommendations have been re-evaluated, and where the goals of different stakeholders or researchers have been considered.

2.1.2 Recommendations, purpose and fairness

This work has already touched upon the goals of recommendations in other fields. Largely, this involves user retention and engagement. Netflix wants users to watch more movies so they are more inclined to keep subscribing. Amazon wants to recommend products users are more likely to buy to increase their sales. And quite often this is also in the interest of the user to some extent. Receiving better movie recommendations is likely to align with user interests. Their enjoyment and satisfaction is linked with their engagement with content they like. Theoretically, for the most part, what the user wants and what the companies want are the same. To deliver the solution that users will enjoy the most. This mentality has historically saturated the academic view of recommender systems, the perspective that they are benevolent win-win tools (Jannach & Jugovac, 2019; Lewis, Ferguson, Wilks, Jones, & Picard, 2022).

However there are caveats, ethical considerations that only increase in other settings. It has recently become a dominant question in the field of recommender systems what ethical responsibilities researchers and developers have, whether we should be considering things like fairness and potential harm, not just efficiency (Deldjoo et al., 2024; Ekstrand, Das, Burke, & Diaz, 2022; Ntoutsis et al., 2020). This is an extremely complex process, as 'fairness' is both vague and multifaceted. Deldjoo et al. (2024) provide a comprehensive overview of what constitutes fairness in the context of recommender systems and how it is tackled in the literature.

Several of the issues they highlight are highly relevant to this thesis. First, they consider fairness as regards consumers and suppliers, formalized as single-sided versus multi-sided fairness. A recommender system or research project designed with improving the function of a recommender system with only one stakeholder in mind would be single-sided and those designed with multiple stakeholders in mind would be multi-sided. Although Deldjoo et al. (2024) do not consider the mental health app context in their review, this is nevertheless directly applicable. Balancing the financial needs of a mental health app company with the needs of the users to

get useful (and not just entertaining) content is a prime example of where multi-sided fairness considerations are applicable. And these are currently lacking, with the vast majority of research concentrating only on a single-sided perspective (Deldjoo et al., 2024).

Second, they highlight that some recommendations have potential societal impact. They use the example of news recommenders and social media recommenders where recommendations are not necessarily benign. News recommenders for example have high potential for societal impact, gatekeeping our consumption of information and informing how we think (S. Y. Lee & Lee, 2023; Nechushtai & Lewis, 2019). The same logic stretches to mental health recommendations. As apps can be potentially harmful due to their content (Anthes, 2016; Torous, 2018), so can their recommendations. And as described in Chapter 1, whilst important at the individual level, this also has a widespread societal impact.

This ties into how we research recommender systems in mental health. When we consider the effects of a recommender system, how do we evaluate them? Deldjoo et al. (2024) highlight that often it is quite abstract and technical. The outcome focus tends to be algorithmic, especially in computer science papers. Furthermore, most studies evaluated their recommender systems *offline*, not in real-time and without consistent human interaction. They argue that this can potentially lead to reduced generalizability and practical impact.

Whilst Deldjoo et al. (2024) do not consider mental health applications at all, many of the issues they explore are relevant to the field. However, there are also certain facets of fairness that they do not explore. This is because user wellbeing was at best a secondary concern in the domains they included in the review. Tourism, e-commerce, video recommendations, all of these things can be argued to, situationally, improve people's quality of life. At least, it is likely that it will be argued in some of the mission statements from the companies who develop them. But it is still qualitatively different from domains where wellbeing is the goal. To put this into perspective, there exist recommender systems in the health domain. One review shines a light on the paradigm difference here (Tran, Felfernig, Trattner, & Holzinger, 2021). The authors systematically reviewed 37 health recommender system studies, with a specific focus on the context in which the health recommender systems were applied. It drew attention to the fact that these have more stringent requirements, for example that allergies and existing medication had to be considered for medicine recommenders. Furthermore, recommending items based on user preferences could often be counterproductive to their health; in a diet recommender context, it is often seen that user preferences are what lead to an unhealthy lifestyle. The contrast can be even more clearly demonstrated - imagine a medicinal recommendation system that suggested the drugs people were most likely to enjoy. This stresses the need in health recommenders to find a balance between user satisfaction and healthiness, a balance that is highly dependent on the recommendation scenario.

Although the adverse consequences of improper recommendations are usually not quite as severe for mental health recommenders as medicinal recommenders, they exist within a similar domain, one where improvement is a primary goal. And it is a unclear *how* to best determine the appropriate balance between improvement and engagement. No matter how good a mental health app is at improving user mental health, it is to no avail if they do not use it. Similarly, it matters little how engaged a user is with the app if it does not help them.

Even broadly, mental health recommender systems are relatively niche, a largely unexplored field. In the domain of health recommender systems reviews commonly do not encounter or do not consider mental health settings (Pincay, Teran, & Portmann, 2019; Tran et al., 2021). Even in one of the largest recent reviews that did, only 6 of 73 articles reviewed were related to mental health (De Croon et al., 2021). And although one conceptual review exists (Valentine, D'Alfonso, & Lederman, 2023), there are no systematic reviews specifically for the field of mental health app recommender systems. In the following sections, I will report the findings of my own systematic review that investigated what the current state of the literature is on mental health recommender systems. It focused on how recent studies are designed and evaluated, how variables were defined, how much of the focus was psychological and how much was algorithmic. It investigated how algorithms were tested, whether through real-time interaction with humans, or offline. It asked, throughout, what the purpose of current research is, who it tries to help, how theoretical or applied it is, and how research methodologies reflect the aims of researchers or developers. It will also lay more specific groundwork for the rest of the thesis. Up until this point the focus has been broad, aiming to clarify the context of mental health apps and mental health recommendations. This next section will look at this niche field specifically. It will still maintain some breadth of perspective, as the aim is not to evaluate the efficacy of any specific solutions. However, it will attempt to identify more specific issues and gaps in research on mental health recommendations which later chapters will focus on filling.

2.2 Methods

2.2.1 Scopus literature search

A single search engine (Scopus) was used to scour multiple databases for related literature using PRISMA guidelines (Page et al., 2021). The final goal was to identify a set of recent studies which implemented and evaluated recommender systems within a mental health setting. To this end the following query was constructed and entered into the Scopus search engine:

("mental health" OR "e-health" OR "wellbeing") AND ("recommender system" OR "recommender" OR "personalization" OR "personalisation") AND ("rct" OR "intervention study" OR "ab test" OR "evaluation" OR "offline evaluation" OR "online evaluation") AND (LIMIT-TO (

PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) AND (LIMIT-TO (LANGUAGE , "english")) AND (LIMIT-TO (SRCTYPE , "j") OR LIMIT-TO (SRCTYPE , "p"))

The search word combinations fell into three categories. The first was designed to reduce the scope of the search to studies involving mental health or digital health. The second was designed to ensure the inclusion of a form of personalization or recommendation. The third was aimed to include only studies that evaluated recommenders with either machine learning or clinical goals in mind. This search was performed on the 10th of May 2022, and yielded a total of 1947 journal or conference articles from 2020 and onwards. The titles and abstracts of these articles were then read through, and discarded if they did not fulfill these first criteria:

1. They must mention a system of recommending mental wellbeing or mental health solutions.
2. Solutions were not delivered by robot or chatbot.
3. They must not be systematic reviews of any kind.

These first three criteria were meant to further narrow down the studies presented from the Scopus search. The second was added seeing as robot and chatbot mental health aids typically present a very different paradigm than is of interest in this study and the thesis as a whole. The third criterion was deemed necessary to filter out any studies that dealt with reviews or meta-analyses of other studies or recommender systems. This study was only concerned with the first-hand testing and presentation of recommender systems, and especially how the original studies chose to evaluate and present their results. This left a remainder of 77 studies. Of these, one was not accessible to the author. The remaining 76 were combed through, with the full articles being read, and retained if they met both the previously mentioned and following criteria (some were only identifiable as review articles from the full texts):

4. They must implement at least one kind of recommender system.
5. Recommender systems must be used explicitly to recommend mental health or wellbeing solutions (or test recommendations offline).
6. Recommendations must be for solutions that can be managed solely by the user (ie. Not requiring or recommending external therapies).

These criteria further narrowed down the list of studies for review, ending up with a list of 10 studies with the desired paradigms to be compared. These provided personalized solutions for improving user mental health, or tested recommender systems offline. The most important factor was that a recommender system targeting mental health was included as part of the study methodology, but the manner in which this was done was not critical. This led to relatively

lenient inclusion criteria in some respects, and a very diverse sample of studies. This was a deliberate choice, seeing as this review is interested in how researchers choose to implement and evaluate these recommender systems across the field. The interdisciplinary nature of the field means that researchers with very different methodological backgrounds can engage with the subject matter and design experiments. Part of the aim of this study was to evaluate the congruence (or lack thereof) between how results are reported in the same field.

2.2.2 Evaluation and comparison

Throughout the comparison, the following five questions were used as focal points:

1. What was the main focus of the study?
2. Was the study offline?
3. What type of solutions were recommended?
4. What type of recommender system was implemented? Was it described in detail?
5. How did researchers choose to evaluate the recommender systems or the outcomes from them?

As mentioned, the focus is clearly broad, to include any study with implemented recommender systems in mental health scenarios. The analysis was meant to provide an overview of all approaches within the field, and the extent to which they leaned towards computer science and psychology. Through these questions, I also aimed to evaluate the balance of value gained for users and developers.

2.3 Results

Figure 2.1 shows a PRISMA flow diagram detailing the selection of the final 10 studies. The Scopus search engine filtered out all duplicates, so no records ended up being removed before screening. The abstracts of 1947 published journal articles or conference papers were read, and 1870 were excluded for further analysis due to not meeting the abstract criteria described in the methods section. Of the 77 remaining studies, one could not be retrieved. The full papers were read for the remaining 76, where 66 were excluded for not meeting the further criteria, leaving a total of 10 studies for final review.

Prisma Diagram

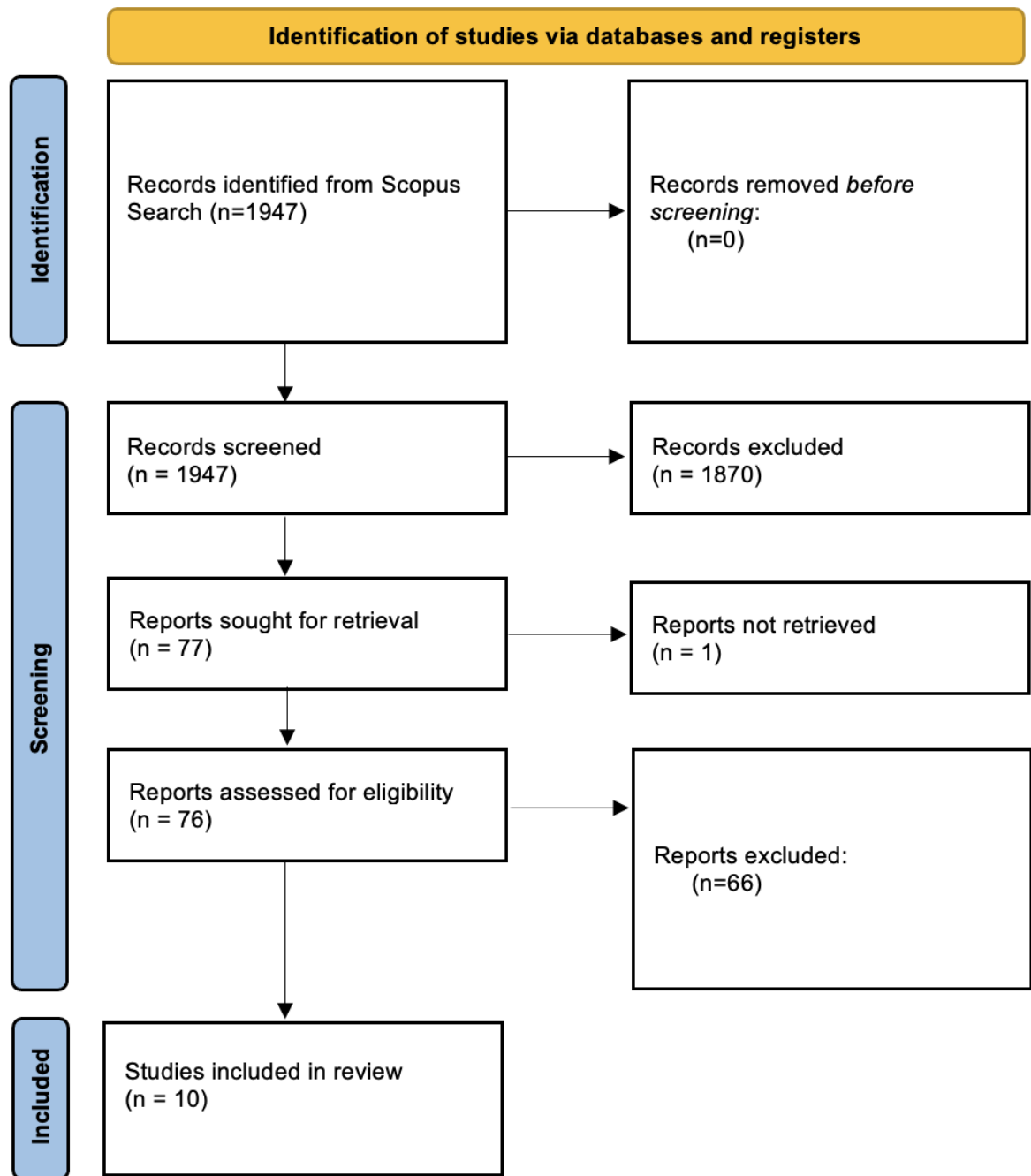


Figure 2.1: Prisma Diagram of filtering procedure.

The key findings of this review can be seen in table 2.1, which includes the main focus of each study, whether it was done offline (meaning it would not include real-time trial and error testing of the recommender system), which algorithms were used, and how both the algorithms and the mental health outcomes were evaluated. There are several notable outcomes of the review. The first is the wide focus. All of the studies recommended some form of item or activity

Table 2.1: Key findings from the systematic review.

Study	Offline?	Main focus	RS description	Evaluation (RS)	Evaluation (long-term mental health)
Ameko et al. (2020)	Yes	Comparison of contextual bandits in emotion regulation strategies	Three types of contextual bandits, Direct Method (DM), Doubly Robust (DR) and Offset Tree (OT) across different weightings of Inverse Propensity Scores. These were compared to a random baseline and actual observed scores.	Mean estimated rewards (denoted as usefulness of a strategy) was compared across recommender systems	None in practice, however the reward levels of recommendations were estimated
Beltzer et al. (2022)	Yes	Comparison of recommender systems when emotion regulation strategies were grouped together by category versus considering them individually	DR method for category approach, DM for individual approach	Mean estimated rewards (denoted as usefulness of a strategy) was compared across recommender systems	None in practice, however the reward levels of recommendations were estimated
Ferré-Bergadà et al. (2021)	Yes	Using multi-attribute utility theory for priority rankings of a broad set of mental health exercises	Marginal Utility Algorithm	None	None
Lewis et al. (2022)	Yes	Comparison of different collaborative filtering and context aware recommender system techniques in identifying enjoyable therapy tasks	K-Nearest Neighbours (KNN), KNN with baseline adjustment, Singular Value Decomposition (SVD), SVD++, Factorization Machine (FM) with context, Field aware FM with context	Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), across multiple experiments to account for random initialization	None
Mayer et al. (2022)	No	Acceptance and viability of SELFPASS app which includes digitalized mental health interventions (mostly CBT based)	Knowledge-based recommendations (algorithm not detailed)	None	None, but perceived quality, practicality and acceptance were measured
Niles and O'Donovan (2018)	No	Attention bias modification for individuals with PTSD with personalized word selection	KNN*	None**	PTSD Checklist for DSM-5, State-Trait Anxiety Inventory Trait version and the Depression Anxiety and Stress Scale
Rohani et al. (2020)	No	Longitudinal feasibility study to test a behavioral activation app for depressive patients	Multinomial Naive Bayes	None	Qualitative assessment, however it was largely general assessment of the app and acceptance of it, with little evaluation on the specific mental health benefits of the recommendations.
Rohani, Springer, Hollis, Bardram, and Whittaker (2021)	Yes	Comparing recommender system model effectiveness in predicting moods to recommend mental health activities with the highest personalized impact	Multinomial Naive Bayes and Support Vector Machine content-based recommender systems	Algorithms were evaluated by holdout cross-validation with incremented training set size, and mean test errors were reported. Additionally, F1-score, precision and recall metrics were included for comparison	None
Suzuki et al. (2021)	Partially***	Evaluation of a recommender system which recommends physical locations based on google maps which may improve user happiness and reduce stress	The algorithm was built on input of multiple scores, using on word2vec, collaborative filtering and random forest algorithms	Normalized Discounted Cumulative Gain	None
Torkamaan and Ziegler (2022)	Study 1 yes, study 2 no	Evaluating the use of gaming matchmaking algorithms in recommending stress reduction activities	ELO, TrueSkill, Glicko2, Rasch Model	The algorithms were compared by the processing speed and certain inherent characteristics, such as the ability to factor in time-decay or initial required effort	None

*The recommender system was not described in this article, however a detailed description and comparison was done in Niles and O'Donovan (2018).

**The recommender system algorithm was not evaluated or compared on any common machine learning parameters here, but RMSE and precision were used in thorough algorithm comparison in Niles and O'Donovan (2018).

*** The recommender system was largely tested offline, but 9 users did evaluate it. However, they did not actually visit the proposed locations, only estimated rankings, and as such the true effect of the recommender system was not measured.

designed towards improving mental health, but that is where the similarities ended. Some had an algorithmic primary focus, some psychological, and the solutions they recommended had little overlap, ranging from stress reduction activities (Torkamaan & Ziegler, 2022) to physical locations that would lead to user happiness (Suzuki et al., 2021). The target populations also differed, with most sampling randomly from the general population, but few were also directly aimed at improving outcomes for specific groups such as individuals suffering from PTSD (Niles et al., 2020), depression (Mayer et al., 2022) or professional caregivers (Ferré-Bergadà et al., 2021).

The second item of note is how early stage a lot of this research seems to be. Of the ten studies, five are fully offline, with two more that are mostly done offline, with a proof-of-concept online component. The final three were mostly focused on mental health or acceptance outcomes, with little evaluation of the algorithms themselves.

There is also a clear pattern of compartmentalization of field-specific methodologies, where studies predominantly had a focus on either computer science or psychology. This can be seen quite clearly from table 2.1, where no study evaluated both machine learning outcomes *and* some form of mental health outcome. Interestingly, in two cases the researchers behind an app or a recommender system did evaluate across both parameters, yet they did so in separate studies. In order to evaluate algorithms, Niles and O'Donovan (2018) conducted an offline study. In the more recent one (Niles et al., 2020), they conducted an online study, yet the focus had shifted entirely to the evaluation of mental health outcomes. Similarly, two studies by largely the same group of researchers, researching the same mental health app (Rohani et al., 2020, 2021) likewise separated the methodologies. Rohani et al. (2020) investigated mostly app acceptance and qualitative outcomes in a longitudinal study, whereas Rohani et al. (2021) conducted an offline study to compare algorithms. It is, however, important to note the distinction between mood and long-term mental health outcomes. Three studies (Beltzer et al., 2022; Lewis et al., 2022; Rohani et al., 2021) included mood or state ratings as inputs, and by extension, in evaluating their algorithms they considered how effectively their recommendations could improve user moods. It could be argued that this qualifies as a mental health evaluation, however, unless a relationship was established with trait-level mental health outcomes, we did not do so in this review. Chapter 5 will further explore the relationship between short-term and long-term outcomes.

Of further note is the broad variety of recommender systems. Content-based recommenders, collaborative filters, knowledge based recommenders and hybrid recommenders are all represented. With the exception of the two studies by the same group of researchers (Rohani et al., 2020, 2021), no two studies utilized the same recommendation techniques. In the studies that included algorithmic evaluations, the methods and metrics were also different. Whilst the most common industry standards are represented in the studies (precision, recall, f1, RMSE) multiple other methods were used as well. Similarly, for those that evaluated mental health outcomes or

acceptance, no outcome metric was used in more than one study.

Finally, the level of detail in describing implemented recommender systems is also widely different. Several of the studies went into great detail, including complete descriptions of algorithms. In several cases the code for replication was also directly available (Beltzer et al., 2022; Niles et al., 2020; Rohani et al., 2021) or stated as available upon request (Ferré-Bergadà et al., 2021; Torkamaan & Ziegler, 2022). In contrast, some papers, especially those with a psychological focus, had less of a focus on the actual recommendation techniques. For example, Mayer et al. (2022) had very little mention of the mechanics by which they recommended mental health solutions. Only by reading between the lines, can it be gleaned that a knowledge-based recommender system was used. Replication would of course be nigh-impossible in this case.

2.4 Discussion

The main takeaway from this review is that in almost all of the studies within one niche field over the course of over two years were very different. In almost all aspects. To sum up the results in a single sentence, the mental health issues they tackled were different, the solutions offered were different, the methodologies were different, the recommender systems were different and the evaluation metrics were different. They described recommender systems for PTSD, depression, stress reduction and general mental wellbeing. They recommended stress reduction techniques, CBT techniques, emotion regulation techniques and physical locations. No two algorithms between research groups were the same, the evaluations tended to be either based in psychology or machine learning but not both, and the level of detail in describing the recommender systems varied widely.

The one thing that was relatively consistent between the studies were that they marked the early stages of research for their respective recommender systems. All of them described relatively recently developed apps or techniques, and most were the first papers detailing the specific interventions they focused on. They all called for further testing of their specific solutions and the ones with an algorithmic focus were mostly offline. Whilst offline evaluations of recommender systems are by far the most common in research settings across the broader field (Deldjoo et al., 2024), it is still notable because offline evaluations often do not reflect real applications, and regularly contradict findings in similar online studies (Beel, Genzmehr, Langer, Nürnberger, & Gipp, 2013; Beel & Langer, 2015; Rossetti, Stella, & Zanker, 2016). Offline evaluations can be useful, but as Deldjoo et al. (2024) pointed out, research on recommender systems without a human in the loop leaves a major gap in our applicable knowledge. This is why they tend to be the first stage of recommender system research, and indeed several of the articles recommended testing in a clinical setting or with RCTs, noting the offline paradigm as a limitation (Ameko et al., 2020; Ferré-Bergadà et al., 2021; Lewis et al., 2022; Rohani et al., 2021; Suzuki et al.,

2021).

However, something that contrasts strongly with the notion that the studies in this review should be considered early stage research was how technically or methodologically complex some of them were. Whether it was the sophisticated evaluations of algorithms seen by Torkamaan and Ziegler (2022) or the strong foundations in theoretical psychology in the study by Rohani et al. (2020), they do not have the appearance of research in uncharted territory. And that is because they draw from a vast repository of knowledge and established tools in different fields of psychology and computer science. Unfortunately, however, as we have already explored several times in this thesis, these supporting fields and established practices carry with them established habits and established issues. Chapter 1 described tendencies in mental health to focus on innovation and novel solutions rather than underlying mechanics or repeated testing (Wampold, 2019), and earlier in Chapter 2 the same thing was established in the broader field of recommender systems (Portugal et al., 2018). This review found the same pattern, with every study on a mental health recommender system focusing on testing a novel intervention or novel application of an algorithm, with mechanisms of change being at most a secondary consideration.

This is not unexpected, and once more it must be emphasized that innovation and novelty are not inherently negative things to pursue, only that there is an opportunity cost. The consequences are the same as highlighted in other scenarios. Having a variety of interventions or recommender systems to choose from makes it more difficult to identify the best ones, much as it does in mental health apps in general (Neary & Schueller, 2018). This difficulty is compounded when the evaluation metrics used in different studies also tend to vary. Whilst the most common industry standards for evaluating machine learning outcomes are represented in the studies in this review (precision, recall, f1, RMSE) multiple other methods were used as well. And these are only for directly evaluating algorithms. When including the diversity of methods for evaluating mental health, engagement or even acceptance outcomes, the techniques are even more diverse, and similar concepts have different operational definitions. There are certainly merits and justifications for each of the chosen methodologies, however it does make comparison across studies very difficult, making it more likely that generalizability concerns will be inherited from other fields.

These concerns can be mitigated by open research practices. Sharing data or code can improve robustness, reliability and understanding of generalizability in psychology, data science and machine learning contexts (McDermott et al., 2019; Nosek et al., 2022; Yu & Hu, 2019). The lack of these open science practices has been identified as an issue in recommender systems (Deldjoo et al., 2024), where by far the majority of studies include no code or data. In this review this was not the case, with half of the included articles either including data directly or specifying it as available upon request. Not sharing data, or in some cases such as in the study

by Mayer et al. (2022) not including a description of how items were recommended, does not necessarily detract from the rest of the content of the study. However, as there is a huge spectrum of personalization techniques and algorithms available, this lack of detail or reproducible code can make it very difficult to estimate the quality and value of recommendations.

The review also highlighted that more work is needed to bind the fields of computer science and psychology together. Several of the included studies provided excellent insights into the mechanics of their chosen research topic, yet few bridged the interdisciplinary gap completely in a single piece of work. There was a tendency to separate psychological and machine learning research and outcome evaluations, even into separate papers (Niles & O'Donovan, 2018; Niles et al., 2020; Rohani et al., 2020, 2021). Perhaps the most interesting example of this separation is seen in the study by Torkamaan and Ziegler (2022). Of all the studies reviewed, they had the most in-depth analysis of engagement outcomes, and a clear algorithmic focus. However, they also clearly defined their goals in a mental health context, laid a strong foundation for their work in theoretical psychology and consistently bound together the interdisciplinary fields conceptually. However, despite the merged conceptual approach, there was no actual evaluation of mental health outcomes. The algorithms described were used in the recommendation of items designed to reduce stress, yet no stress related outcomes were reported. In a sense, they partially avoided the tendency described by Deldjoo et al. (2024) for work in computer science to have an overly narrow or technical scope, as they grounded all their technical results in a mental health context, yet the study still missed the final step of psychological evaluation to thoroughly merge the different disciplines.

There are different ways to do this, and much can be incorporated within the design of the recommender systems. It must be considered what the recommender systems are actually trying to recommend, what is the input and output. This brings us back to the assumption described in Chapter 1, that mental health recommenders will recommend activities that are useful to users in improving mental health. It can be seen that academic studies of recommender systems tend to have a focus biased towards the user experience (Abdollahpouri et al., 2020; Jannach & Jugovac, 2019), but in this review we can see that even within academic publications, that focus can manifest differently. Whilst none had a commercial focus, the inputs varied for the recommender systems. Lewis et al. (2022) built an algorithm to recommend enjoyable therapy tasks. Similarly, Rohani et al. (2021) based recommendations on the prediction of their effect on user moods. Torkamaan and Ziegler (2022) designed theirs to estimate levels of required effort and user ability, and recommend items based on the optimal pairings. Mayer et al. (2022) used a knowledge based recommender system, which used psychological theory to recommend items best suited to the classifications of the mental health issues users had. Suzuki et al. (2021) recommended locations likely to make the user happiest.

These are clear examples of how recommender systems in mental health contexts can, but not

necessarily will, incorporate user mental health variables directly in the algorithms. If they do not, then user improvement relies on the assumption that the solutions themselves *will* have a positive effect on user mental health. This is not always an implicit assumption. Lewis et al. (2022) and Rohani et al. (2021) used inputs that directly pertained to user moods, likert ratings where users could indicate how they felt after engaging with a therapy task. This is an excellent example of how expressions of short-term mental health input can be used in recommender systems, however they did not evaluate any longer-term mental health outcomes. Despite this, Lewis et al. (2022) made the explicit assumption that these would improve long term mental health benefits. In their introduction they stated:

"...an RS algorithm could instead be optimised to recommend therapy tasks to clients that might help them to feel happiest, most productive, or most relaxed given their current context. On finding such tasks, a client is more likely to improve their well-being in the short-term and persist with the treatment in the long-term, thus increasing the likelihood that they recover to a state of better mental health."

As there does not appear to be any evidence for this for this in the field of recommender systems yet, this is an untested assumption. We do not currently know the relationship between short-term and long-term improvement from mental health apps, or what the effects of retention are on mental health. The use of user states as inputs in algorithms by Beltzer et al. (2022), Lewis et al. (2022) and Rohani et al. (2021) is a step in the right direction for interdisciplinary research, yet their evaluations still ended up focusing largely on machine learning metrics, once more highlighting the disparity between fields. It also highlights the need for an understanding of mechanisms of change, because if relationships are assumed and not tested we will remain in the dark regarding *which* of our solutions are most effective for *which* individuals, or what the relationship is with engagement metrics.

The tone of this review has perhaps seemed relatively negative. This is because there are several issues in the field, some of which have carried over from research practices in different areas. Yet this should not be taken as a poor reflection of the quality of individual research. The studies themselves all contributed in a meaningful way, and no study can address all aspects of a focus area. However, when considered from a birds eye perspective, this review highlighted both a sparsity and lack of congruency in research surrounding recommender systems in mental health settings. This is perhaps no surprise, as it is a relatively new application from an older (yet still young) research field. And as research on recommender systems in general increases, the specific application of mental health will follow, and the sparsity issue will likely resolve itself given time.

However, time has not solved the other issues highlighted in this review. Psychologists have been aware of replication and generalizability issues for decades, and still the problem remains.

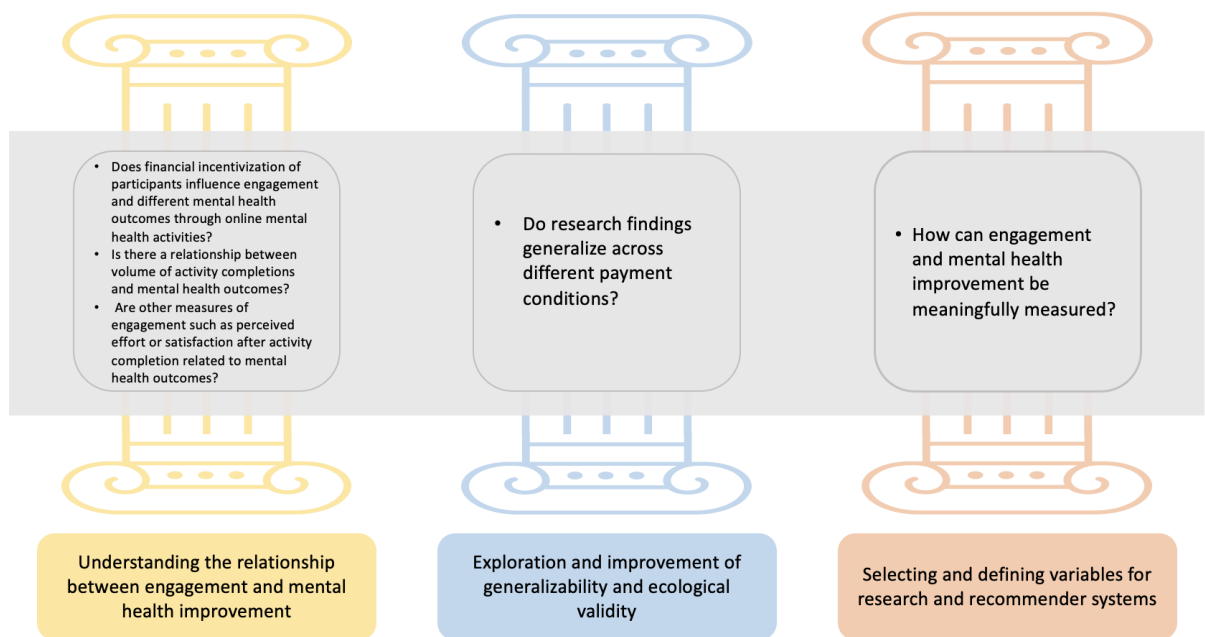
It is a tendency in both the parent fields of psychology and computer science to neglect in depth research of solutions in favor of developing new ones. The research is also potentially limited quite naturally by the setting most of it is conducted in. Namely, the offline nature of much of the research may be due to the inherent difficulty of researching recommender systems in academia. Where industry has a constant influx of new users, academic studies rarely do. Since the field demands a high number of users, or at the very least a high number of entries per user, preferably over time, this can be costly. Yet academic-style studies are often more effective at separating confounds and isolating variables.

The purpose of this review, in terms of the PhD project, was to establish what the existing literature was on mental health recommender systems. Originally it was intended as a starting point for understanding the interplay between user engagement and user improvement. It was intended to build towards an end-goal of designing recommender systems that balanced user needs with commercial needs and generate frameworks for research that combined theoretical interests with ecological validity. However, by developing a greater understanding of the field, the purpose shifted. My background in research is primarily psychology, with knowledge of computer science being picked up later. Rushing towards the goal of building new algorithms would simply make me one more researcher in the headlong race towards innovation and new solutions. Worse, it would likely leave me gasping for breath in the tail-end, because another thing this review revealed was that there are people far more qualified than I, with far more technical expertise in the field of machine learning, developing these solutions.

The research instead turned to different questions, highlighted by the gaps now clear in the research. What is the relationship between user engagement and user mental health improvement? What variables should be included in these recommender systems, based on their potential for increasing engagement and mental health improvement? How do we measure these variables? How much difference is there between individual activities in improving mental health? How can we design studies to evaluate mental health apps or recommender systems with a focus on usability and ecological validity? Do academic findings reflect those from data gathered in the wild? These are all questions that the rest of this thesis will attempt to answer through collected data.

Chapter 3

Financial incentivization and its implications



3.1 Introduction

3.1.1 The consequences of compensation

This chapter begins the second part of this thesis. The mental health context is hopefully clear to readers, as is how personalizing ubiquitous mental health solutions fit into this context. We have explored how we have advanced tools at our disposal in a relatively novel field, how technology is far outpacing science, and how this has led to gaps in our fundamental understanding of recommender systems in mental health. One area that may be particularly important, is understand-

ing the nature of the relationship between user engagement and mental health improvement, and how our research methods can influence these outcomes.

The first step was to investigate one prevalent way of incentivizing participation in research in academia. Money. The effect of payment on participation has been long studied in research due to both ethical and experimental concerns (Bentley & Thacker, 2004). The ethical side of the debate largely focuses around the potential coercive effects of financial incentives (Grady, 2019; Largent & Lynch, 2017), and is more commonly explored than the experimental side. However, there have been some studies which attempted to uncover the confounding influence on results. These tend to focus mostly around data quality, whether payment incentivizes users to take responses more seriously, or sampling considerations, how many participants, and what type of participants payment attracts. Generally it is this latter factor that appears to be most influenced by monetary incentives. Unsurprisingly, compensating participants for their time improves collection rates, but an equally notable benefit is that it potentially increases the diversity of samples (Hsieh & Kocielnik, 2016; More, 2022). On a further positive note, there does not seem to be an influence on data quality stemming from payments, although individual and group differences in participants can have an effect (More, 2022).

Yet the question of the effect of financial incentivization remains largely unexplored in research on recommender systems. This is an oversight. Where increased collection rates due to participant payment is described as a positive in the cases where it is explored (Hsieh & Kocielnik, 2016; More, 2022), this is not necessarily the case in this context. One of the core goals of recommender systems in commercial settings is to improve user engagement and retention. They certainly bring qualitative benefits to the users, but their primary value in business settings is that they increase retention or purchases, and in turn revenue (Abdollahpouri et al., 2020; Jannach & Jugovac, 2019).

Here there is a clear divergence between the academic and commercial settings. People pay for access to mental health apps in the wild. Sometimes through direct purchases or subscriptions and sometimes through indirect methods such as through in-app advertisements. In academia it is often the opposite - research participants are paid to engage with materials, or incentivized in some other way. This illustrates the divide described in Chapter 2, where academic research tends to have a biased focus on user experience (Abdollahpouri et al., 2020; Jannach & Jugovac, 2019). Payment may or may not influence this aspect of recommender systems, yet it almost certainly influences the extent to which participants engage with the content of a study (Hsieh & Kocielnik, 2016; More, 2022). Therefore studies that pay people to engage with recommendations have a diminished ability to estimate its value in a commercial setting.

The extent to which studies on mental health recommender systems address this issue differs. In the studies included in the review in Chapter 2, most did not describe how participants were

compensated. One compensated participants with a gift voucher upon completion of the study (Rohani et al., 2020), one had an incremental compensation plan for the different stages of the study (Niles et al., 2020) and two described datasets collected 'in the wild', meaning from a sample of users naturally engaging with the recommender systems outside of academic settings (Rohani et al., 2021; Torkamaan & Ziegler, 2022). The latter studies, those sampling naturally from the population, automatically avoid the issues related to financial incentivization. However, for the those that did in some way compensate participants, this is a potential confounding influence on engagement outcomes. In fact, monetary compensation is likely the most influential predictor of engagement in research (Linardon & Fuller-Tyszkiewicz, 2020). Whilst engagement outcomes were not of interest in all of the studies, it remains important. In part because it is likely a priority for any companies looking to utilize the algorithms, but also because a beneficial recommendation or activity only works if people engage with it.

And it is not a given that users will engage with app content. Low user retention and compliance is a pervasive issue in the research of mental health apps. Even in commercial settings, numbers are shocking. Approximately 70% of users are lost within the first week of downloading a new mental health app (Sigg, Lagerspetz, Peltonen, Nurmi, & Tarkoma, 2016), and only somewhere between 13.1% to 23.3% of users are still active past the 28-day mark (Lattie et al., 2016). These numbers, whilst concerning, are offset by a number of factors in industry. Generally, in commercial settings, though high attrition rates may be an issue, there will typically be ongoing marketing and onboarding efforts to land new users to compensate for those leaving. Furthermore, depending on the business plan, a company may well accept user losses as long as a core of active users remain.

However, academic research has different requirements, and equivalent dropout rates in mental health app studies pose a serious problem. Usually these studies have a set start and end date or minimum number of users, and each lost participant either means a poorer final dataset, or extra time and resources spent on rebuilding sufficient sample sizes. This is why the academic research paradigm is inherently different. Whereas for most companies attracting and retaining users is the primary purpose, as this is how they generate their income, in academic studies, participants are more akin to employees than customers. They generate value through the data and information they help provide, and, as employees do, they often get compensated. This likely explains why attrition in academic research, although high, is usually considerably lower than in commercial settings. Several reviews have looked at engagement with mental health, and have found that the mean dropout rates ranged from 21% to 24.1% in the short term (Leech, Dorstyn, Taylor, & Li, 2021; Linardon & Fuller-Tyszkiewicz, 2020) and 35.5% in long term follow ups (Linardon & Fuller-Tyszkiewicz, 2020).

This disparity between settings poses several problems. First and foremost, when experimental results differ considerably from what is observed outside of academia, questions of ecological

validity and generalizability arise. If practical applications do not match controlled conditions it can introduce confounds. For example, participants motivated by financial incentive who would not consistently engage with commercial mental health apps may be inherently different to participants who are motivated by internal factors, and if so, results would reflect that. This is not a given, however it is not currently explored in recommender system or mental health app literature, and so remains unknown. Understanding these confounds and individual differences may go part of the way in explaining why results on the effects of engagement on mental health improvement are inconsistent across the literature (Molloy & Anderson, 2021). Other factors are also likely to contribute to the variability in results, and have been better explored, such as the breadth of intervention types (Inkster, Sarda, & Subramanian, 2018) or chosen engagement metrics (Ng, Firth, Minen, & Torous, 2019; Weingarden et al., 2020). However, the choices of if and how to incentivize participants is one every researcher must tackle, and to our knowledge no study has as of yet attempted to isolate and investigate the effect of monetary compensation in mental health recommender systems.

Beyond the considerations of financial incentives for participation, there is also the question of whether engagement in itself has an effect on mental health change. In the broader field of mental health apps, results are unclear. Several systematic reviews have evaluated its role across the literature (Lekkas, Price, McFadden, & Jacobson, 2021; Linardon & Fuller-Tyszkiewicz, 2020; Molloy & Anderson, 2021), and across different fields and types of mental health apps, they make it clear that congruence is low. Custom criteria are often used to evaluate both engagement and mental health outcomes and there is considerable heterogeneity in how studies investigate engagement and its effects, if they even include them in the study at all. Results are also often contradictory, with some studies finding that compliance and app usage were significant predictors for some measure of mental health improvement (Inkster et al., 2018; Lekkas et al., 2021; Schlosser et al., 2017), yet Molloy and Anderson (2021) found that only four of nine studies that evaluated the effect of engagement found that it reduced depressive symptoms. Continuing to investigate this relationship will allow for a better understanding of how types of engagement may influence mental health outcomes, and has a direct application to the field of mental health recommender systems, where engagement is often a core focus, especially in commercial settings.

3.1.2 The current study

The main purpose of the current study was to investigate the effect of financial incentivization on engagement and mental health improvement within mental health apps. To this end a research app was created from a subset of 'mindfulness and meditation' mental health activities included in an existing mental health app by Koa Health called 'Foundations'. The app is commercially available for companies who wish to provide support for employees, and has shown promis-

ing results across a range of mental health issues and populations (Catuara-Solarz et al., 2022; Gnanapragasam et al., 2023; Schulze et al., 2024). Researchers involved with the current study were primarily interested in mental health change across two parameters, mental wellbeing and perceived stress.

Mental wellbeing was of interest for multiple reasons. First, it is one of the main mental health outcomes the parent app Foundations targets, and was a primary focus in two of their trial studies (Catuara-Solarz et al., 2022; Schulze et al., 2024). It is also conceptually one of the most interesting variables for mental health apps, as it is one of the broadest. Individuals with low mental wellbeing are more susceptible to other mental health disorders (van Agteren et al., 2021), and high mental wellbeing can be a strong protective factor against developing other issues (Arango et al., 2021). Furthermore, mindfulness based interventions tend to be some of the most effective in both clinical and non-clinical populations (van Agteren et al., 2021; Zollars et al., 2019). This makes it a good candidate for mental health apps to target, as the ubiquitous nature meshes well with broad issues and broad populations.

Stress was chosen as the second measure in part to address generalizability of findings. Mindfulness and meditation have also been found to be effective tools for combatting stress (Zollars et al., 2019), and a strong link has been established between stress and mental wellbeing (Coffey, 2004; Teh, Archer, Chang, & Chen, 2015). However, stress can also be quite situational and prone to change (Lebois, Hertzog, Slavich, Barrett, & Barsalou, 2016) and has seen variable results from trials with the parent app compared to mental wellbeing (Catuara-Solarz et al., 2022). It is therefore a good starting point to address questions of generalizability by investigating the differences between two mental health outcomes, that are theoretically both improved by mindfulness interventions, interventions developed in a similar style by the same company. It also increases in relevance for Chapter 5, where we consider the state and trait variations in stress and mental wellbeing, considering generalizability from a temporal perspective.

The main purpose of this study was to determine to what extent financial incentivization influenced user engagement (measured through number of completed activities) and mental health change (measured through self-report data) when completing online mental health activities. We were also interested in the relationship between these two variables, whether increased engagement with mental health activities was in itself a predictor of mental health improvement. However, to better understand mechanisms of change, we also collected additional information to see what other factors may have an effect, including subjective engagement ratings, demographic data and post-activity ratings. The research carried out in this study (as in the rest of the thesis) was entirely exploratory.

3.2 Methods

3.2.1 Participants

Eligibility for this study was determined on two parameters. First, participants had to be above 18 years of age. Second, as mental health levels can influence user experience and acceptance of mental health apps (Mayer et al., 2022; Schulze et al., 2024), we decided to only include users who scored below a certain threshold of mental wellbeing. Therefore, 302 users were included in a screening study where they completed two questionnaires (detailed under materials). 80 participants were invited to the main study after screening, and were allocated randomly into the test group or control group (the qualities of which are described in the the 'procedure' subsection). Seventeen of these either withdrew from participation or did not complete all the study components.

Ethics approval for this study was received from the College of Medical Veterinary and Life Sciences at the University of Glasgow (application 200220250).

We first screened a gender-balanced sample of 300 UK adults who completed at least 10 Prolific studies with an approval rate of at least 95%. Second, as mental health levels can influence user experience and acceptance of mental health apps (Mayer et al., 2022; Schulze et al., 2024), we decided to only include users who scored below a certain threshold of mental wellbeing. A total of 98 participants who scored below 50 on the WHO-5 were eligible and therefore invited to participate in the main study. The main study offered 40 places for each of the two randomly assigned experimental conditions, where they had to complete at least one mental health activity and an exit questionnaire. After excluding 17 participants who either did not complete any activities or the exit questionnaires, the final sample included 63 UK adults (24 male; 39 female) with a mean age of 40.54 years (SD = 11.55).

3.2.2 Design

For this between-subjects, longitudinal experiment, participants completed an online study that lasted over a period of 21 days maximum. Prior to study start, participants were randomly assigned to one out of two incentivization groups, either baseline payment or per-activity payment. Participants in the baseline payment condition were offered £10 for completing and rating at least one activity. Participants in the per-activity payment condition were offered £3 for completing and rating the first mindfulness activity as well as £1 each for up to ten additional mindfulness activities. The maximum payment for participants in the per-activity condition was therefore £13. Participants had access to the app for the duration of the study with no upper limit to how many activities they could complete, only a cap to the ones they were paid for.

Regardless of their experimental condition, all participants self-reported their mental wellbe-

ing and perceived stress pre- and post-experiment. During the 21-day online study, all participants had the opportunity to complete and rate online mindfulness activities. In addition to self-reported activity ratings, we collected standard user analytics via the platform. We further collected sociodemographic information.

3.2.3 Procedure

Prior to conducting the main study, we identified eligible participants through running a screening study. Participants were sampled via Prolific and referred to the Qualtrics platform where they completed the screening questionnaires (see, figure 3.1). Participants with a WHO-5 Well-being Index score < 50, indicating reduced mental wellbeing, were randomly split into the two experimental groups and subsequently invited for participating in the 21-day online experiment via Prolific.

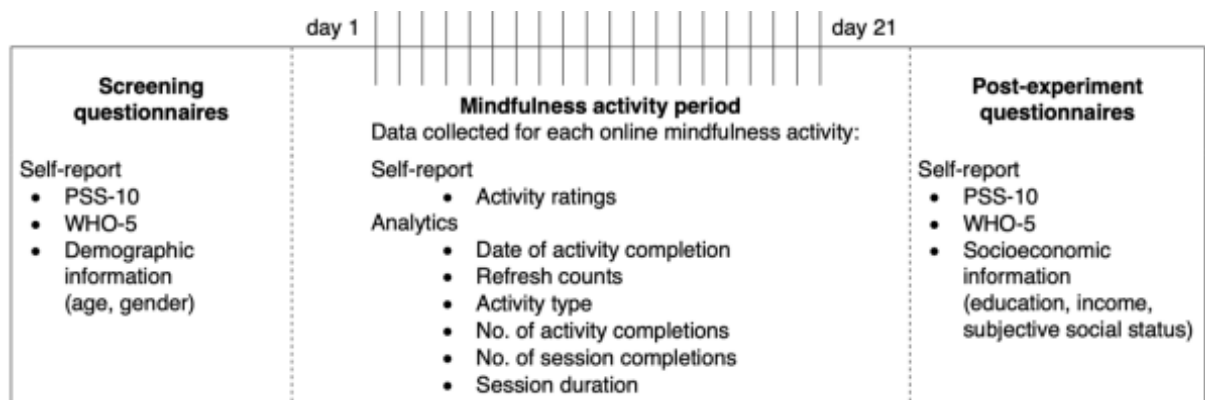


Figure 3.1: Procedure for data collection in Chapter 3.

Individuals who agreed to participate in the main study provided written informed consent and basic demographic information on Qualtrics. Participants then received access to the online platform (implemented in *shiny r*) hosting the mindfulness activities for the duration of the study. After logging onto the platform with their Prolific ID, participants were presented with a choice of three online mindfulness activities at a time, and with a button to refresh the choice of presented activities. Each activity was presented with its title and approximate length in minutes. Once participants chose an activity, they listened to it from start to finish without being able to skip parts or increase the playback speed. Upon completing the online mindfulness activity, participants provided their ratings and were guided back to the home screen where they could choose to complete another mindfulness activity presented, or to click the refresh button. Activities were sampled with replacement so that participants could complete them as many times as they liked.

During the mindfulness activity period, participants received biweekly reminders on Prolific about study participation and reimbursement regulations. On day 21, participants completed the

post-experiment questionnaires; all participants were reimbursed and thanked for their participation. After day 21, their access to the online platform was removed.

3.2.4 Materials

Questionnaires

The WHO-5

The 5-item Wellbeing Index of the World Health Organization (WHO-5) assesses mental wellbeing in individuals over the past two weeks. Participants rated items such as “Over the last two weeks, I have felt cheerful and in good spirits” on continuous self-report scales with one decimal point precision, ranging from 0 (at no time) to 5 (all of the time). The WHO-5 has shown satisfactory reliability across nationally representative samples of 35 European countries (Sischka, Costa, Steffgen, & Schmidt, 2020) and is a sensitive and specific screening tool for depression (Topp, Østergaard, Søndergaard, & Bech, 2015).

The PSS-10

The 10-item Perceived Stress Scale (PSS-10; Cohen, Kamarck, and Mermelstein (1983)) assesses how much distress individuals experienced in the last month. Participants rated items such as “In the last month, how often have you been upset because of something that happened unexpectedly?” on continuous self-report scales with one decimal point precision, ranging from 0 (never) to 4 (very often). As gold standard in the field, the PSS-10 showed acceptable internal consistency and test-retest reliability across various studies (E.-H. Lee, 2012).

Note: It is important to remember when reading results that scores on the two scale are scored in opposite directions. Higher WHO-5 scores indicate higher mental wellbeing, whereas higher PSS-10 scores indicate higher stress. For the discussion these will be referred to generically and should not be an issue, however the results section may be confusing if this is not taken into account.

Socioeconomic characteristics

Participants provided socioeconomic information on their highest educational level, annual income, and subjective social status. Participants self-reported their years spent in formal education ranging from 0 to 20 years or more. Participants categorised their annual income into a £10,000 increment, ranging from “below £10,000” to “above £50,001”. Participants lastly reported their subjective social status on the MacArthur Scale (Adler, Epel, Castellazzo, & Ickovics, 2000). This measure consists of a 10-rung ladder representing where people stand in the United Kingdom, with a higher rung indicating a higher subjective social status. In a multi-ethnic sample, the MacArthur Scale exhibited adequate test-retest reliability (Operario, Adler,

& Williams, 2004).

Activities and in-app ratings

The web app was developed in Rstudio, using R version 4.3.1 using the R “Shiny” package. It was deployed via <https://www.shinyapps.io>. Once participants were given the link to the platform, they were required to log in using their prolific ID each time they accessed the application. When participants completed activities and submitted ratings, the session data would be sent to a password protected google sheets document which could only be accessed by the researchers in this study.

Once users accessed and logged in to the app, they would be sent to a page showing 3 non-identical recommendations for mental health activities (see figure 3.2). These recommendations were randomly sampled, and a new set could be generated (with replacement) by clicking the refresh button. We were not interested in any specific algorithm, so recommendations remained entirely random. Participants could select an activity by clicking an icon, after which they would be redirected to the relevant activity page.

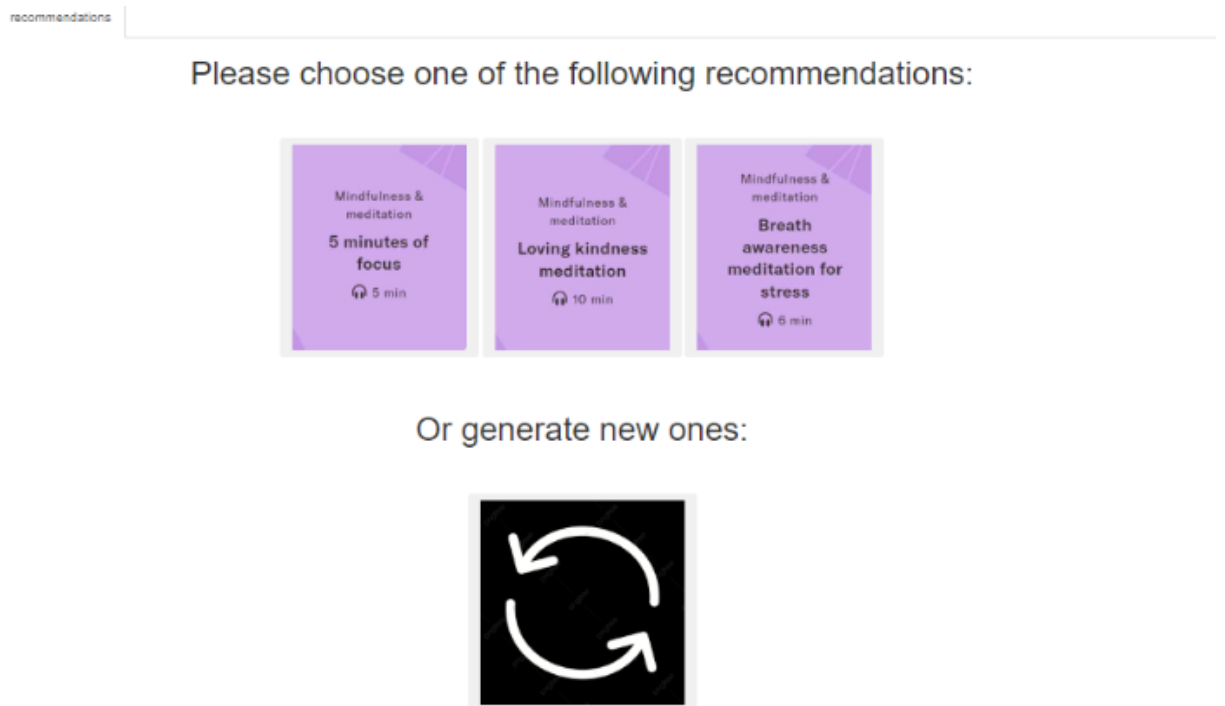


Figure 3.2: Screenshot of the recommendations page for the app described in Chapter 3.

There were a total of 21 different online audio mindfulness and meditation activities, extracted from the Foundations app developed by Koa Health. Mindfulness activities were presented in audio format and targeted various topics, such as self-appreciation, mindful breathing, and meditations. Mindfulness activities varied in length, advertised as ranging from 4 to 11 minutes. For an example of how the activity screen looked, see figure 3.3.

After they were finished listening to the activity, participants also provided six ratings for each mindfulness activity they completed. Ratings concerned participants themselves (i.e., wellbeing, stress), the content of the activity (i.e., enjoyment, meaningfulness), and subjective engagement (i.e., perceived effort, intention to repeat the activity). Activity ratings were scored on continuous self-report scales with one decimal point precision, ranging from 0 (not at all) to 10 (extremely) with slightly adjusted scale labels for the item on perceived effort. The rating order was randomized each time. See figures 3.4 and 3.5 for how the ratings appeared in the app.

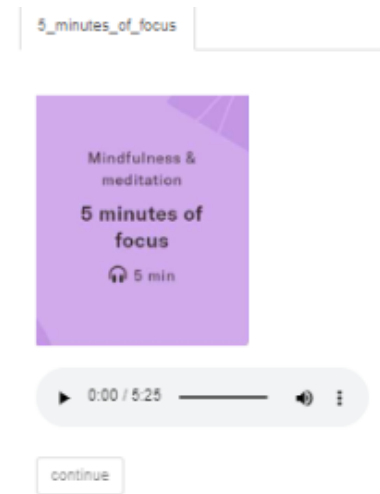


Figure 3.3: Example of a screenshot of the activity page for the app described in Chapter 3.

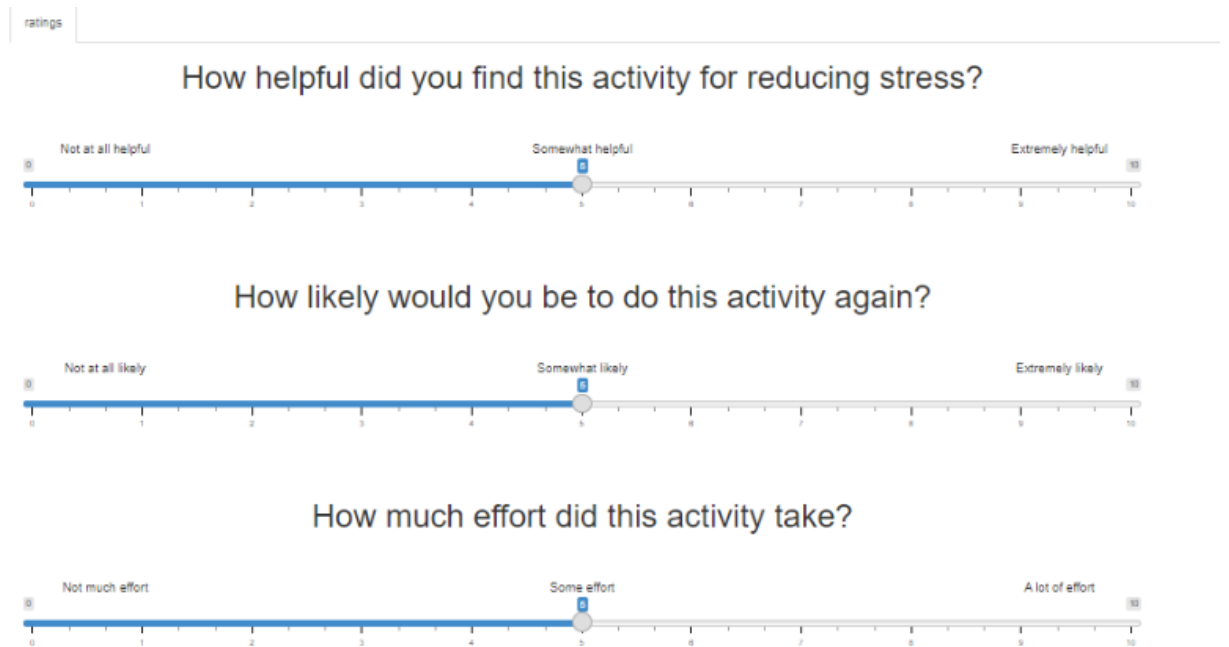




Figure 3.4: First half of the ratings page for the app described in Chapter 3


How meaningful did you find this activity?



How enjoyable did you find this activity?



How helpful did you find this activity for your general wellbeing?



Thank you! Please click the button below to record your response. You will then be redirected back to the recommendations page in case you wish to do another activity

Figure 3.5: Second half of the ratings page for the app described in Chapter 3

3.2.5 Analysis

First we explored surface level user behaviors using descriptive statistics and visualizations. These focused on understanding when users completed activities, how many they completed and if anything stood out from activity properties. We also ran a simple regression to see if there was any difference between screener WHO-5 and PSS-10 scores for both the per-activity and control groups.

We then look at our post-activity ratings to identify overlap and commonalities through correlation analysis and Principal Component Analysis (PCA). This was in part to understand the properties and usefulness of this type of measurement, but also to extract principal component scores for any overlapping ratings. We used varimax rotation to ensure independence of variables, and the PCA scores were then used instead of individual ratings for the following regression analyses. This was done to streamline analysis with fewer variables and avoid collinearity issues. We extracted PCA scores for the socioeconomic variables (income, education and subjective social status) for the same reason and purpose.

We then ran a series of regressions to investigate what variables were useful in predicting our post-activity ratings, engagement metric (number of completed activities) and PSS-10 and WHO-5 exit scores. For all analyses we standardized all continuous predictors and performed mean-centred deviation-coding on all categorical predictors. We ran multiple linear regression

analyses where post-activity ratings or mental health outcomes were our dependent variables, Poisson regression models where activity completions was our dependent variable.

For these regressions we were interested in the models as a whole, and so used model selection approaches designed to include the maximum number of informative predictors without sacrificing model accuracy, to find the simplest model with the highest predictive power. We therefore included all potential independent variables in an initial model. We then removed the predictor with the highest p value from a new otherwise identical model. We compared criteria for the two models, the adjusted R^2 for our multiple regressions and Akaike Information Criterion (AIC) for our generalized linear model, and kept the model with the best goodness of fit. We repeated this process until removing variables no longer improved the criterion. This also had the benefit of reducing the chance of overfitting our models, as we had a dataset rich in predictors.

3.3 Results

3.3.1 Descriptives and completion metrics

To begin with there were 40 participants in each condition, but by the end of the study 33 remained in the per-activity payment condition and 30 in the control condition, 63 total. Across all participants there were a couple of common trends in activity completions. First, figure 3.6 illustrates that they engaged considerably more with shorter activities. However, the longer activities were not ignored completely, indicating some level of interest or curiosity across the board.

Activity completions by duration

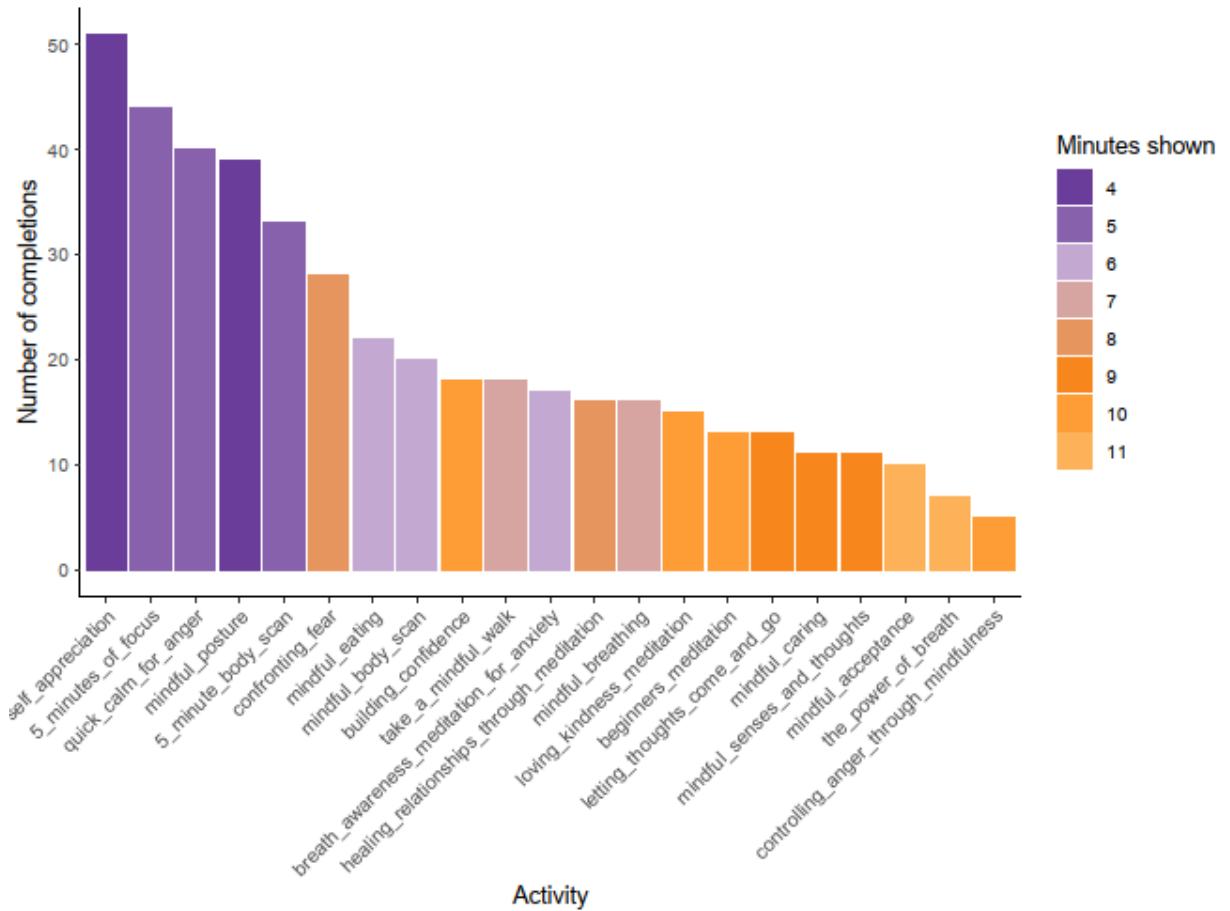


Figure 3.6: Bar chart showing the total number of individual activity completions and their stated duration

Second, the effect of reminders were noticeable (see figure 3.7). Whilst most activities were completed during the onboarding days, completions spiked on days when participants received a reminder about the mindfulness activities (i.e., every Monday and Friday), encouraging them to complete activities and reminding them about participant reimbursement.

Activity completions by date

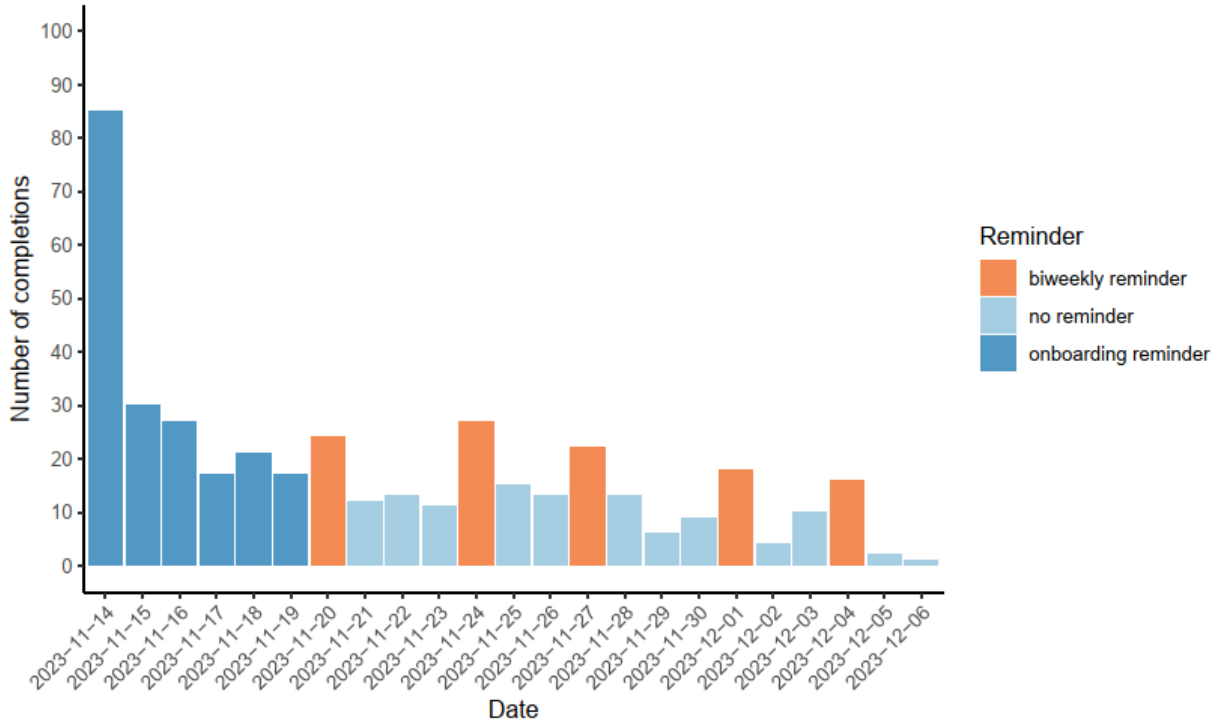


Figure 3.7: Bar chart showing the total number of activity completions by date and whether a reminder was sent

Group differences were also immediately visible. Figure 3.8 shows that activity completions were higher for the per-activity condition compared to the control condition. The mean number of completions for the control group was 4.33 (sd = 4.25, median = 2.5) and the mean for the per-activity group was 8.64 (sd = 4.58, median = 8). There was a difference in participant dropoff between groups, with more rapid user loss in the control group (see figure 3.9). However, in both groups the majority of users completed more than one activity (80% for the control group, 97% for the per-activity group), despite no further incentivization in the control group, suggesting that whilst potentially important, financial incentivization may not be the *only* reason for engagement. This is further supported by the fact that 27.27% of participants in the per-activity condition also completed more than 10 activities, the maximum they could be paid for. Conversely, compensation is also not necessarily enough, as only 32.5% of participants in the incentive group completed 10 activities.

Percentage of users in each group and number of activity completions

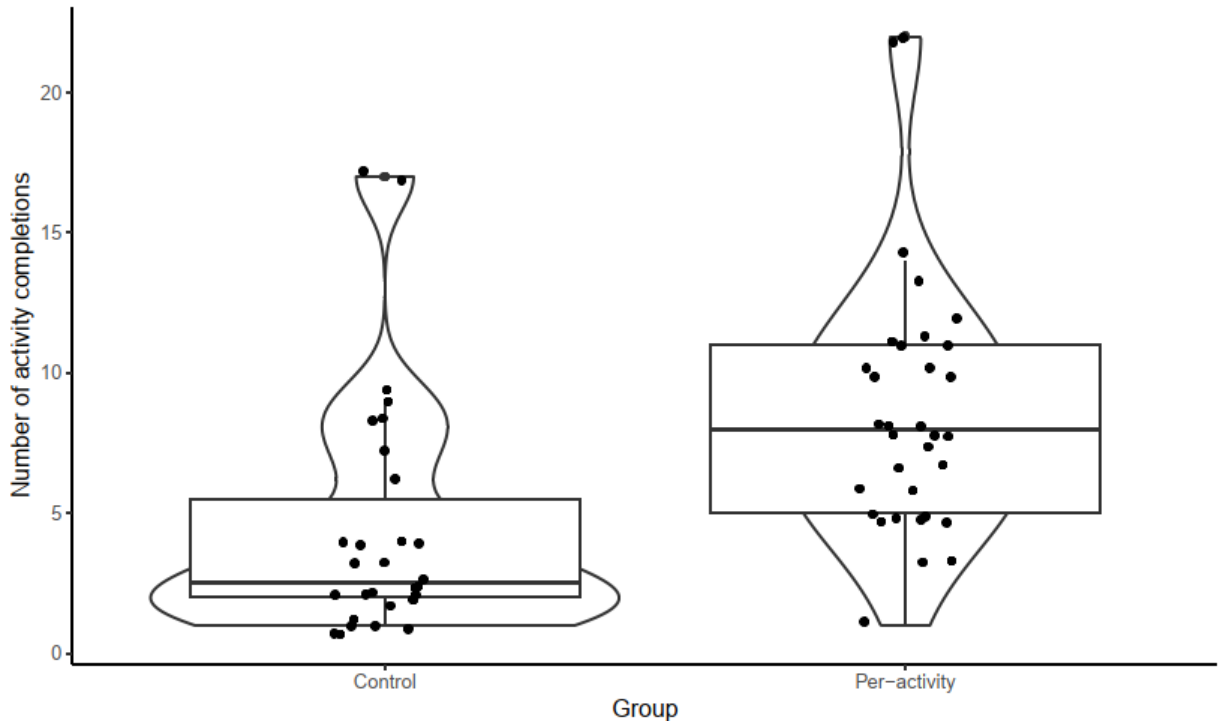


Figure 3.8: Plot showing the 'retention' of users. It describes the percentage of participants in each group remaining at each activity interval.

3.3.2 PCA and analysis of post-activity ratings

Socioeconomic PCA

Income, education and perceived social status loaded onto two principal components, which explained 84% of the variance. Income and social status loaded strongly onto a single component (over 0.8 for both) and education loaded strongly onto its own component (0.98). The component scores were used for all subsequent regression analysis. We will refer to these as the "*education component*" and the "*income and status component*" for the remainder of this study.

Post-activity ratings correlations and PCA

Ratings between all variables showed strong, positive correlations with Spearman correlation coefficients $> .80$ with the exception of perceived effort, which did not correlate above 0.09. We therefore performed principal component analysis to reduce the dimensionality of the five correlated activity ratings and left effort in its raw form (although scaled in the cases where it was used as an independent variable for regression). These five ratings all loaded strongly (> 0.9) onto a single principal component that explained 92% of the variance. For this study we will refer it as the "*usefulness and satisfaction*" component. As with the socioeconomic PCA,

we used the component scores for regression, in this case as both dependent and independent variables.

Predicting ratings

We ran two multiple linear regression models for this section using the model selection method described previously. Our dependent variables were the raw mean effort ratings across all activity completions per participant, and the mean of the *usefulness and satisfaction* component scores. Both models included as independent variables WHO-5 screener scores, PSS-10 screener scores, the *education* component scores, the *income and status* component scores, participant age, gender and group condition. The model with the *usefulness and satisfaction* component scores as the dependent variable also included the scaled mean effort ratings, as we were interested in whether effort predicted satisfaction.

The results for both final models can be seen in table 3.1 and table 3.2. Results for the full models including all variables are included in the appendix. The most notable findings are that adjusted R^2 values are relatively low for both models indicating poor goodness-of-fit. The PSS-10 screener scores and group conditions were consistent predictors of higher ratings across both dependent variables. The only significant predictor in either model was group condition, and only for predicting the *usefulness and satisfaction* component scores.

Table 3.1: Final model predicting effort ratings (adjusted $R^2 = 0.094$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
PSS-10 screener scores	0.417	0.225	1.855	0.069
Gender	0.774	0.461	1.678	0.099
Group condition	0.745	0.449	1.660	0.102

Table 3.2: Final model predicting *usefulness and satisfaction* component (adjusted $R^2 = 0.084$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
PSS-10 screener scores	0.173	0.088	1.970	0.0535
Group condition	0.356	0.175	2.036	0.0462

3.3.3 Analysis of engagement

A single generalized regression model (family = "poisson") was run to investigate the variables that affected predicted number of activity completions. The same model reduction technique was used, with original independent variables being mean effort scores, mean *usefulness and satisfaction* component scores, WHO-5 screener scores, PSS-10 screener scores, the *education* component scores, the *income and status* component scores, participant age, gender and group condition. The can be seen in table 3.3, and the model including all original independent variables is included in the appendix.

Table 3.3: Final model predicting activity completions

Independent variable	Coefficient	Std. Error	z value	p value
Mean effort scores	0.205	0.050	4.073	< 0.001
Mean <i>usefulness and satisfaction</i> scores	-0.092	0.0534	-1.717	0.086
PSS-10 screener scores	-0.135	0.054	-2.511	0.012
<i>Education</i> component scores	-0.101	0.055	-1.837	0.066
Age	0.161	0.051	3.145	0.002
Group condition	0.706	0.117	6.015	< 0.001

The final model included 6 variables, 4 of which significantly predicted number of activity completions. The *usefulness and satisfaction* component scores and the *education* component scores both negatively predicted the engagement outcome, but were not significant. Lower initial stress scores, higher mean effort ratings, higher age and being in the per-activity payment condition were all significant predictors of a higher number of activity completions.

3.3.4 Analysis of change in WHO-5 and PSS-10 scores

The first model we ran for this section was simply to see if there was a significant difference between onboarding scores and exit scores. For this we ran two simple linear regressions with raw WHO-5 and PSS-10 scores as our two dependent variables, and occasion as our independent variables. Results were positively significant for WHO-5 ($\beta = 2.63$, $p < .001$, adjusted $R^2 = 0.12$) but not for PSS-10 ($\beta = -1.17$, $p = .078$, adjusted $R^2 = 0.02$). Both models suggest improvement in mental health over time (the PSS-10 being scored negatively), yet not significantly for the PSS-10.

We ran two more multiple linear regression models, with change in WHO-5 scores and change in PSS-10 scores from onboarding to exit as dependent variables. The same model reduction technique was used, and original independent variables included number of completed activities, mean effort scores, mean *usefulness and satisfaction* component scores, WHO-5 screener scores, PSS-10 screener scores, the *education* component scores, the *income and status* component scores, participant age, gender and group condition. Table 3.4 and 3.5 show the results of the final model, and findings from the models with all original variables can be seen in the appendix

The final WHO-5 model included 6 variables, but only effort was significant, with higher mean effort ratings negatively predicting mental wellbeing change. Of the non-significant variables, higher mean *usefulness and satisfaction* component scores positively predicted WHO-5 change, whereas WHO-5 screener scores, PSS-10 screener scores, the *income and status* component and being male all negatively predicted WHO-5 score change.

In the PSS-10 model, Group condition was the only non-significant variable of the three final

Table 3.4: Final model predicting change in WHO-5 scores (adjusted $R^2 = 0.202$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Mean effort scores	-0.878	0.381	-2.307	0.025
Mean <i>usefulness and satisfaction</i> scores	0.527	0.374	1.410	0.164
WHO-5 screener scores	-0.671	0.372	-1.804	0.077
PSS-10 screener scores	-0.489	0.392	-1.247	0.218
<i>Income and status</i> component	-0.574	0.392	-1.464	0.149
Gender	-1.310	0.786	-1.668	0.101

Table 3.5: Final model predicting change in PSS-10 scores (adjusted $R^2 = 0.272$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
WHO-5 screener scores	-0.904	0.393	-2.301	0.025
PSS-10 screener scores	-1.831	0.390	-4.697	< 0.001
Group condition	0.925	0.774	1.195	0.237

variables included, with participants in the per-payment condition being more likely to see an increase in stress levels. Both WHO-5 and PSS-10 screener scores were significant, negatively predicting PSS-10 change.

3.4 Discussion

3.4.1 Main findings

The primary goals of this study were to examine whether compensating users financially influenced their levels of engagement with an online mental health app, and whether it influenced mental health improvement metrics. We found that financial incentivization strongly predicted how many activities users completed, more so than all the other independent variables in the final model together. The average number of completions was almost twice as high in the per-activity payment condition than in the control condition. We also found that it was not useful in predicting mental wellbeing or stress change. It was not significant in the WHO-5 model or the PSS-10 model, and only marginally improved the fit for the PSS-10 model. These findings carry positive and negative insights for those looking to do academic research on mental health apps and recommender systems. That researchers can utilize financial incentives to increase engagement without influencing mental health outcomes is largely a positive. Recommender systems can be data hungry, often performing better with more information, and payment can be a way to ensure minimum requirements are fulfilled in a timely manner. This, however, is old news. Taken as a single statement, "paying participants increases the likelihood that they will engage more with the content", is unsurprising. And whilst there is merit in testing unspoken assumptions, that alone hardly merits as much attention as it got in this study. What does merit consideration is the extent to which payment influences participants, and what side-effects there

may be.

The first side-effect is that studies that pay for participation cannot give a clear insight into the engagement value of a recommender system. The business viability and benefits are therefore hard to estimate, and may influence whether companies are likely to implement recommender systems from these studies into their commercial apps. However, our study also found that financial incentivization was not the alpha omega of engagement. In the control group, most participants did more activities than the minimum single one necessary for payment, and more than a quarter of participants in the per-activity payment group did more than the maximum of ten. This aligned with another study using the Foundations app where participants on average did over double the amount of required activities (Khwaja, Pieritz, Faisal, & Matic, 2021). However, the implementation of compensation based on non-mandatory engagement in our current study showed that the reverse was also the case. The majority of participants in the per-activity group completed less than ten, suggesting that money, at least in these amounts, was not enough to ensure complete engagement. This, at least, means that it may be possible, despite payments, to estimate *relative* engagement outcomes in mental health apps or recommender systems. The absolute values will likely be skewed, but if comparing two apps or recommender systems, engagement values can be compared in an incremental payment scheme, or if the minimum engagement requirements are not set too high.

Another potentially negative side effect we considered was whether it acted as a confounding variable for mental health outcomes. This study did not find that to be the case, which is good news for mental health app research. If participant payment does not influence mental health outcomes or data quality (More, 2022), results will have higher ecological validity and better robustness. For those interested mostly in mental health outcomes, monetary compensation should be a safe way to increase engagement without compromising the main findings.

However, this also carries the unfortunate implication that engagement does not influence improvement. Number of activity completions was not included in the final models for WHO-5 or PSS-10, indicating that it was not a useful predictor for either. This speaks to the assumption highlighted in Chapter 2, that persistent use of mental health apps would help long-term mental health improvement (Lewis et al., 2022). In terms of volume, this did not appear to be the case. As mentioned previously, the congruence in the literature is low for the effect of engagement on mental health, but it may come down to how engagement is measured. Volume of activity completions may not influence mental health outcomes, and results vary regarding duration of engagement (Henwood, Guerreiro, Matic, & Dolan, 2022; Ruiz de Villa et al., 2023), but there are at least some signs that different subjective engagement measures do (Graham et al., 2021). It is also possible to interpret Lewis et al. (2022)'s assumption of persistence as a measure of frequency, or consistency. This would explain why one study found that the activities users were most likely to engage with were also the ones that benefitted them the most (Ruiz de Villa et

al., 2023). Ruiz de Villa et al. (2023) looked at the influence of individual activities and activity modules, and speaks to a level of differentiation of improvement based on content. The current study only looked at overall engagement metrics, and may not have captured these nuances. This would be a strong argument for the usefulness of recommender systems - if general engagement does not influence mental health outcomes, but specific engagement does, personalization tools are a way to ensure intended effects. However, there may also still be general benefits that could not be identified through the current study. There may be a minimum number of beneficial activities with diminishing returns, or short-term improvements, that if continuous, can predict long-term improvement. Understanding and delineating these facets of engagement would be a valuable focus area for future research.

3.4.2 Further exploration of variables

The current study was designed with the intention of addressing, the relationship between immediate effects of mental health activities and long term effects. The post-activity ratings included both measures of subjective engagement and short-term improvement. However, neither the model predicting the *usefulness and satisfaction* component nor the model for predicting effort ratings were particularly strong, with low goodness of fit parameters, with no significant predictors of effort and only group as a positive significant predictor of *usefulness and satisfaction*. This may suggest that payment influences how participants perceive how rewarding activities are to engage with, although at a low adjusted R^2 (0.09) it is unlikely to be a major influencer. It is also interesting that both stress screener scores and group were included in the final model selections for both models, and may be worth exploring in further studies. However, based on results in this study, little can be confidently inferred about what may predict subjective engagement and perceived benefits from mental health activities.

Activity completions, however, were easier to predict. There was, as mentioned, the strong influence of group condition, yet this did not stand alone. Participants were more likely to complete more activities if they were older, less stressed and if they found the activities more effortful on average. This last finding is particularly interesting, as intuitively, participants should do *less* activities if they found it more effortful, however there are some potential explanations to be found in theoretical psychology. Completing harder tasks (to a point) can increase self-efficacy from achieving goals (Locke & Latham, 2006; Rosenstock, Strecher, & Becker, 1988). This may also be the case here, where users who perceived activities as more effortful may have experienced a higher sense of reward. This notion has even been leveraged in a mental health recommendation setting, where algorithms were designed to optimally match user ability with task difficulty (Torkamaan & Ziegler, 2022), improving user engagement. One caveat however, is that we would then also expect effort to predict perceived *usefulness and satisfaction*, which it did not in this study. This may come down to the averaged scores. It may be occurring at

an activity level, and information is being lost in reduction to central tendencies. In this study the sample size of participants was too small relative to the sample of activities to perform any meaningful analysis, however we determined that it would be worthwhile exploring in future studies. This will be further explored in a larger study in Chapter 5.

Our models predicting mental health outcomes showed mixed results. On one hand, the models themselves were adequate, with moderate adjusted R^2 values, suggesting that they may be useful in determining stress and wellbeing change. However, very few of the predictors were significant, indicating that a larger sample size may be necessary to gain accurate estimates of degree. The variables included in final models may still be able to offer some insights for future studies. First, mean effort ratings were the only significant predictor of WHO-5 change, and higher perceived effort of activities predicted a *decrease* in WHO-5 scores. These are early findings, and must be taken with a pinch of salt, but if replicated they suggest that the algorithms explored by Torkamaan and Ziegler (2022) may be walking a fine line. Our results suggest that recommending more effortful activities to users may increase engagement levels, but if effort levels are too high it may negatively impact mental health outcomes. This does not contradict the premise that matching user ability to activity difficulty can increase mental health, but it does indicate a delicate balance. Finding that balance is exactly the purpose of the algorithms described by Torkamaan and Ziegler (2022), but the current study stresses that it cannot be a default assumption that increasing engagement with mental health activities will improve mental health outcomes. This is a relationship that is often not explored simultaneously in the literature, but one that may be necessary to align the different goals of mental health apps and recommender systems.

However, selecting measures for mental health outcome evaluation may not be simple. In this study we used self-report questionnaires. For both our WHO-5 and PSS-10 final models, the screener scores were included as predictors after model selection, though they were only significant for PSS-10. In all cases they were negative predictors of change. This is notable, because they are scored in opposite directions, and therefore the results are contrasting. Better original mental wellbeing predicted that users would get worse by the end of the study, but higher original stress scores predicted they would get better. These are distinct variables, and so there might be several explanations for this, but it does potentially indicate floor or ceiling effects. These are well known issues in self-report measures (Bot et al., 2003), where participants who score lower have more room for upwards than downwards change and vice versa. These effects are also exacerbated in samples focusing on only one end of the distribution, as the current study did. So whilst the notion that individuals with lower mental health scores are more inherently motivated to engage with mental health apps is well placed, sampling from this population may have unintended methodological consequences and must be kept in mind for future studies.

3.4.3 Future directions

Our findings show that future research into mental health apps and recommender systems need not worry too much about payments confounding measures of mental health change. It also showed that although absolute measures of engagement are likely to be skewed, researchers who are interested in comparing apps or algorithms can still compare the relative effects on engagement.

More research is needed to determine the precise relationship between engagement and improvement. It is by no means a foregone conclusion that higher activity in mental health apps automatically leads to mental health improvement. This is worth further testing and replication, however even more pertinent is understanding the different facets of engagement. Beyond objective measures, this study included measures of subjective engagement, however we found significant overlap with perceived stress benefits, mental wellbeing benefits, activity meaningfulness, enjoyability and repeatability. Understanding whether these are truly tapping into the same construct, or whether this is simply an artefact of participant behavior would be a worthwhile endeavour for future studies. Furthermore, this study did not leverage information about consistency or minimum requirements of engagement, and these are still potentially influential on improvement. After all, we did see a significant change in both WHO-5 and PSS-10 scores, and RCTs using the parent application have determined an effect of access to the app (Catuara-Solarz et al., 2022; Gnanapragasam et al., 2023; Schulze et al., 2024). It may be that there is a minimum number of completed activities that determine improvement, and beyond that a sharply diminishing return on investment. It is also possible that persistent use over time can improve mental health, but the process cannot be rushed. It could even be that the app itself does not matter, and that improvement relies on a placebo effect. Future studies should investigate exactly what metrics of engagement are useful in predicting mental health outcomes, which would also contribute to resolving incongruency issues in the field (Lekkas et al., 2021; Linardon & Fuller-Tyszkiewicz, 2020; Molloy & Anderson, 2021).

Effort also appears to be a worthwhile variable to explore when it comes to engagement and mental health change. It was an informative predictor in both our models predicting number of completions and WHO-5 change and could play an important role in user experience and behavior when engaging with mental health apps. Whilst it is already being utilized as an input for recommender systems in mental health (Torkamaan & Ziegler, 2022), understanding its influence across multiple outcomes may benefit the design of recommender systems, or even the development of the mental health resources themselves.

Generalizability of results should also be a consistent question in future studies. Our sample was relatively limited, with a medium sized total across both groups, and looking only at a subset of the population. Mental health apps are intended to be a ubiquitous solution, and

understanding whether apps are useful in a broader sample or across different questionnaires or target mental health outcomes is critical for our overall understanding of how they work. It is also necessary to investigate the distinction between individual activities and solutions. The current study included had a high ratio of activities per user, and these activities were all derived from a single field, mindfulness and meditation. This would have made it difficult to investigate their individual effects, and future studies should attempt to understand how different solutions in mental health apps might drive change across different parameters or how they might vary in effect for different users.

The same is true for users. Individual differences should be investigated further, to better understand what might help a user improve more, or lead to higher engagement. This study included demographic, mental health and behavioral variables, however there are many more potential influencers, and a rich literature to draw from for inspiration from general research into mental health. Several of these will be explored further in Chapter 5.

3.4.4 Limitations

The app used for this study had some minor issues on launch. It loaded slowly on startup for some users, but was mostly resolved by refreshing page or switching browsers. The browser issues were known and reported to users in the information sheet. Some few did not see this, or forgot, and were reminded if they contacted researchers during the study, yet it may have influenced some users' perception of difficulty or usefulness.

It may also have made a difference that the app was designed as a web app. It was possible to access through mobile web pages, yet information on the medium users utilized was not collected. The extent to which this influenced results (if at all) is therefore unknown, as there is little research on the effect of medium on any of our chosen variables. Further research could address the generalizability of results across platforms and mediums.

Other limitations have already been alluded to through the discussion. We investigated the effect of financial incentivization, yet the scalar effect of payment is still uncertain. We used estimated hourly wages and prolific guidelines to determine what constituted moderate but fair payments, however incentives change from study to study. Some may pay more, some may pay less and some may use alternative compensation methods such as gift vouchers, or credit as part of university courses. Similarly, for reasons of fairness we did compensate our control group users, but not by completion. This still allowed us to isolate the effect of payment, as users would be completing activities for free after the first, yet it still may skew engagement outcomes compared to unpaid participants (or users in the wild). Once again this speaks to generalizability issues, as we do not know whether our findings replicate across different payment types or levels. These are things that can only be addressed through continued research, but must be kept in mind for

any individual study.

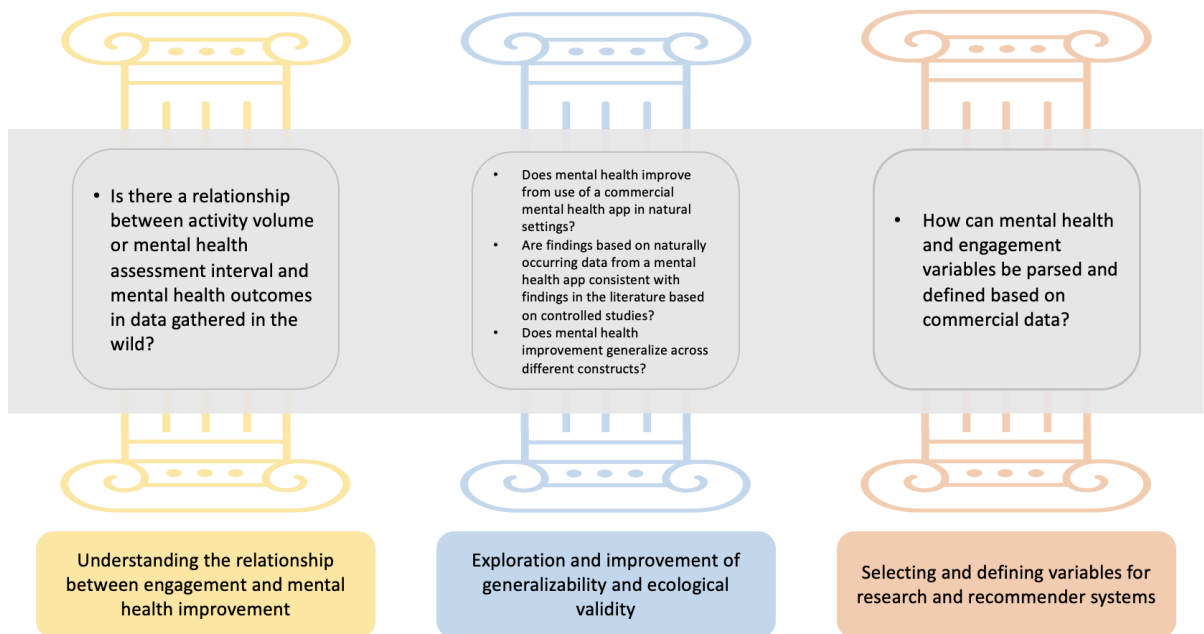
Finally, our post-activity ratings, specifically the five that loaded onto a single factor, merit more consideration. Even if they are naturally interwoven, we were interested in isolating user perspectives across different metrics such as subjective engagement and mental health states, and the high overlap did not allow us to do this. Further experimentation with these variables is required to allow researchers to better capture real-time effects of mental health activities.

3.4.5 Conclusion

This study investigated how financially incentivizing users influenced their engagement and improvement from online mental health activities, and what the relationship was between engagement and mental health outcomes. It found that per-activity payments strongly influenced how many activities users completed, but did not influence mental health outcomes. It was also not the only predictor of engagement. Future research can therefore safely compensate participants for their time without worrying about negative impacts on mental health results, but absolute measures of engagement, and potentially commercial value estimates are likely to be skewed. However, the relative engagement metrics may still be useful, meaning that apps and algorithms can be compared with each other within paid studies. This, however, also tied together with the fact that the current study did not find engagement to be a useful predictor of mental health. This is critical, as it contradicts an assumption in current research, and stresses that future studies of engagement in mental health apps cannot make the assumption that tools that improve engagement will automatically lead to mental health benefits from the user. It stresses that further research on mechanics of change are needed to improve mental health apps for all stakeholders.

Chapter 4

Exploring real world data



4.1 Introduction

This chapter looks at data collected naturally over a long period of time in a commercial mental health app. Designing experimental conditions that capture real world processes can be very difficult (Barsalou, 2019; Czaja & Sharit, 2003), perhaps especially in psychology where this a prominent concern for generalizability of findings (Miller et al., 2019). This PhD was done in collaboration with an industrial partner (Koa Health), and therefore had the opportunity to investigate user behavior, engagement and mental health change "in the wild" through their mental health app 'Foundations'. There were several overlaps between data gathered for this study and the other studies presented in this thesis, allowing for direct comparative insights

between the academic and commercial world.

The efficacy of the Foundations app has been evaluated in multiple trials, largely showing positive results. Equally importantly, it has been evaluated across different population samples, showing efficacy for individuals with increased levels of stress and anxiety (Catuara-Solarz et al., 2022), for students (Schulze et al., 2024) and in a large high-risk sample of workers in the National Health Service (Gnanapragasam et al., 2023). These trials also evaluated mental health outcomes on a number of different parameters, including mental wellbeing, stress and depression. There have also been studies investigating the influence of recommendations within the app on engagement (Ruiz De villa, Sottocornola, Coba, Lucchesi, & Skorulski, 2024) and the effect of individual activity types on mental health outcomes (Ruiz de Villa et al., 2023). However, these studies examined data collected, through various means, specifically for research purposes and therefore controlled several elements of the procedure such as mental health evaluation intervals or app access. The current study will examine data in natural settings where every decision from downloading the app and onwards was autonomously made by the users.

There are advantages and disadvantages to this. Attempts can (and will) be made through filtering procedures to wrangle data into categories that allow for direct comparison, but it still does not reach the degree of control possible in experimental conditions, and reducing control increases potential bias. For example, the data for this study was collected over several years, and this allows for a larger sample size. The data can be grouped in terms of time, for instance comparing users who have taken wellbeing assessments two weeks apart. Or, within the data, we could select only users within a certain month of a certain year, or within a single version update of Foundations. However, applying all of these filters to exert maximal control over variables occurring naturally, reduces sample sizes to the point where robust or informative findings are unlikely. These elements are, for the most part, equalized quite naturally in academic experimental designs. But each additional way variables are managed makes for a more synthetic paradigm, and research can often go overboard in creating the perfect theoretical setting (Barsalou, 2019). Ideally, both research avenues are explored, and reviews or syntheses of results across methodologies provide insights regarding overlaps and ecological validity. However, currently the literature on mental health apps is predominantly from experimental settings, mostly RCTs (Neary & Schueller, 2018; Tønning et al., 2019), and this study aims to help balance the scales.

The current study aimed to further our understanding of mechanisms of mental health change, and aligns with the main themes in the rest of the thesis. The Foundations application has an inbuilt voluntary mental health assessment users can complete at will, which is comprised of smaller assessments within multiple mental health domains. Koa also collects some user data passively, logging activity completions and activity related data. This allowed us investigate how engagement metrics and mental health change are related in a naturally occurring sample.

Due to the variability in assessment intervals, it also allowed us insights into temporal variations of change, which have been highlighted in previous research as important. In a study using the Foundations app (Schulze et al., 2024), participant mental health improvement was significant over a 4 week interval but not over a 2 week interval. By parsing the data available, we could investigate the generalizability of results across different mental health constructs, over different intervals and all in a naturally occurring setting.

4.2 Methods

4.2.1 Participants, assessments, activities and filters

No sensitive or demographic information was collected or included in any of the datasets made available for this study.

The raw wellbeing assessment dataset consisted of 15,382 assessment completions by 7438 users stretching back to the 19th of April 2019 with the last completion on the 1st of July 2024. A number of filters were applied for more meaningful analysis. First, we only kept users who did their first wellbeing assessment within two days of onboarding. The reasoning for this was that prolonged access to the app, and potentially a higher number of activity completions before the first assessment, could create a distorted baseline, where mental health change had already happened. Second, as we were largely interested in mental health change and its predictors, data from users with less than 2 separate mental health assessments were discarded. For the purpose of this study we were also mainly interested in the initial change from engaging with the app, and therefore only kept data from the first two assessments per user.

We also wanted to limit the length of time between assessments for results that were more comparable across existing literature and academic studies. Furthermore, we wanted to retain the ability to investigate whether interval lengths influenced mental health change, and therefore included assessments from set interval thresholds. Finally, we wanted to minimize potential bias from individual users completing high numbers of wellbeing assessments, and so limited the number of possible assessment inclusions from each user at each threshold. To this end, we isolated the first wellbeing assessment users completed, and bucketed assessments completed at the closest date nearest to 14 days, 28 days, 42 days and 56 days from the initial completion date. To avoid overlap in sampling, and to reduce excessive influence from individual users, we collected maximum one assessment per interval, and set sampling ranges to include only results from 7 days before to 6 days after each 2 week marker. This left 1089 unique users, with two assessments each, totalling 2178 assessments.

The raw activity dataset, which was generated independently from the wellbeing dataset consisted of 678,694 interactions with activities. This was filtered to include only the users present

in the wellbeing dataset, and removed any activity data points where users had interacted with or viewed the activity, but not completed it. This left 65,442 activity completions by 1059 unique users (some users had completed wellbeing assessments, but no activities).

We then proceeded to join these datasets together and applied a further set of filtering parameters in an attempt to remove confounds that might influence our results. First, we created a synthetic control group (keeping their data separate for further analysis), consisting of 57 unique users who had not completed any activities before their second wellbeing assessment. We removed these 57 users from the main group. This reasoning behind this was that users who did not engage with any of the mental health solutions the app offered, theoretically would not experience any benefits. In the remaining 1032 users in the main sample, we removed 103 users who were within the top 0.1 quantile of total activity completions (having completed over 37 activities). This left a final sample of 929 users in the main dataset with 11,296 activity completions between them.

These were our final datasets. Some transformations and subgroup sampling was done for specific analysis, but unless otherwise directly stated, these were the samples used for analysis described in the remainder of this chapter.

4.2.2 Materials

The Foundations mental health assessment consists of 16 questions drawn from five different commonly used mental health questionnaires capturing different constructs. These questionnaires include the the 5-item Wellbeing Index of the World Health Organization (WHO-5), the 4 item Perceived Stress Scale (PSS-4), the 3 item Minimum Insomnia Symptom Scale (MISS-3), the 2-item Generalized Anxiety Disorder questionnaire (GAD-2) and the 2-item Patient Health Questionnaire (PHQ-2). These questionnaires originally assessed symptoms or qualities over different quantities of time, however to ensure congruence, the Foundations version slightly changed wording to measure all constructs in only the previous week. Also to ensure congruence, they normalize final scores on the original scales to a scale of 0-20. These measures will be briefly described below.

WHO-5

The 5-item Wellbeing Index of the World Health Organization (WHO-5) assesses mental wellbeing in individuals. Users rated items such as “In the last week, you have felt cheerful and in good spirits?” on a six point likert scale, ranging from 0 (at no time) to 5 (all of the time). The final score is calculated from the sum of the individual items, and a score of 12 or below has been used as a threshold to indicate low mental wellbeing (Topp et al., 2015). The WHO-5 has shown satisfactory reliability across nationally representative samples of 35 European countries

(Sischka et al., 2020) and is a sensitive and specific screening tool for depression (Topp et al., 2015).

PSS-4

The 4-item Perceived Stress Scale is the short-form version of the PSS-10 described in the previous chapter (Cohen et al., 1983). It assesses how much distress individuals experienced. Users rated items such as “In the last week how often have you felt difficulties were piling so high that you could not overcome them?” on a five point likert scale, ranging from 0 (never) to 4 (very often). The final score is calculated from the sum of the individual items, and although no official threshold has been established, a score of 6 or above has been used to indicate high stress based on population norms (Malik et al., 2020). The PSS-4 has shown acceptable psychometric properties, although poorer than its longer counterparts (E.-H. Lee, 2012).

MISS-3

The 3-item Minimum Insomnia Symptom Scale (Broman, Smedje, Mallon, & Hetta, 2008) was developed as a concise measurement of sleep issues. It assesses the perceived quality of sleep and wakefulness in individuals. Users rated items such as "In the last week have you had problems with feeling like you are refreshed by sleep?" on a five point likert scale ranging from 0 (none) to 4 (very severe). The final score is calculated from the sum of the individual items, and although no official threshold has been established, a score of 6 or above has been proposed as a cutoff score to indicate potential insomnia. The MISS-3 has shown acceptable psychometric properties across different populations and have been found to be useful screening tools for insomnia (Hedin et al., 2022; Hellström, Hagell, Fagerström, & Willman, 2010).

GAD-2

The Generalized Anxiety Disorder 2-item is a short-form version of the original GAD-7 (Spitzer, Kroenke, Williams, & Löwe, 2006). Both items have shown good psychometric properties and usefulness in screening for generalized anxiety disorder (Plummer, Manea, Trepel, & McMillan, 2016). The final score is calculated from the sum of the individual items, and a score of 3 or above has been found to be optimal for screening for depression (Plummer et al., 2016). Participants rated items such as "In the last week, how often have you been bothered by feeling nervous, anxious, or on edge" on a 4 point likert scale ranging from 0 (not at all) to 3 (nearly every day).

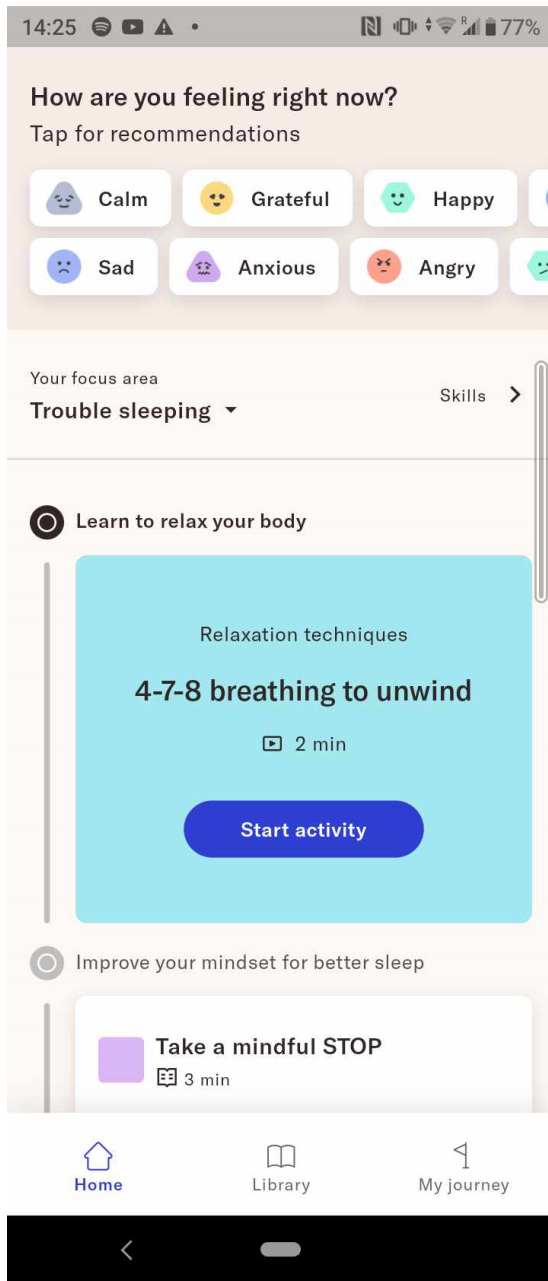
PHQ-2

The 2-item Patient Health Questionnaire (Kroenke, Spitzer, & Williams, 2003) is a short-form version of the PHQ-9 (Kroenke, Spitzer, & Williams, 2001) designed to detect symptoms of de-

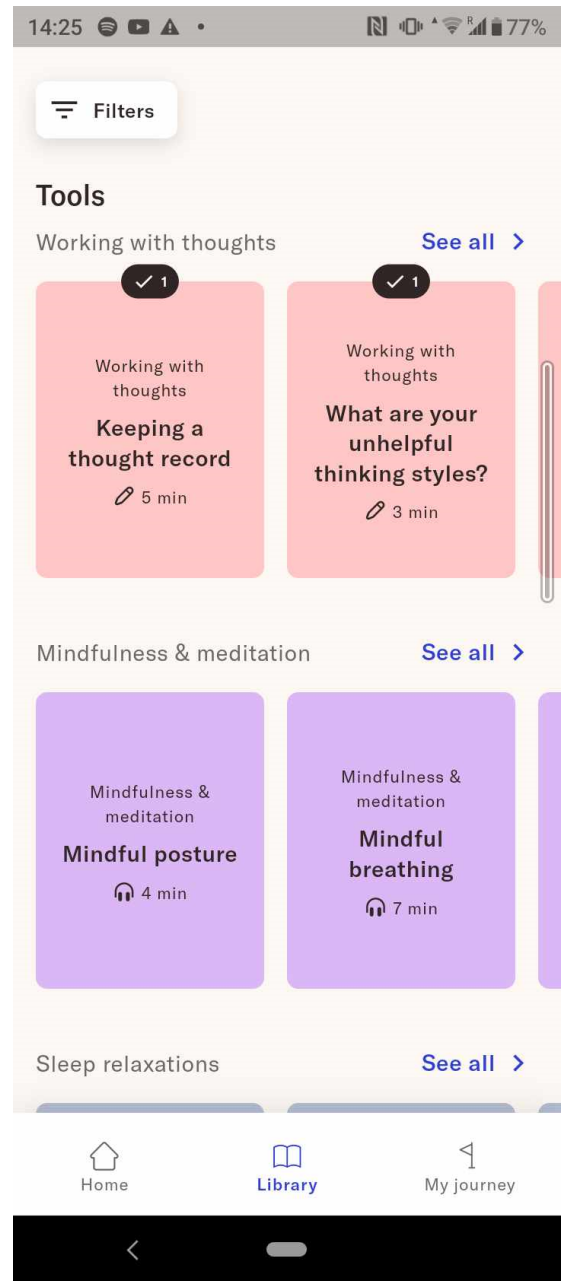
pression. Participants rated items such as "In the last week, how often have you been bothered by feeling down, depressed or hopeless" on a 4 point likert scale ranging from 0 (not at all) to 3 (nearly every day). The final score is calculated from the sum of the individual items, and a score of 3 or above has been found to be optimal for screening for depression (A. J. Mitchell, Yadegarfar, Gill, & Stubbs, 2016). It has shown acceptable psychometric properties, and has been identified as a useful tool in first-step assessment of depression, but not for clinical confirmation (Levis et al., 2020; A. J. Mitchell et al., 2016).

App and activities

The Foundations app includes a range of activities designed to help improvement across a wide range of mental health issues based on different techniques, such as CBT, relaxation techniques or meditation. In our final dataset, users had completed 363 unique activities between them. More are available in the application, and the selection has changed slightly over time, however not every activity was represented in our sample. Users can access these activities through various means. Figure 4.1a shows the home screen, where users can indicate their mood and receive knowledge-based recommendations generated by mental health experts. They can also select a focus area, where different programs (groupings of activities) or individual activities are recommended. They can also access them through the library page, where exercises are categorized into 9 different modules based on type, such as 'mindfulness and meditation' or 'positive psychology' (see figure 4.1b). They can interact with specific activities by pressing the icon, after which they will be redirected to the activity page (see figures 4.1c and 4.1d). In addition to the various activity types, the activities also present in different formats, some requiring listening and others writing.

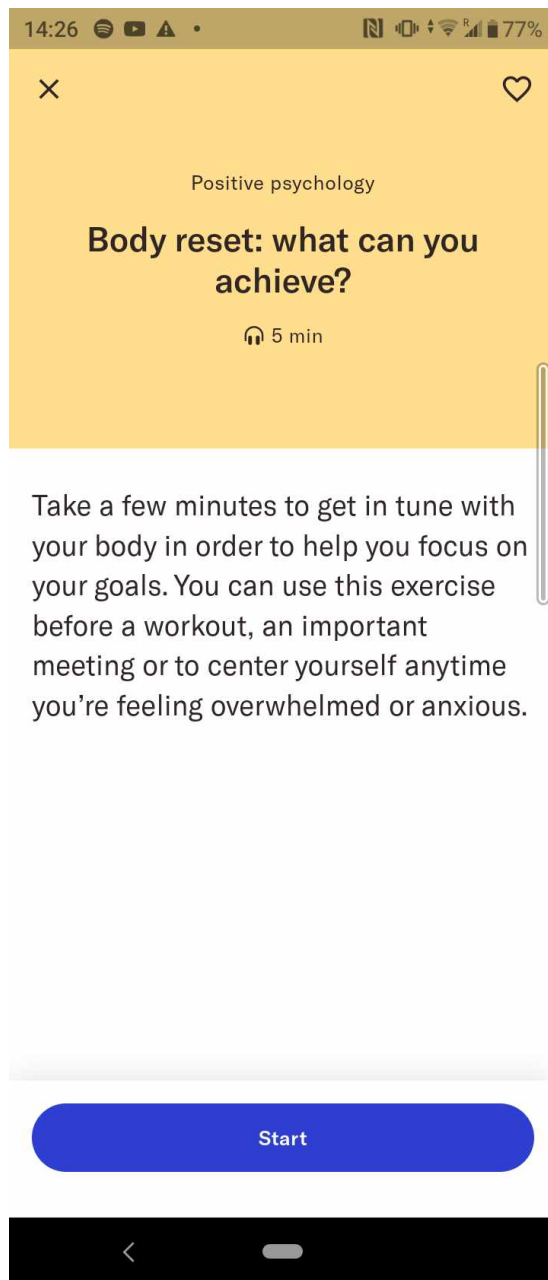


(a) Home screen for Foundations app.

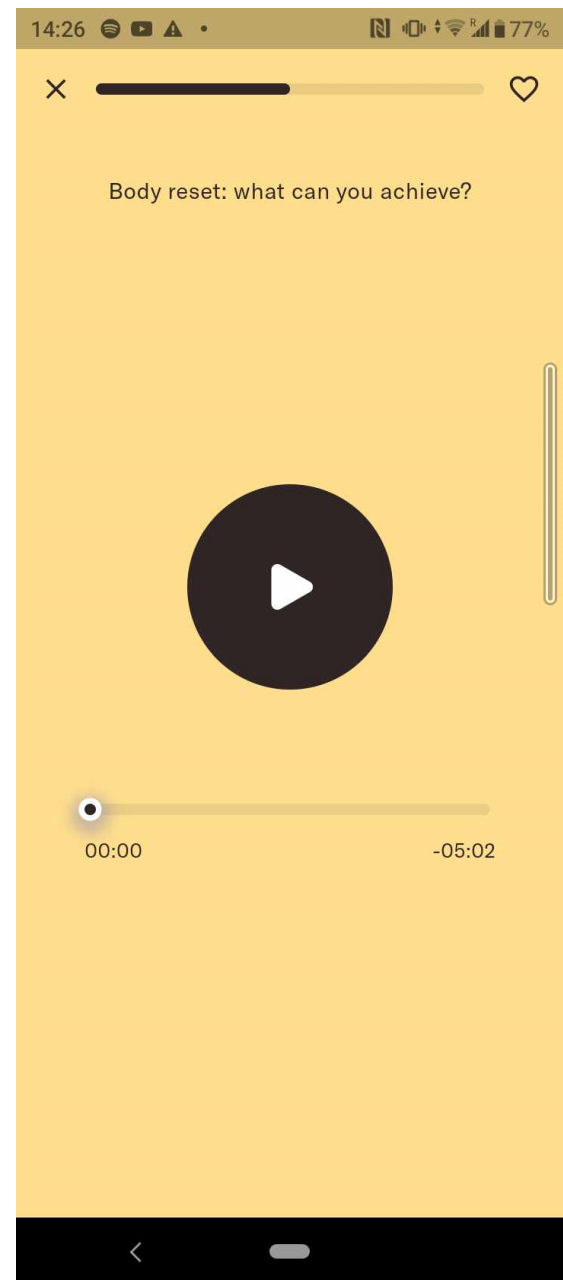


(b) Library screen for Foundations app.

Figure 4.1: Home and library screens of the foundations app



(c) First page for the "Body Reset" activity in Foundations.



(d) Main activity page for the "Body Reset" activity in Foundations.

Figure 4.1: Activity pages for the "Body Reset" activity in Foundations

4.2.3 Analysis

Analysis was done primarily on the main dataset, excluding the 'control' users. Unless explicitly stated where groups were compared, or the control group was analyzed independently, it can be assumed that analysis concerned the sample of 929 users with 2 assessments each.

First we calculated some descriptive statistics regarding how many activities users in our main sample completed, what the distribution of mental health scores were and how they changed.

Within the app, all scores are normalized to a scale of 20 and inverted for the same directionality (higher scores equal better mental health), and these are the parameters considered when describing the "raw" scores. We then proceeded to investigate the relationships between the different questionnaires through various psychometric analyses. We calculated the correlations between the various mental health scores, ran a PCA analysis and conducted various generalizability studies using Generalizability Theory (Cronbach, Rajaratnam, & Gleser, 1963), focusing on how generalizable scores were across questionnaires and time.

We then ran a series of regression models. For all analyses we standardized all continuous predictors and performed mean-centred deviation-coding on all categorical predictors. For all analysis where mental health change was the outcome variable of interest, we ran 5 separate models, one with each questionnaire as a dependent variables. Each model or set of models are described below.

Investigating change over time in main sample

Here we used the main sample of 929 users. In order to investigate whether user mental health scores changed over time, we ran 5 models with raw mental health scores as our dependent variable, and the measurement occasion (whether it was a user's first or second assessment) as our independent variable.

Investigating change over time in control group

Here we used the control sample of 57 users. To investigate whether users who had not completed any activities in between their assessments experienced change in mental health scores, we ran 5 linear regression models with raw mental health scores as our dependent variable, and the measurement occasion as our independent variable.

Investigating differences between groups

We also wanted to see whether there was a difference between the two groups. First we ran a regression model with change in mental health measures as dependent variable, and group as independent variable. However, as samples were heavily imbalanced, we also bootstrapped our analysis for more robust results to compare with and to reduce chances of bias. We ran 25,000 iterations, in each iteration sampling 57 users from the main data frame (with replacement). In each case a linear regression model with change in the mental health score of interest was run, with group as a predictor. The slope coefficient was saved across the 25,000 iterations to generate a distribution of effects, and we calculated the two-tailed p-value from the number of times it was above or equal to 0. This process was repeated 5 times, once for each mental health outcome.

This was the last analysis where the control group was included or investigated, the remainder were only based on the main sample of 929 users.

Investigating whether interval predicted change

To investigate whether the length of interval between assessments (in the buckets described previously), we ran a regression model for each of the mental health outcomes. In each case, change in mental health scores was the dependent variable, and the interval buckets were the independent variables. However, for these models we found that assumptions of collinearity were violated. We therefore ran new Generalized Least Squares (GLS) models (as opposed to Ordinary Least Squares) to account for correlation between variables. Only the GLS model results will be reported in the analysis section.

Investigating predictors of change

This analysis was intended to investigate whether any of our included values predicted mental health change. Change in scores for each of the mental health outcomes were the dependent variables. Our independent variables consisted of change in mental health on the other parameters, number of activity completions, and original mental health scores, in the latter two cases with variables separated into buckets. We bucketed the activity parameter to address findings described in Chapter 3. Number of activity completions as a continuous parameter did not predict mental health outcomes, however we theorized that there may be a minimum threshold of activity completions which may be beneficial, with diminishing returns. We therefore separated users into categories based on whether they had completed 1-3 activities ($n = 155$) or more than 4 activities ($n = 774$). We separated original mental health scores into buckets based on the thresholds described under materials (section 4.2.2), scaled to fit the Foundations scoring system. Each users was categorized as "low" or "high" for the WHO-5 (low: $n = 525$, high: $n = 404$), the PSS-4 (low: $n = 521$, high: $n = 408$), the MISS-3 (low: $n = 174$, high: $n = 755$), the GAD-2 (low: $n = 240$, high: $n = 689$) and the PHQ-2 (low: $n = 152$, high: $n = 777$).

Investigating predictors of engagement

The final analysis was done to investigate whether original mental health scores (bucketed as described above) predicted number of activity completions as a continuous dependent variable. As this was a count variable, we ran a Poisson regression model.

4.3 Results

4.3.1 Descriptives and psychometrics

Activity completions

Users in our main sample completed a total of 11296 activities between them, ranging from 1 to 37 completions (mean = 12.16, median = 10, SD = 9.15). The distribution of activity completions can be seen in figure 4.2. The most popular activity module was "Working with thoughts", with 4207 total completions and the least popular was "Workplace science" with 328 completions. The relative popularity of each module can be seen in figure 4.3.

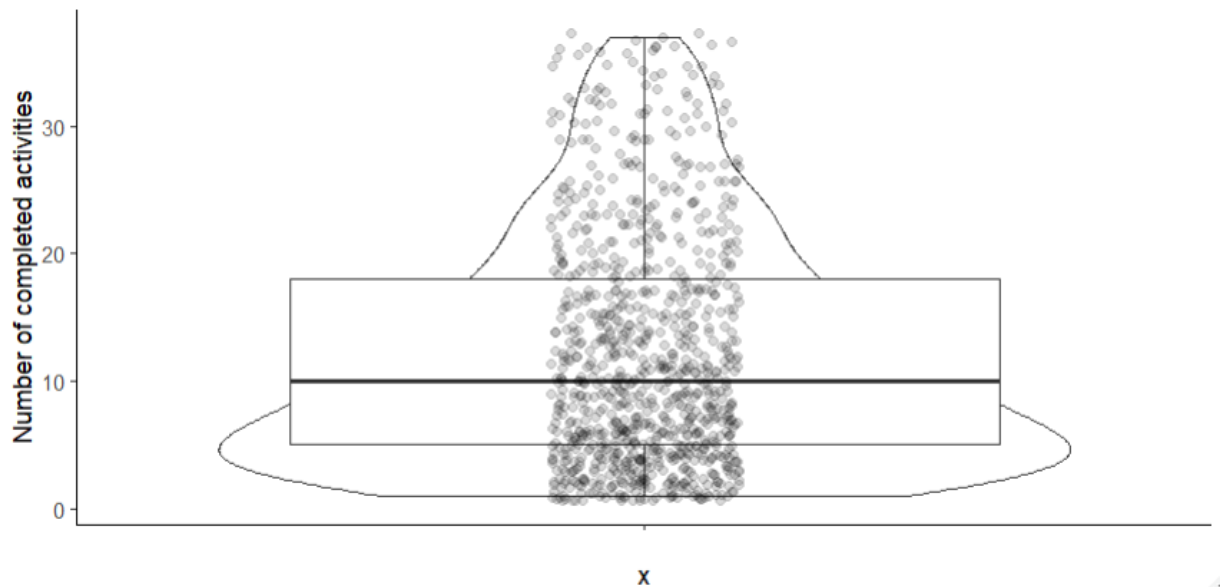


Figure 4.2: Distribution of total activity completions by user

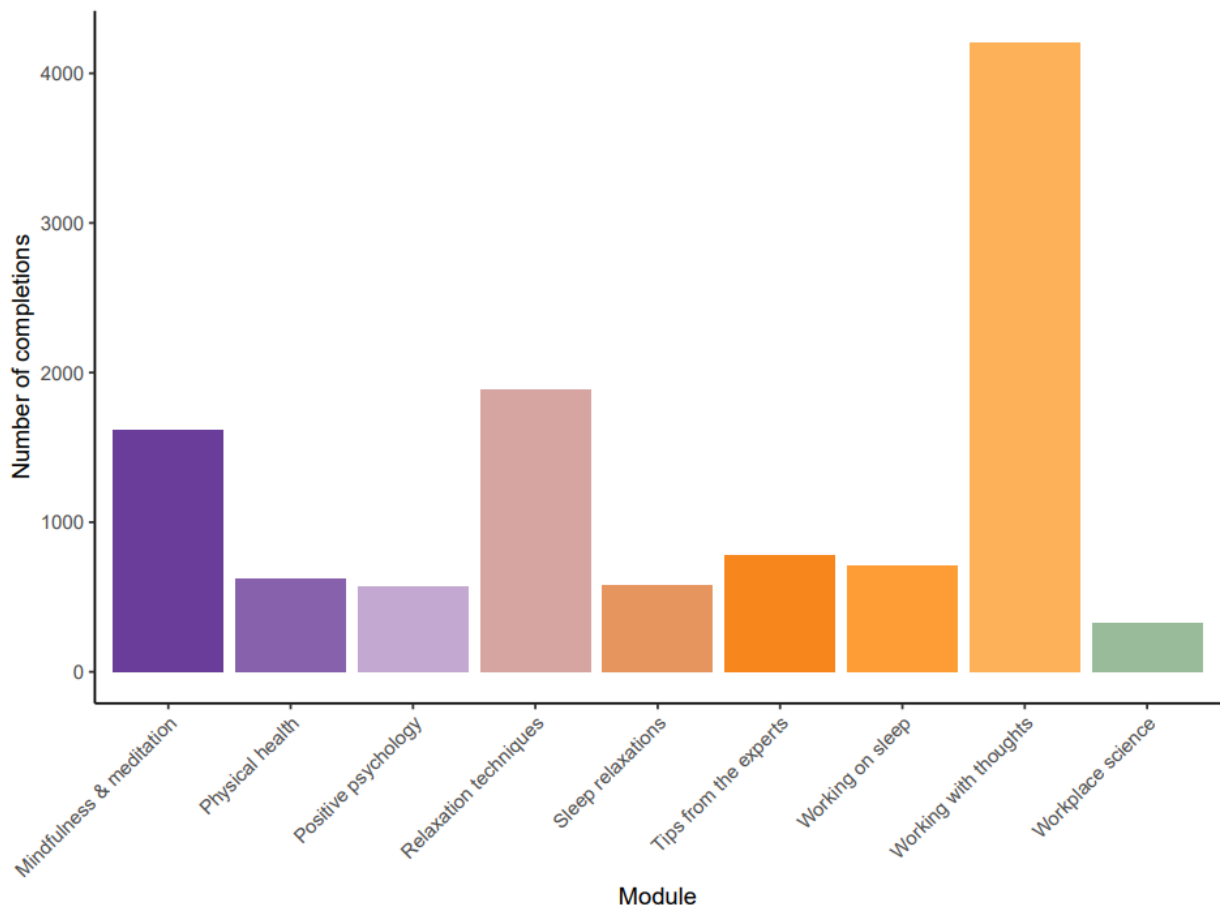


Figure 4.3: Activity completions by module

Descriptives of original scores

Table 4.1 shows the means and standard deviations of the different mental health scores for the first assessments users did, for both the main group and the control group. Notably, the means for the control group are higher for all of the questionnaires. The mean scores also differ considerably for the separate measures, ranging from 8.49 to 13.96 in the main group and from 9.61 to 15.60 in the control group. The standard deviations are high for both groups, suggesting a diverse sample as regards mental health. Figure 4.4 shows the distribution of original wellbeing scores for the main group and figure 4.5 shows the same for the control group. The diversity of scores and different means are once more clear, but the graphs do highlight that distribution trends are similar across both groups.

Table 4.1: First assessment scores for both groups

(a) Main group

	WHO-5	PSS-4	MISS-3	GAD-2	PHQ-2
Mean	8.49	11.42	12.71	12.22	13.96
SD	4.29	4.22	4.35	6.15	5.67

(b) Control group

	WHO-5	PSS-4	MISS-3	GAD-2	PHQ-2
Mean	9.61	12.68	13.07	14.09	15.60
SD	4.93	4.66	5.70	6.75	5.44

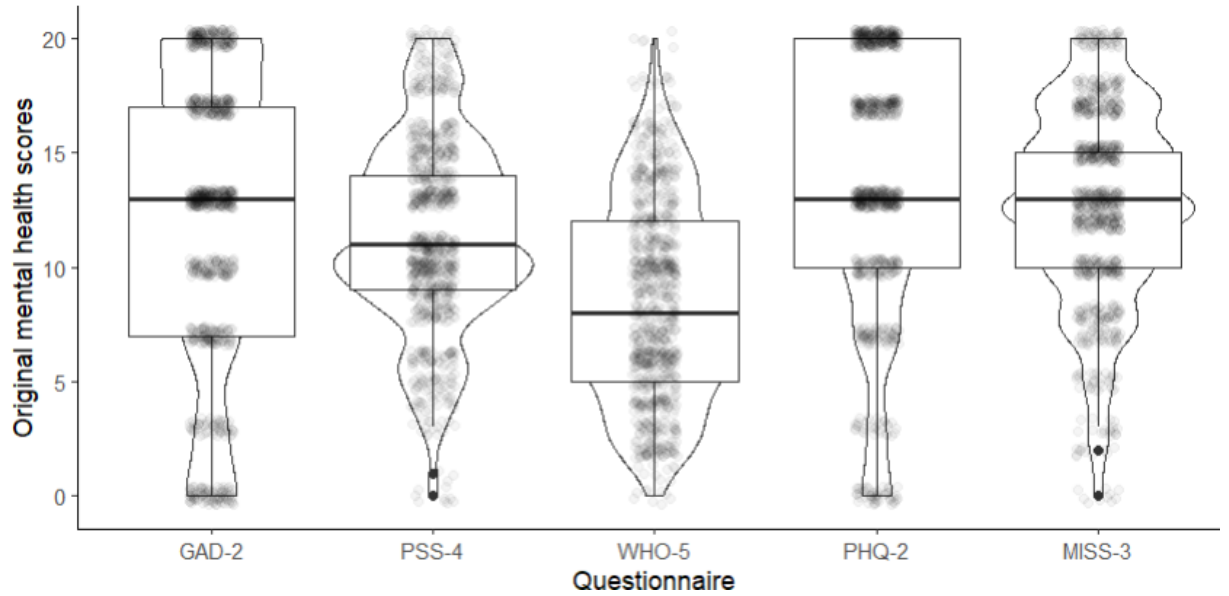


Figure 4.4: Distribution of user mental health scores in the main group for their first assessment

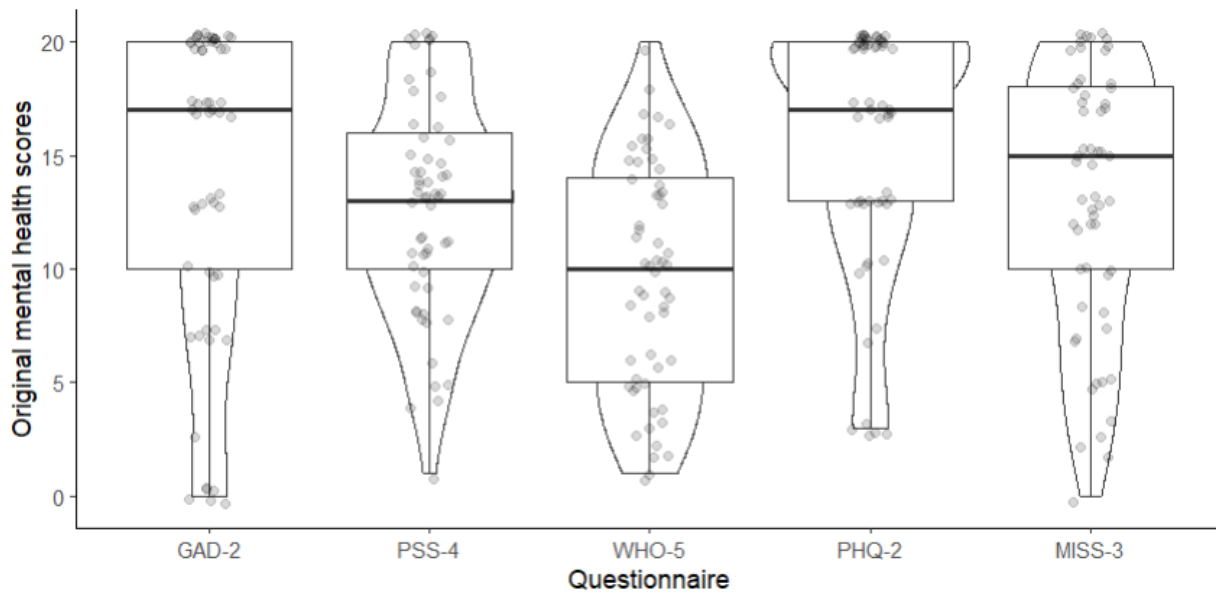


Figure 4.5: Distribution of user mental health scores in the control group for their first assessment

Ratings between all mental health scores showed moderate to strong positive correlations with each other through Spearman correlation coefficients (0.39-0.68). We further investigated their relationships through a PCA with 3 principal components that explained 87% of the variance. WHO-5 scores and PSS-4 scores loaded strongly onto component 1 (0.88 and 0.69 respectively) with a moderate loading from the PHQ-2 (0.55). The second component was comprised mostly of GAD-2 scores (with a loading of 0.90), followed by PHQ-2 (0.64) and the PSS-4 (0.52). The MISS-4 loaded onto its own component (0.95).

Change in mental health scores

Table 4.2 shows the means and standard deviations for the *change* in mental health scores between the first and second wellbeing assessment users did. The mean changes are small, but for the main group they are all positive, the first indication of mental health improvement. For all questionnaires, the mean changes were higher in the main group than in the control group, and in the control group, the PSS-4 scores and the PHQ-2 scores decreased. Figure 4.6 shows the distribution of score change for the main group and figure 4.7 shows the same for the control group. Here the spread is once more notable, but the trends are relatively similar across each of the mental health measures.

Table 4.2: Change in assessment scores for both groups

(a) Main group

	WHO-5	PSS-4	MISS-3	GAD-2	PHQ-2
Mean	0.94	0.84	0.42	1.08	0.85
SD	3.46	3.34	3.86	5.52	4.96

(b) Control group

	WHO-5	PSS-4	MISS-3	GAD-2	PHQ-2
Mean	0.05	-0.28	0.58	0.42	-0.21
SD	3.07	2.98	3.93	5.54	3.93

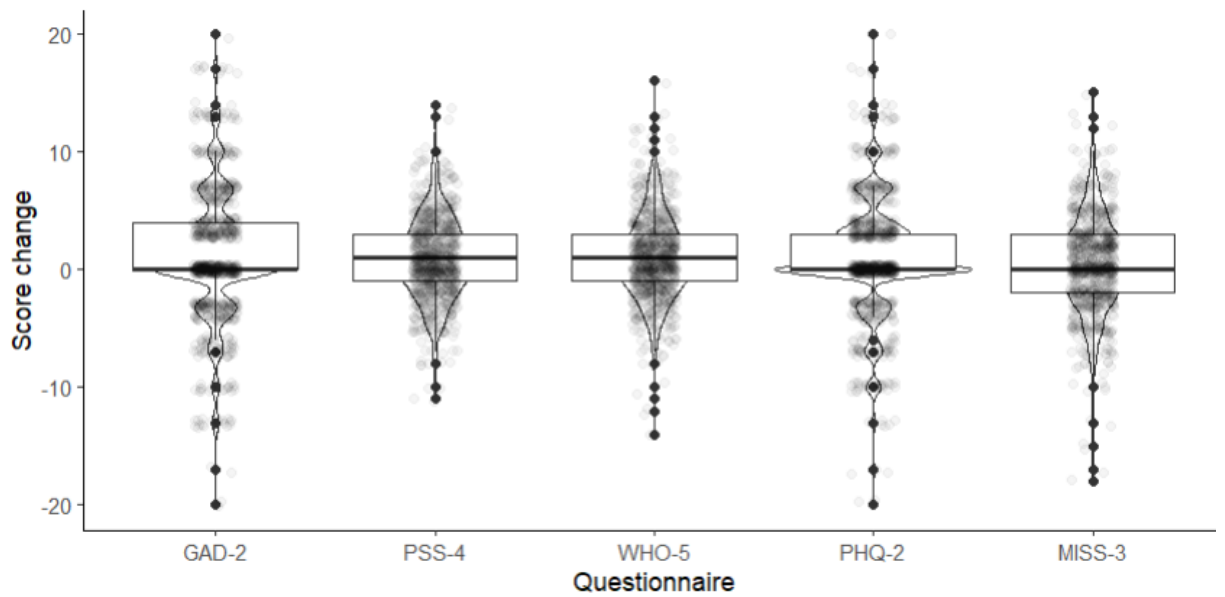


Figure 4.6: Distribution of user mental health score change in the main group

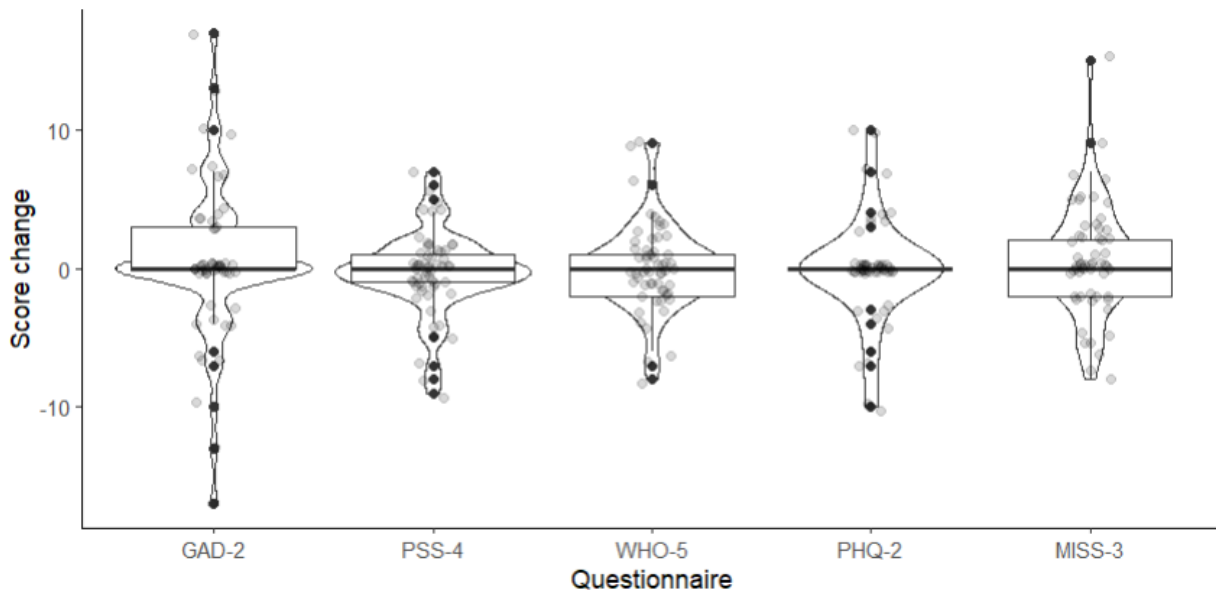


Figure 4.7: Distribution of user mental health score change in the control group

For both individual scores and change in scores, the variance was quite high, and to better understand what contributed to this variance we ran a Generalizability study (Gstudy). We used calculations as described by Brennan (2001) to calculate coefficients that described how well scores generalized across across timepoints and ratings, as well as the relative and absolute generalizability of results. We defined individual variance as Person (P) variance, which was also our facet of generalization, variance over time as Occasion (O) variance and variance across measures as Questionnaire (Q) variance for a PxOxQ Gstudy. The variance profiles can be seen in figure 4.8.

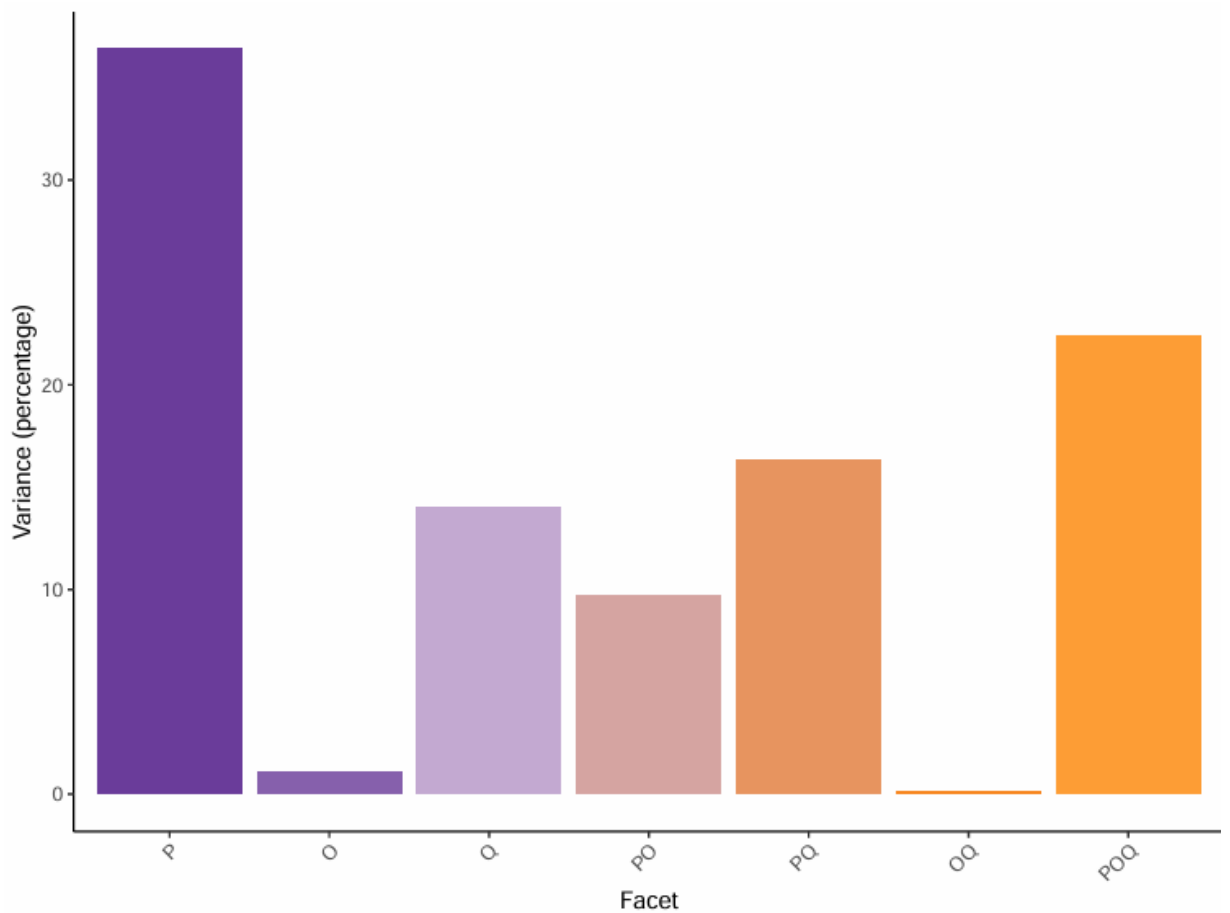


Figure 4.8: Gstudy variance profiles

Coefficients ranging from 0-1 describe how generalizable scores are across measured variables, with 1 meaning perfect generalizability. In our study, we calculated 4 coefficients. Our occasion coefficient (0.77) indicated a relatively high generalizability over time. Mental health scores, in this study, did not fluctuate massively across measurement occasions. Our questionnaire coefficient (0.55) indicated moderate generalizability across the different questionnaires. This aligns with our other analyses which showed clear differences, yet considerable overlap between mental health domains. Our relative G-coefficient (0.43) was also moderate. This is a measure of how well scores can generalize to similar settings after accounting for included facets. Typically it is used to assess the generalizability of studies in experimental settings, however here it provides an insight into the generalizability of results across these mental health domains in a natural setting, potentially useful for companies offering mental health apps catering to similar issues. Finally, our absolute G-coefficient (0.36) was low-moderate. This indicates how generalizable results are across potentially different domains. For instance, here, how much we could trust these results to translate across different related mental health domains, or potentially different questionnaires within the same mental health domains.

4.3.2 Regression models

Models for change over time

Regression results for both groups can be seen in table 4.3. For the main group, occasion was a positive significant predictor of mental health change across the board, suggesting improvement was likely in all areas of mental health. For the control group, occasion did not significantly predict mental health change for any measure.

Table 4.3: Regression results for mental health change over time in each group

(a) Main group

	WHO-5	PSS-4	MISS-3	GAD-2	PHQ-2
β	0.938	0.839	0.418	1.081	0.852
p-value	< .001	< .001	0.041	< .001	< .001

(b) Control group

	WHO-5	PSS-4	MISS-3	GAD-2	PHQ-2
β	0.053	-0.281	0.579	0.421	-0.211
p-value	0.955	0.755	0.579	0.732	0.836

Models for differences between groups

Despite the differences in the earlier models, there was only a significant difference between groups for the PSS-4 for both the bootstrapped models and models based on the imbalanced samples, although results approached significance for the WHO-5. In all cases except the MISS-3, the main group were more likely to see a higher degree of improvement. The bootstrapped results were similar to the raw results, indicating that in this case the imbalanced sample did not overly bias the findings. The regression results for both raw and bootstrapped analysis can be seen in table 4.4.

Table 4.4: Regression results for group comparison of mental health change

(a) Raw scores

	WHO-5	PSS-4	MISS-3	GAD-2	PHQ-2
β	0.886	1.119	-0.161	0.660	1.062
p-value	0.060	0.014	0.760	0.381	0.113

(b) Bootstrap samples

	WHO-5	PSS-4	MISS-3	GAD-2	PHQ-2
Mean β	0.884	1.116	-0.162	0.661	1.064
SD of β	0.462	0.441	0.509	0.732	0.653
p-value	0.056	0.011	0.747	0.375	0.105

Model for whether interval predicted change

No interval duration was significant in predicting mental health change for any measure.

Models for predictors of change

All findings from the regression models can be seen in table 4.5, with subtables for each individual wellbeing measure. There were several notable findings. First, similar to findings in Chapter 3, lower scores in a mental health measure significantly predicted positive change in the same construct. Second, the activity bucket users fell within did not significantly predict change in any of the mental health measures. Third, for every measure, positive change in the other mental health areas predicted improvement within the measure of interest, indicating that improvement occurs holistically. Finally, whilst it varied which thresholds predicted which change in mental health outcomes in different domains, there was one consistent pattern. With a single exception (WHO-5 scores negatively predicting PHQ-2 change), all original thresholds were positive predictors of change in other constructs.

Table 4.5: Regression results for predictors of mental health change

a) Predictors of change in WHO-5 (adjusted $R^2 = 0.270$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Activity bucket	0.047	0.262	0.180	0.857
WHO-5 bucket	-2.228	0.238	-9.346	< .001
PSS-4 bucket	1.317	0.252	5.221	< .001
MISS-3 bucket	0.175	0.288	0.609	0.543
GAD-2 bucket	0.593	0.302	1.960	0.051
PHQ-2 bucket	-0.089	0.342	-0.261	0.794
PSS-4 change	1.00	0.115	8.699	< .001
MISS-3 change	0.411	0.108	3.813	< .001
GAD-2 change	0.327	0.127	2.568	0.010
PHQ-2 change	0.481	0.124	3.888	< .001

b) Predictors of change in PSS-4 (adjusted $R^2 = 0.335$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Activity bucket	-0.194	0.240	-0.808	0.419
WHO-5 bucket	0.981	0.227	4.321	< .001
PSS-4 bucket	-2.344	0.222	-10.552	< .001
MISS-3 bucket	-0.273	0.264	-1.032	0.302
GAD-2 bucket	0.581	0.278	2.089	0.037
PHQ-2 bucket	0.361	0.314	1.148	0.252
WHO-5 change	0.879	0.101	8.699	< .001
MISS-3 change	0.213	0.100	2.137	0.033
GAD-2 change	0.767	0.115	6.694	< .001
PHQ-2 change	0.580	0.113	5.130	< .001

c) Predictors of change in MISS-3 (adjusted $R^2 = 0.190$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Activity bucket	0.361	0.307	1.177	0.239
WHO-5 bucket	0.201	0.293	0.685	0.493
PSS-4 bucket	0.315	0.300	1.049	0.295
MISS-3 bucket	-3.369	0.319	-10.566	< .001
GAD-2 bucket	0.598	0.355	1.684	0.093
PHQ-2 bucket	1.163	0.400	2.910	0.004
WHO-5 change	0.507	0.133	3.813	< .001
PSS-4 change	0.300	0.140	2.137	0.033
GAD-2 change	0.528	0.149	3.548	< .001
PHQ-2 change	0.437	0.146	2.997	0.003

d) Predictors of change in GAD-2 (adjusted $R^2 = 0.415$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Activity bucket	-0.164	0.373	-0.439	0.661
WHO-5 bucket	0.683	0.355	1.923	0.055
PSS-4 bucket	0.626	0.365	1.717	0.086
MISS-3 bucket	1.159	0.409	2.835	0.005
GAD-2 bucket	-5.699	0.389	-14.634	< .001
PHQ-2 bucket	2.481	0.482	5.153	< .001
WHO-5 change	0.418	0.163	2.568	0.010
PSS-4 change	1.117	0.167	6.694	< .001
MISS-3 change	0.546	0.154	3.548	< .001
PHQ-2 change	1.790	0.168	10.656	< .001

e) Predictors of change in PHQ-2 (adjusted $R^2 = 0.387$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Activity bucket	0.313	0.343	0.914	0.361
WHO-5 bucket	-0.646	0.327	-1.978	0.048
PSS-4 bucket	0.677	0.335	2.021	0.044
MISS-3 bucket	0.762	0.377	2.023	0.043
GAD-2 bucket	1.114	0.396	2.816	0.005
PHQ-2 bucket	-4.934	0.418	-11.796	< .001
WHO-5 change	0.578	0.149	3.888	< .001
PSS-4 change	0.794	0.155	5.130	< .001
MISS-3 change	0.425	0.142	2.997	0.003
GAD-2 change	1.683	0.149	3.888	< .001

Model for predictors of engagement

Findings from the regression model investigating predictors of activity completions can be seen in table 4.6. Notably, higher WHO-5 and MISS-3 scores positively predicted number of completions, whilst higher PHQ-2 scores negatively predicted number of completions.

Table 4.6: Regression results for predictors of activity completions

Independent variable	Coefficient	Std. Error	t-statistic	p value
WHO-5 bucket	0.188	0.023	8.206	< .001
PSS-4 bucket	0.022	0.023	0.953	0.341
MISS-3 bucket	0.058	0.027	2.141	0.032
GAD-2 bucket	-0.019	0.027	-0.702	0.483
PHQ-2 bucket	-0.105	0.031	-3.368	0.001

4.4 Discussion

4.4.1 Comparing mental health assessments

The current study is a good example of how it can be difficult to align real world data with that collected in experimental settings. In experimental settings, efforts are often made to design studies such that they mimic phenomena and situations in the wild as much as is possible whilst still investigating variables of interest. This study aimed to come at the same problem from the opposite direction, methodically filtering natural data to resemble what might be collected by psychologists in controlled environments. One of the drawbacks of this approach was that it reduced the sample size from thousands of participants who had completed over a hundred thousand activities to only a fraction of those numbers. However, it still left 986 users (including the control group) who had completed multiple wellbeing assessments and activities of their own volition at times entirely of their choosing. We will first look at the properties of their orig-

inal scores, how they overlapped and how different mental health constructs relate in a natural environment.

The first thing of note is the differences in distribution of mental health scores. Both the main and the control group scores showed similar trends in how they performed on each questionnaire, and this trend showed remarkable disparity between mental health constructs. For both groups, the central tendencies for MISS-3, GAD-2 and PHQ-2 were similar. The PSS-4 was slightly lower, and the WHO-5 noticeably lower. Furthermore, whilst the thresholds indicating lower mental health were the same for four of the questionnaires (50% of the total possible score), the PSS had a different cutoff (below 62.5% of the total possible score when inverted to match the Foundations scale). These differences combined mean that the MISS-3, GAD-2 and PHQ-2 are more stringent screeners for their respective mental health issues. This can be seen in our sample, as a smaller percentage of the sample had scores indicating potential insomnia, generalized anxiety disorder or depression (19%, 26% and 16% respectively) than those who had scores indicating low mental wellbeing or high stress (57% and 56% respectively). These findings are relatively consistent with the literature for studies including two or more of the same questionnaires (Ghazisaedi, Mahmoodi, Arpaci, Mehrdar, & Barzegari, 2022; M. Lara-Cabrera et al., 2022; M. L. Lara-Cabrera, Betancort, Muñoz-Rubilar, Rodríguez Novo, & De las Cuevas, 2021), and indicate either higher prevalence of mental wellbeing and stress difficulties than issues in the other domains, or a difference in how psychologists tend to quantify the issues.

Despite these differences there was also considerable overlap between the questionnaires. The Gstudy found that scores generalize moderately over the measures in this study, although there is likely more overlap between some than others. The PCA found that mental wellbeing, stress, anxiety and depression could be explained by two principal components, stress and depression contributing at least moderately to both. MISS-3, however, loaded onto its own component, without much contribution from any of the other measures. Overall, this blend of independence and overlap is to be expected. Each questionnaire measures a different construct, and they are typically used in different scenarios for different purposes. However, mental health issues often present comorbidly, and suffering from one disorder can make individuals more susceptible to another (McGrath et al., 2020; McLafferty et al., 2017).

What is perhaps more interesting is the difference between the main group and the control group in this sample. The control group, on average, had higher baseline scores than the main group for every mental health measure. These are only descriptive statistics, and the samples were highly imbalanced, so results must be taken with a grain of salt, but it does merit consideration for future research. For the study described in Chapter 3, we sampled only users with low mental wellbeing scores, as it was theorized they would be inherently more motivated to engage with the app. This has also been done in other studies for different reasons (Catuara-Solarz et al., 2022), and the present findings can be viewed as mild evidence in support of this assumption.

4.4.2 Mental health change

For the main group of participants, this study found a significant change in scores over time for all included mental health measures. This is in line with other studies investigating the efficacy of Foundations in trial studies (Catuara-Solarz et al., 2022; Gnanapragasam et al., 2023; Schulze et al., 2024), and suggests that, at least in this case, results from controlled environments reflect natural settings. This study, however, did not find that the interval duration influenced score changes for any measures, contrasting with results in Schulze et al. (2024), which found significant improvement in mental health metrics over two weeks, but not four. The methodologies differed in multiple respects, however, which may account for these differences. It may be that the difference between natural and controlled settings changed the degree to which users benefitted from app usage. Or perhaps our filtering methods, which were not in precise two-week intervals, may have lead to different outcomes. The cadence of assessments and intervals between them may play an important role in mental health change, and warrant further investigation in future studies.

The current study also did not see any significant change in mental health outcomes for the control group, suggesting that self-assessment is not enough to enact mental health change, likely requiring some interaction with activities or other solutions. However, group did not significantly predict change in scores, despite the fact that one group saw significant positive change whilst the other did not. This is likely due to the fact that whilst improvement was significant for the main group, it was small on average. The sample size of the control group (which was mirrored in the main group through bootstrapping) was likely not large enough to detect small effects. However, despite this, the findings in this study still support that there is a minimum necessary interaction with the app to enact mental health improvement. Yet we also found that beyond this minimum, it was not significant how many activities users did. This aligns with the results described in Chapter 3. It suggests once more that it may not be particularly important how many activities users do when it comes to mental health improvement, but it may matter which ones they do (Ruiz de Villa et al., 2023) or how they engage with them (Graham et al., 2021).

Even though volume was consistently non-significant in predicting mental health change across measures, there were some other notable findings. First, for each measure, change in scores was positively and significantly predicted by change in scores across the other measures. This suggests that improvement occurs simultaneously across mental health domains. This is not entirely unsurprising, as many mental health issues present overlapping symptoms (American Psychiatric Association, 2013) or occur simultaneously (McGrath et al., 2020; McLafferty et al., 2017), but the consistency across *all* measures is still noteworthy. It does have some direct parallels in the literature, as a similar effect was also stated by Ruiz de Villa et al. (2023), who found that completing activities from the sleep module of Foundations was effective in

improving mental health in all their included domains. Taken together, this speaks well for the broad potential of mental health apps.

There was also an interesting trend in how scores from the first wellbeing assessment predicted change. Here it differed which measures predicted outcomes, but there were two patterns that stood out. First, in each case, change in scores on a measure was significantly predicted by whether users scored above or below the threshold for the same measure, and in each case in the same direction. Users were, in all cases, more likely to experience a higher degree of change if they were in the low mental health bracket originally for the target variable. This was the same pattern as was noted in Chapter 3, and once more there are two likely explanations, possibly working in tandem. It may be that users with more to gain are more likely to benefit when provided with relevant solutions. This is a logical line of thought, and also why it is common for mental health apps to target specific mental health subsamples (Marshall et al., 2020; Rodriguez-Paras et al., 2017; Schulze et al., 2024). It is also possible that measurement bias exists, as users who score lower have less downward movement potential, and conversely that users who score higher have less upward movement potential.

The second apparent trend was that mental health change was significantly predicted by the original scores in other questionnaires. This was not as consistent as previously stated results. For example, WHO-5 change was only significantly predicted by original PSS-4 scores whilst PHQ-2 change was significantly predicted by scores on all the other questionnaires. However, what is notable is the directionality. When significant, in all cases but one, users were more likely to improve in a mental health domain if they scored highly in other domains originally. This suggests that better mental health levels (possibly general, possibly domain specific) facilitate improvement across different measures, and that mental health improvement generalizes across constructs. This also aligns with the notion that change in one area coincides with change in another, and potentially indicates a positive spiral of enabling further improvement.

4.4.3 Limitations

Working with data obtained naturally and commercially has its value, yet also comes with several limitations. First, every choice made had an opportunity cost. In this sample we chose to filter users and assessments by week-based parameters, yet many other choices could have been made that might have lead to different results. Many users had completed assessments between the two we had selected, as we picked those closest to two-week markers. It is possible that these repeated measures contorted results as users familiarized themselves with the assessments. However, as there has not always been a strictly reinforced minimum waiting period between assessments in Foundations, we also may have experienced issues had we simply picked the first two assessments users did, as in some cases they are less than a day apart. Selecting specific thresholds allowed for more control in some respect and less in others.

There was also the potential of activities completed previous to the first wellbeing assessment acting as a confound. We partially accounted for this, as we included only users who had completed an assessment within two days of onboarding, which minimized access to the app but didn't completely remove it. Again, a different filtering method could have been chosen, such as only selecting users who had finished a wellbeing assessment before an activity, however as activities are usually one of the first things users engage with in the app, this would have severely limited our sample size.

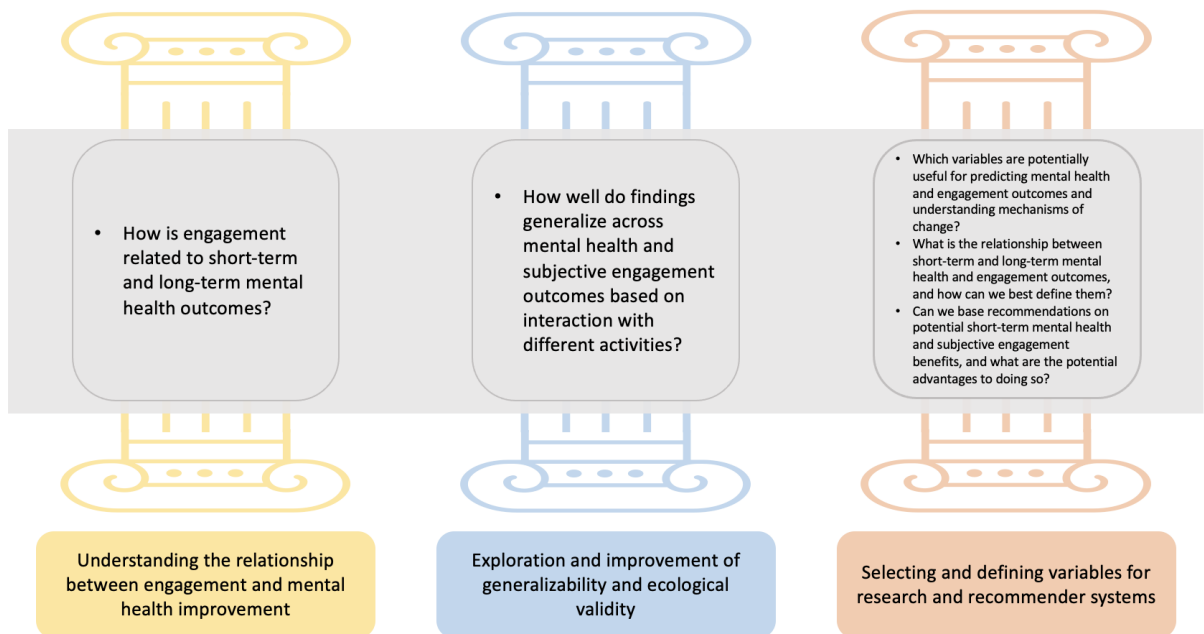
The final confound that merits mentioning is the breadth of time this data was collected over. Some of the first wellbeing assessments were taken during the Covid-19 epidemic, which has generated an entire sub-field of mental health research, and lead to a spike in the popularity of mental health apps. There is certainly a chance that this influenced users' mental health, the time they had available to complete activities or any number of other factors. Furthermore, the time range meant that data was collected across different versions of Foundations. Any of this could have made a difference, and therein lies the strength of more controlled research settings. A balance is required to investigate specific phenomena whilst ensuring real world value.

4.4.4 Conclusion

This study found that results regarding mental health efficacy and predictors of improvement in a commercial setting largely align with those in academia. App usage significantly predicted positive mental health change across multiple variables, suggesting that when mental health apps work, they work broadly. It found that whilst mental health measures are distinct, there is sufficient overlap that a holistic approach may be possible to leverage small, but consistent improvement across domains. However, it also found that volume of activities did not significantly influence improvement, and more research is therefore required to understand the specific features of mental health apps that enable this change.

Chapter 5

Establishing predictors of improvement



5.1 Introduction

This chapter describes efforts to better understand what drives mental health change. Understanding what factors contribute to improvement (or lack thereof) from engaging with mental health interventions has many benefits. However, the range of potential factors is also extremely wide. There are person-based factors, such as personality traits, differing levels of mental health and wellbeing, different disorders and motivation. There are intervention-based factors such as medium (whether delivered through phone or laptop), the type of technique (such as mindfulness or cognitive restructuring), the method of delivery (audio or video) and duration of activity. Situational factors may also vary and things like the time of day, emotional states or surround-

ing distractions could all change the way a user interacts with an app. And the interactions themselves are also potential predictors of improvement, with the frequency and duration of engagement all adding to the list of things that might help a user maintain or improve their mental wellbeing.

However, this also raises the additional question of lasting improvement. Collecting momentary snapshots of user moods or mental states has become increasingly possible and increasingly popular (Becker et al., 2018; van Breda et al., 2016) yet the relationship between short-term and long-term improvement is still unclear. Whilst some research has shown that it may be possible to predict long term mental health outcomes based on momentary assessments (van Breda et al., 2016), others have found that the temporal connection is tenuous (Knight & Emery, 2022). Beyond the link between short-term and long-term improvement, it is also worthwhile understanding and targeting momentary improvement alone. Short term relief and symptom alleviation can be valuable in its own right, even if it does not lead to consistent improvement. Some studies focus on this short-term value. For example Beltzer et al. (2022) evaluated an algorithm which recommended emotion regulation techniques which were predicted to be the most effective in the moment, however the premise often seems to be that the purpose of short-term improvement is to effect long-term change (Lewis et al., 2022; Rohani et al., 2021).

The current study attempted to evaluate these relationships. It investigated whether personality traits, socioeconomic factors and mental health factors contribute to mental health outcomes at different points of time. The primary long-term mental health outcomes were mental wellbeing and stress. Several of the included variables were similar to those included for the study described in Chapter 3. The present study, however, intended to dive deeper, and new potential predictor variables were selected or refined through exploring related literature. The first choice was to include measures of personality. Different personality traits have been strongly linked with both mental wellbeing (Schneider, Matic, Buda, & Dolan, 2024; Sirgy, 2021; Zotova & Karapetyan, 2018) and stress outcomes (Liu, Lithopoulos, Zhang, Garcia-Barrera, & Rhodes, 2021; Luo, Zhang, Cao, & Roberts, 2023). Furthermore, personality has also been researched within mental health apps. Different personality traits have been successfully used to predict which activity types and formats users were more likely to engage with (Alqahtani, Meier, & Orji, 2022; Khwaja et al., 2021), in many cases outperforming self-reported preferences (Khwaja et al., 2021). The reverse is also possible, predicting user personalities based on engagement and behavior patterns (Xu, Frey, Fleisch, & Ilic, 2016). Different combinations of personality traits have also been used to identify user engagement archetypes. Aziz et al. (2023) described how high neuroticism, moderate extraversion and low life satisfaction was a common combination in "help-seeking users", who tended to launch a mental health app almost daily. In contrast, extraverted, emotionally stable "maintenance users", launched the app infrequently, and Aziz et al. (2023) suggested these users only engaged when they felt the need. These studies, however,

focused primarily on engagement outcomes, and positive mental health outcomes were either discussed in theoretical and potential terms, or simply assumed as a result of increased engagement. The relationship with mental health improvement via apps is still relatively unexplored, something the current study aimed to address.

It also aimed to clarify mechanisms of change over time, considering some of the findings in previous chapters. Volume of activity completions was found to be an ineffective predictor of improvement in the studies described in both Chapter 3 and Chapter 4, and was included in the current study to see if this finding could be replicated once more. However, it also aimed to explore other factors of engagement, specifically subjective engagement and long term effects. Participants were therefore asked to rate perceived effort and satisfaction after each activity completion as measures of subjective engagement which have previously been linked to mental health outcomes (Graham et al., 2021; van Agteren et al., 2021). Participants were also asked to complete a 4-week mental health follow-up, after they had their access to the app removed. This was included to investigate whether app benefits were maintained even after users stopped using it, and to begin exploring what the potential minimum engagement criteria are to improve mental health through apps.

We also aimed to refine the post-activity ratings described in Chapter 3 to gain a better idea of user states, and specifically the perceived effect of mental health activities on these states, and how these were linked to long-term change. Some studies have already attempted to leverage information on how individual activities affected user moods to generate recommendations in mental health recommenders (Niles et al., 2020; Rohani et al., 2021), however they did not include measures on trait-level mental health outcomes. Personalized recommendations tailored to improving user moods is a worthwhile endeavour, however it needs to be established whether this translates into long-term gain, especially as moods and states are volatile, prone to fluctuation. (da Fonseca, Maffei, Moreno-Bote, & Hyafil, 2023). The current study also aimed to go beyond a single rating of user mood, and capture information on the effect of mental health activities on participant states related to mental wellbeing and stress traits. There was a large number potentially significant variables identified, however only a few could realistically be included without inducing fatigue in the participants and drawing focus away from the activities themselves. In the end we settled on seven. These included perceived effort and satisfaction, for reasons described above, direct ratings of perceived stress and wellbeing improvements based on the completed activity, and measures of cheerfulness, social connectedness and coping efficacy based on findings by Lebois et al. (2016) and by Ruggeri, Garcia-Garzon, Maguire, Matz, and Huppert (2020).

This study aimed to provide a better understanding of mechanisms of change, short-term and long-term, to address the gaps identified in chapters 1 and 2. It aimed to do this by exploring different operational definitions of variables, both as predictors and outcomes, or potential in-

puts and outputs for recommender systems. By including both engagement and mental health metrics, it also aimed to understand the relationship between the two, and where potential recommendations might synergize or deviate in their ability to facilitate an improvement across the separate goals. Finally, it aimed to understand how mental health outcomes were generalizable over time, and across different operational definitions, and whether completion of different activities would lead to different results.

5.2 Methods

5.2.1 Participants

This study aimed for a gender-balanced sample of 150 participants above 18 years of age, who had not taken part in any related studies on Prolific. Participants must have completed at least 10 Prolific studies with an approval rate of at least 95%. There were no mental health criteria for this study, as opposed to that described in Chapter 3, as we were interested in results from the wider population. To qualify for inclusion in the final dataset, they had to complete a series of onboarding questionnaires, 4 short mental health activities through an experimental app over the course of 2 weeks, and a set of exit questionnaires. There was also an optional (paid) follow-up 4 weeks after completion of the exit questionnaires. Participants were paid £8 for completing the main part of the study, and £2 for the optional follow-up. 113 participants (58 female, 52 male, 1 non-binary and 2 "other") completed all mandatory requirements, and of those, 86 (40 female, 45 male and 1 "other") completed the optional follow-up. The mean age for the main group was 41.51 years (SD = 12.63) and the mean age for users who completed the optional follow up was 42.33 (SD = 12.63).

Ethics approval for this study was received from the College of Medical Veterinary and Life Sciences at the University of Glasgow (application 200220250).

5.2.2 Procedure

Participants were sampled via Prolific and referred to the Qualtrics platform where they completed their first set of questionnaires, were briefed on the nature of the study, and provided written informed consent. Participants were sampled over the weekend, starting on Friday, but only received access to the online platform (implemented in "shiny" R, hosted by <https://www.shinyapps.io>) the following Monday, and kept access for two weeks. They received a study participation reminder on the day they got access along with their login information, and biweekly reminders after that via Prolific. After logging on to the platform, participants were shown a library of 8 possible audio activities they could choose from, presented with its title and approximate length in minutes. Once they chose an activity and pressed play, they would be able to pause, but not

skip or increase the playback speed. Upon completion of the activity, they would click "continue" and provide 7 ratings of their state change and perception of the activity. After clicking "finish" their answers would save and they would be redirected back to the library page. They were required to complete a minimum of 4 activities to qualify for payment, but they were free to complete those activities at any time of their choosing during those two weeks.

On day 14 participants would complete the exit questionnaires, and would have the opportunity to provide qualitative feedback about the app, the activities or the study. They were thanked for their participation, debriefed and reimbursed with a reminder that they could complete the optional follow-up 4 weeks later. For the participants who chose to return to complete the follow-up, they were reminded of the study purpose, filled in questionnaires, were thanked for their continued participation and reimbursed.

5.2.3 Measures

The HEXACO-60

The 60-item HEXACO (Ashton & Lee, 2009) instrument assesses personality traits along the six dimensions of honesty, emotionality, extraversion, agreeableness, conscientiousness, and openness. Items were rated on a continuous self-report scale with one decimal point precision, ranging from 1 (strongly disagree) to 5 (strongly agree). All subscales showed satisfactory test reliability ($\alpha = 0.73-0.80$) and test-retest reliability ($r = 0.82-0.89$) in a large adult sample (Henry, Thielmann, Booth, & Möttus, 2022).

The WEMWBS-14

The Warwick-Edinburgh Mental Well-being Scale 14-item version (Tennant et al., 2007) covers both hedonic and eudaimonic aspects of mental health including positive affect (feelings of optimism, cheerfulness, relaxation), satisfying interpersonal relationships and positive functioning (energy, clear thinking, self-acceptance, personal development, competence, and autonomy). Items were rated on a continuous self-report scale with one decimal point precision, ranging from 1 (none of the time) to 5 (all of the time). Both scales (7-item version and 14-item version) are considered to be robust and valid when applied in population, community, educational, occupational, and clinical settings (Stewart-Brown, 2015). The original 14-item WEMWBS provides a fuller picture of mental wellbeing with a better balance of feeling and functioning items than the 7-item WEMWBS.

The PSS-10

The 10-item Perceived Stress Scale (Cohen et al., 1983) assesses how much distress individuals experienced in the last month. Participants rated items such as "In the last month, how often have you been upset because of something that happened unexpectedly?" on continuous self-

report scales with one decimal point precision, ranging from 0 (never) to 4 (very often). As gold standard in the field, the PSS-10 showed acceptable internal consistency and test-retest reliability across various studies (E.-H. Lee, 2012).

Note: It is important to remember when reading results that scores on the two scales are scored in opposite directions. Higher WHO-5 scores indicate higher mental wellbeing, whereas higher PSS-10 scores indicate higher stress. For the discussion these will be referred to generically and should not be an issue, however the results section may be confusing if this is not taken into account.

Socioeconomic characteristics

Participants provided socioeconomic information on their highest educational level, annual income, and subjective social status. Participants self-reported their years spent in formal education ranging from 0 to 20 years or more. Participants categorised their annual income into a £10,000 increment, ranging from “below £10,000” to “above £50,001”. Participants lastly reported their subjective social status on the MacArthur Scale (Adler et al., 2000). This measure consists of a 10-rung ladder representing where people stand in the United Kingdom, with a higher rung indicating a higher subjective social status. In a multi-ethnic sample, the MacArthur Scale exhibited adequate test-retest reliability (Operario et al., 2004).

Activities and in-app ratings

The web app was developed in Rstudio, using R version 4.3.1 using the R “Shiny” package. It was deployed via <https://www.shinyapps.io>. Once participants were given the link to the platform, they were required to log in using their prolific ID each time they accessed the application. When participants completed activities and submitted ratings, the session data would be sent to a password protected google sheets document which could only be accessed by the researchers in this study.

The app was based on the prototype described in Chapter 3, however some modifications were made. First, to avoid influencing participants’ selection of activities, no random recommendations were made, and they were instead presented with the library seen in figure 5.1. There were also only 8 activities, compared to the original 21, and some of them had changed. This change was for a few reasons. First, we wanted to limit the sample size of activities to get more data for the remaining ones. Second, in the study in Chapter 3, we saw that participants were far more likely to select the shortest activities, and we therefore only included activities less than 10 minutes long. Finally, to be able to compare the effect of different activity types, we included options from two new modules. Three items from "Meditation and Mindfulness" remained, but there were also three from "Relaxation Techniques" and two from "Positive Psychology".

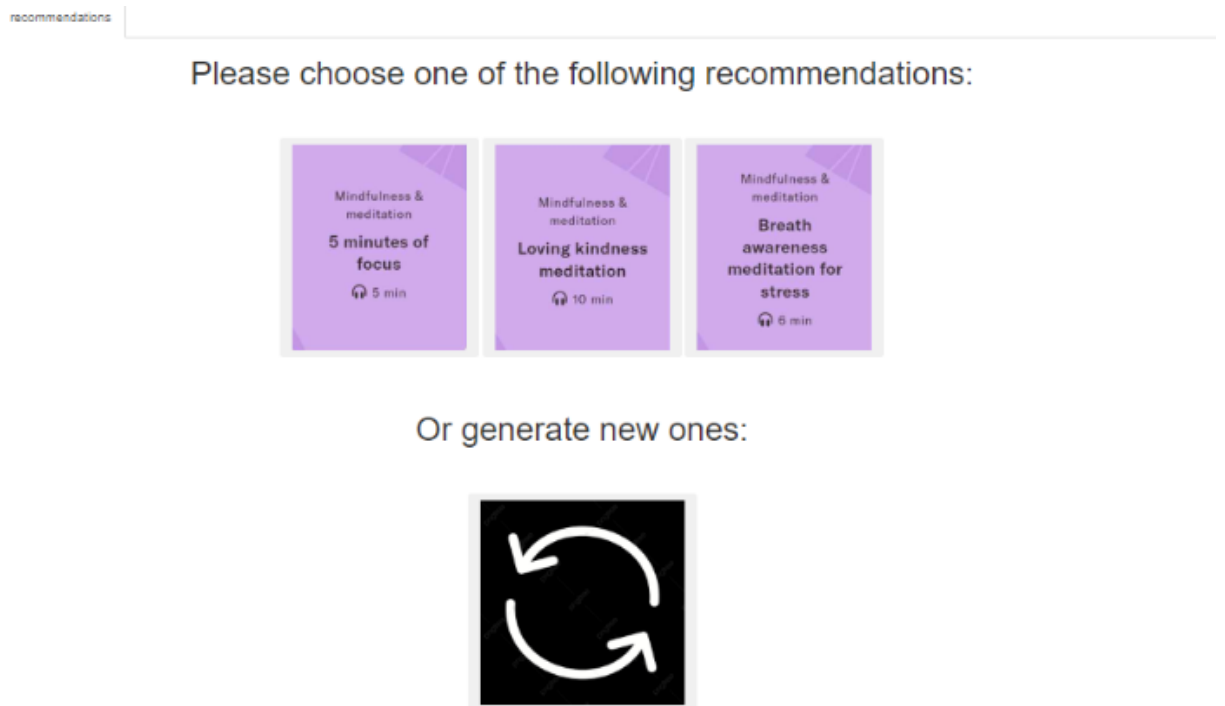


Figure 5.1: Screenshot of the library page for the updated app described in Chapter 5.

The visuals for the activity page and the ratings page remain as shown in Chapter 3. However, the ratings themselves changed as explained in the introduction. With the exception of the ratings pertaining to effort and satisfaction, responses were also phrased to indicate perception of *change* due to the activity, so we could investigate perceived improvement from the activity. In this way we differed from the mood ratings used for recommendations by Lewis et al. (2022) and Rohani et al. (2021), and this was done to isolate the perceived value of the activity rather than participant states, which might be more heavily influenced by situational factors. The stress rating was the only rating that was negatively scored (with higher scores indicating a deterioration not improvement of a mental health state) and so stress scores were inverted for analysis. As in Chapter 3, the order of the ratings was randomized, but the new ratings themselves can be seen in figure 5.2 and 5.3.

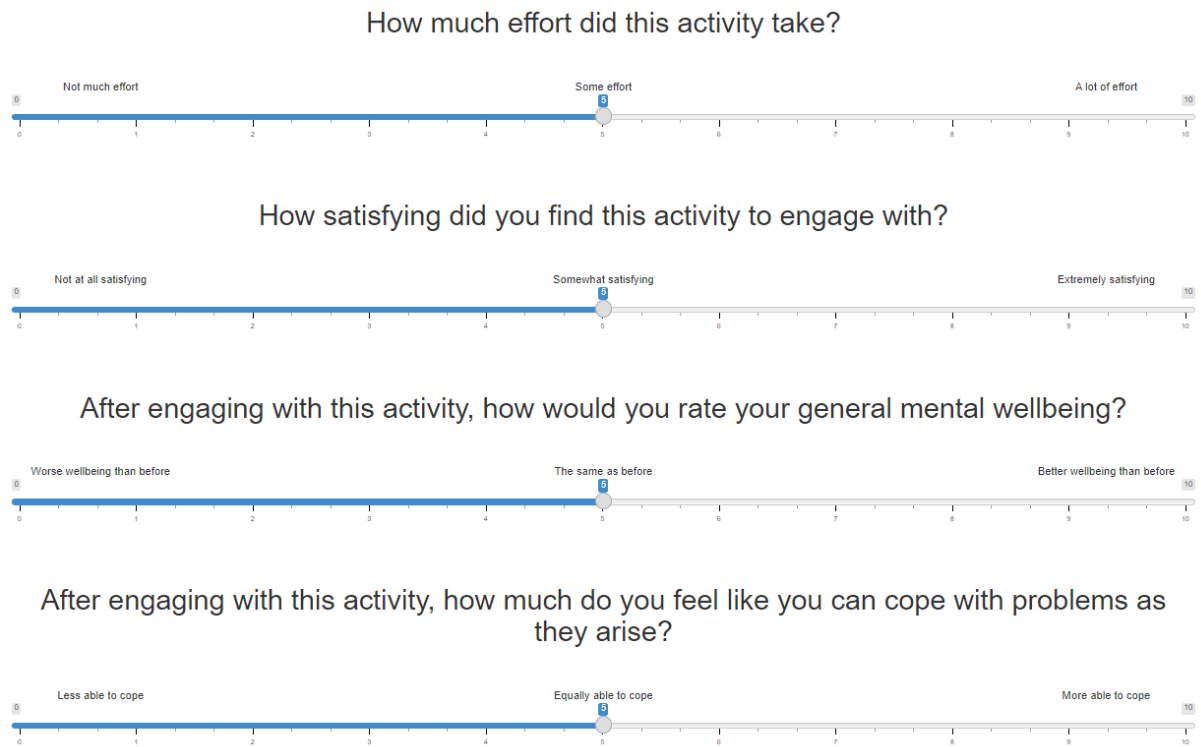


Figure 5.2: First half of the ratings page for the updated app described in Chapter 5

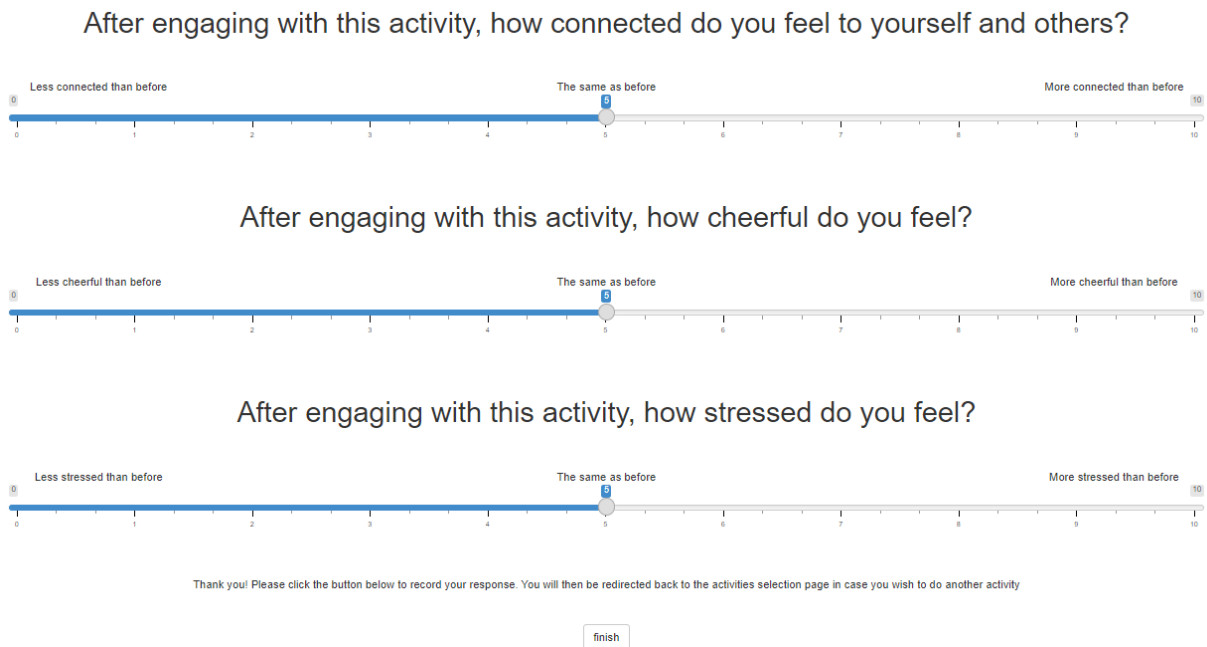


Figure 5.3: Second half of the ratings page for the updated app described in Chapter 5

5.2.4 Analysis

A wide range of different models were used in this study. Outside of descriptive statistics employed to understand the sample, they fell into four categories of analysis:

1. Psychometric analysis of the post-activity ratings.
2. Analysis of post-activity ratings and their potential predictors.
3. Analysis of offboarding data, and potential predictors of change between onboarding and offboarding.
4. Analysis of follow-up data and potential predictors of change between follow-up scores and onboarding or offboarding scores.

For the psychometric analysis, we calculated the Intraclass Correlation Coefficients (ICCs) ICC2 and ICC2k for the post-activity ratings for each activity to examine inter-rater agreement agreement. We then ran a PCA to investigate the overlap between ratings (also extracting component scores for use in subsequent analysis), and finally a Person x Activity x Rating (PxAxR) Gstudy.

For the analysis of post-activity ratings, we ran a number of regression models in two parts. For the first part we investigated whether information pertaining to the individual activities predicted the ratings for the individual activities. For this analysis, we were also interested in whether models including activities were significantly better than models that did not, and therefore ran an omnibus test comparing model coefficients. For the second part, mean scores were calculated for post-activity ratings, and our regression models focused on whether these could be accurately predicted using information from the onboarding questionnaires.

The analysis of offboarding and follow-up data was done via regression models, which will be presented individually in the results section. For all regression analyses we standardized all continuous predictors and performed mean-centred deviation-coding on all categorical predictors. We ran multiple linear regression analyses where post-activity ratings or mental health outcomes were our dependent variables, and a generalized linear model with family set to "binomial" where activity completions (separated into two buckets, the minimum requirement of 4 completions and above 4 completions) was our dependent variable. Furthermore, similar to what was described in Chapter 3, we ran a PCA using varimax rotation for our socioeconomic variables, and used component scores instead of raw scores any time they were included in subsequent analysis. Finally, for some of the latter models, we utilized the same model selection methods described in Chapter 3. This will be highlighted in the relevant parts of the results section.

5.3 Results

5.3.1 Descriptives and completion metrics

113 participants completed all mandatory parts of the study. Between them, they completed 685 activities (mean number of completion per person: 6.06, $sd = 2.23$, median = 6). Figure 5.4 shows the distribution, and figure 5.5 shows the relative popularity of each activity.

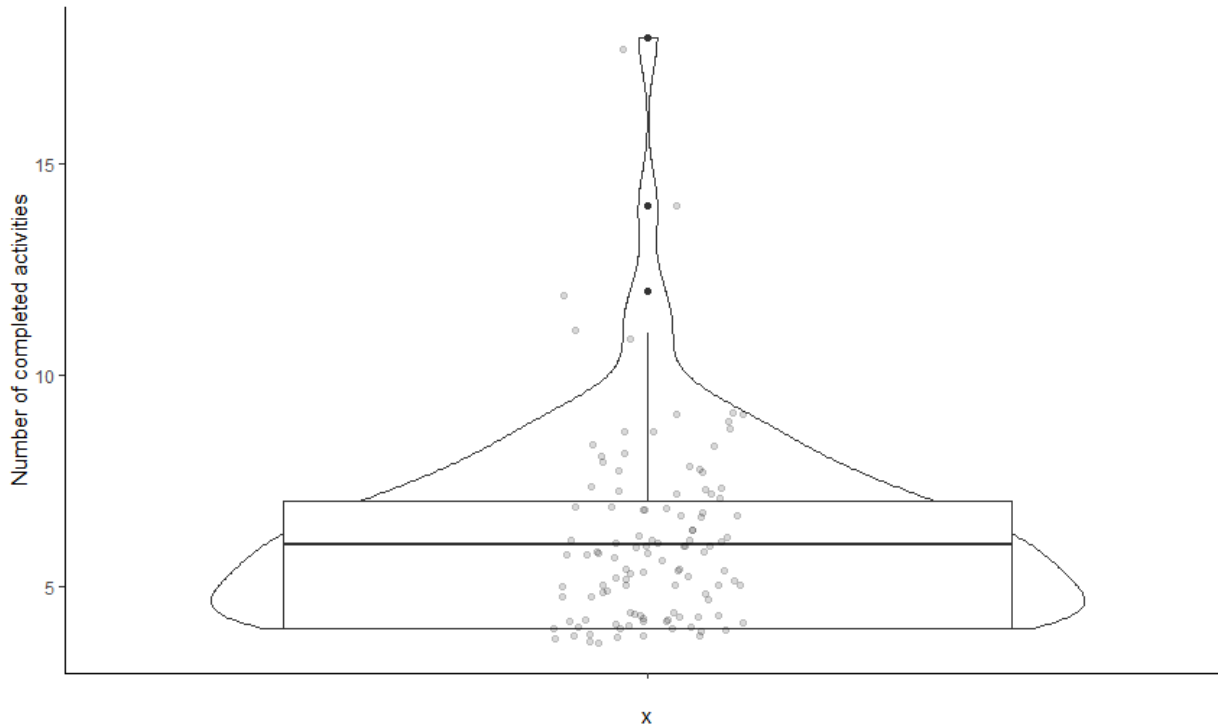


Figure 5.4: Distribution of activity completions

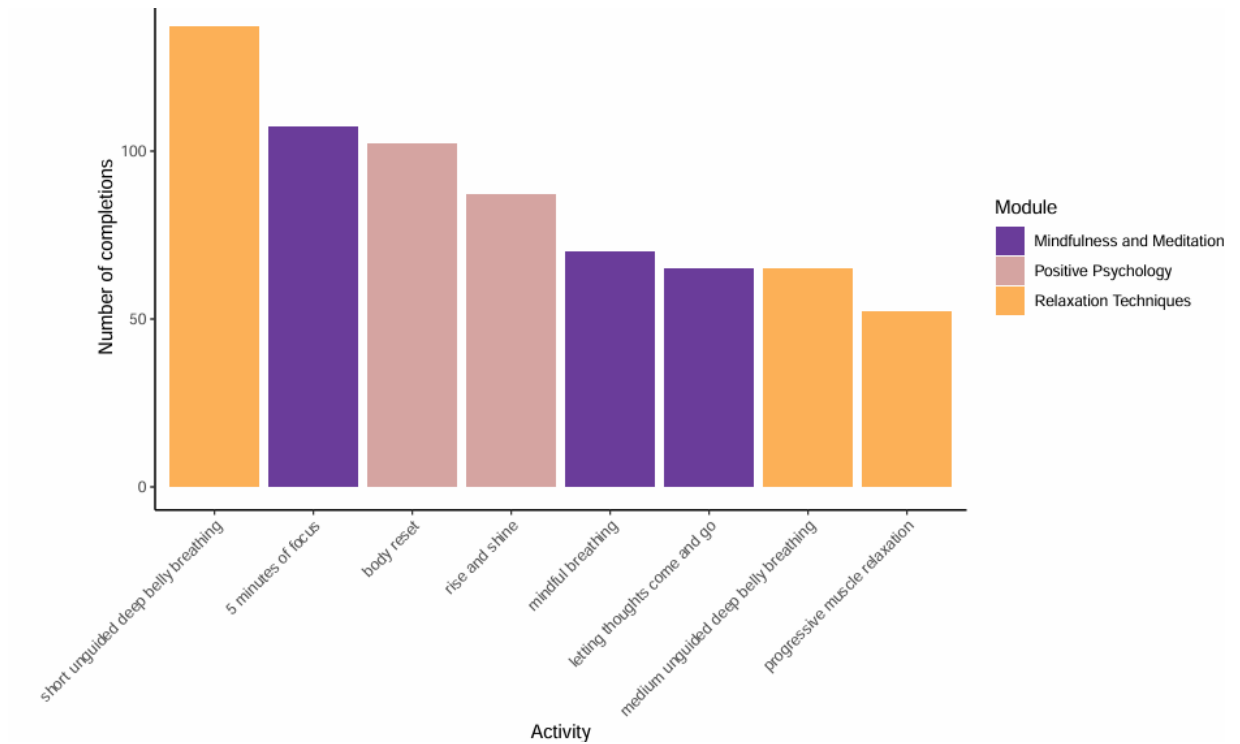


Figure 5.5: Bar chart showing the total number of activity completions by activity and type

The different post-activity ratings (figure 5.6) showed some interesting patterns. For each rating, 5 represented a neutral midway point, but the distributions and central tendencies differed. Effort was the clear anomaly with the only mean score below 5, and an argument could have been made for reversing the effort scores, however due to the findings presented in Chapter 3, we did not consider higher required effort to be inherently negative. However, beyond the central tendency it also differed in spread, with a wider distribution of scores than the other ratings. Activity satisfaction also saw some lower scores compared to the five mental health related ratings, yet it still had the highest mean score. It is also positive to note that relatively few participants rated any activity below 5 for cheerfulness, connection, coping, satisfaction, stress relief and wellbeing, as scores below 5 would have indicated a perceived deterioration in these factors stemming from the activity completions. It would, however, be worthwhile trying to understand what made those few participants feel worse as a result of completing activities.

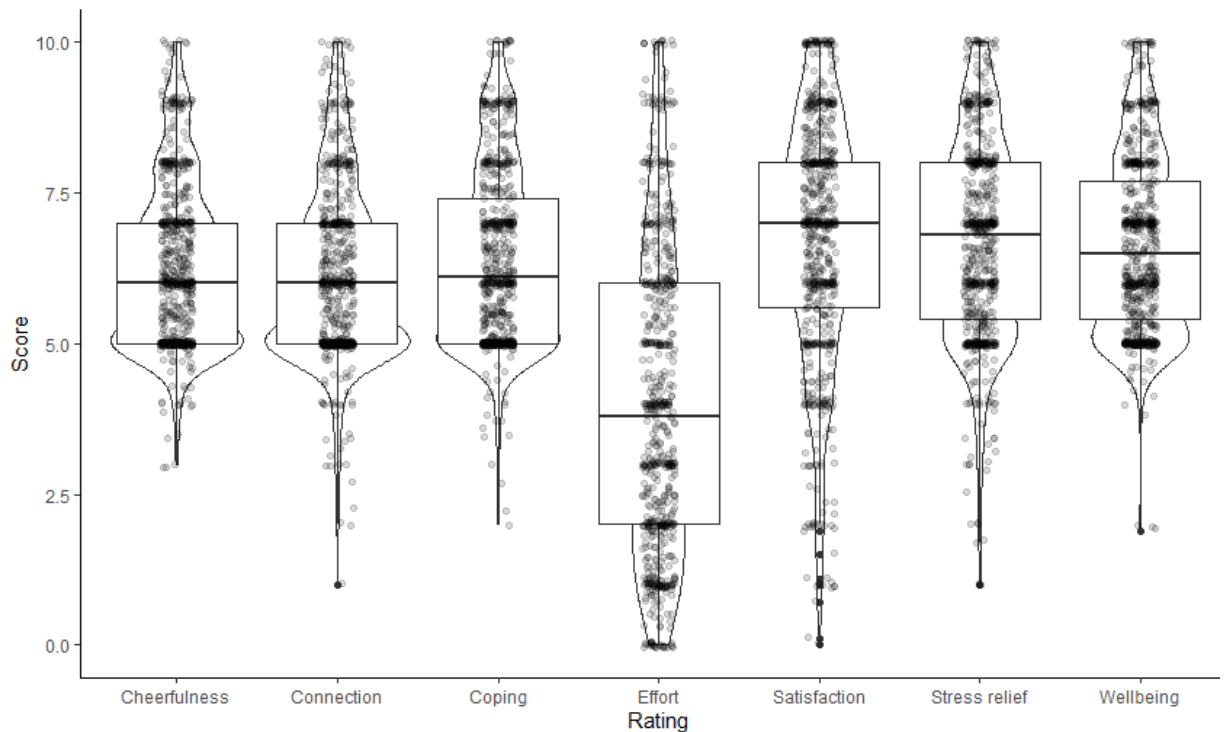
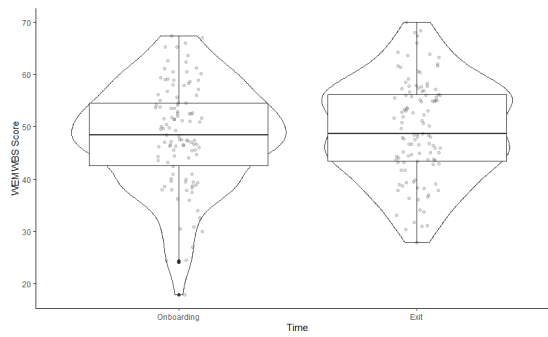
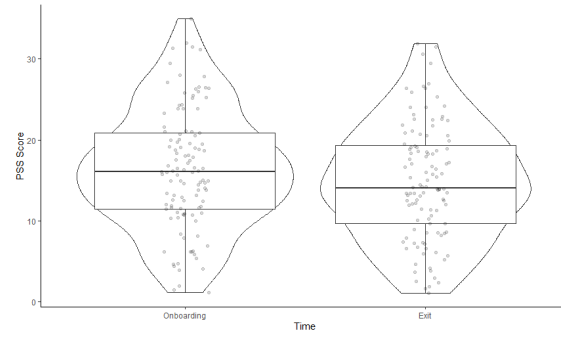


Figure 5.6: Distribution of scores for post-activity ratings

The distribution of mental health scores from onboarding to exit two weeks later can be seen in figure 5.7a and 5.7b. In both cases there was an average improvement from the first to the second occasion, with the mean WEMWBS-14 score changing from 48.26 (sd = 9.99, median = 48.4) to 49.45 (sd = 9.42, median = 48.7) and the mean PSS-14 score changing from 16.39 (sd = 7.40, median = 16.1) to 14.83 (sd = 7.12, median = 14.1). The distribution of scores in the smaller subsample of participants who also completed the follow-up questionnaires (n=86) can be seen in figure 5.7c and 5.7d. Here, the mean WEMWBS-14 score increased from onboarding (mean = 48.06, sd = 10.28, median = 48.85) to offboarding (mean = 49.33, sd = 9.59, median = 48.6) and again at follow-up (mean = 49.79, sd = 9.77, median = 51). For the PSS-10, however, mean scores improved from onboarding (mean = 15.92, sd = 7.30, median = 15.4) to offboarding (mean = 14.56, sd = 7.17, median = 13.9), but deteriorated at follow-up (mean = 15.29, sd = 7.09, median = 13.9). For a representation of the distribution of score changes for the individual participants in a similar manner, see figure 5.8. These differences will also be further evaluated later on in regression analyses.

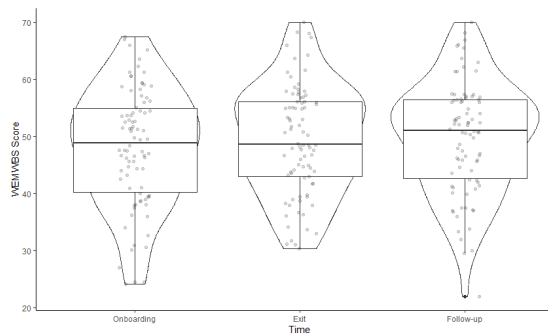


(a) WEMWBS-14 scores at onboarding and exit.

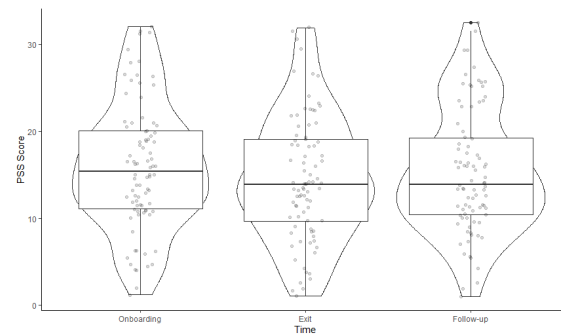


(b) PSS-10 scores at onboarding and exit.

Figure 5.7: Mental health scores at onboarding and exit

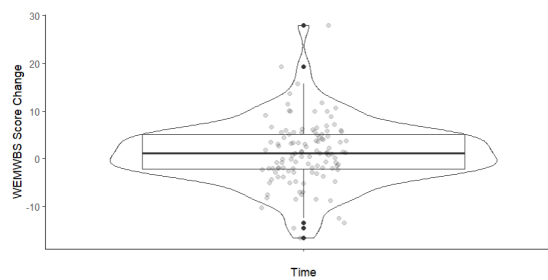


(c) WEMWBS-14 scores at onboarding, exit and follow-up.

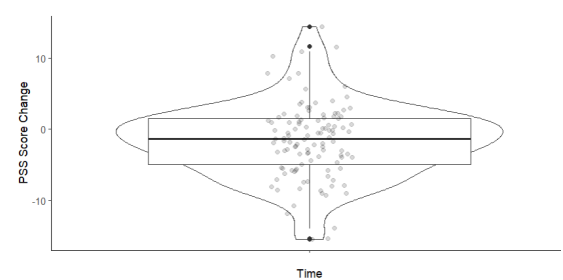


(d) PSS-10 scores at onboarding, exit and follow-up.

Figure 5.7: Mental health scores at onboarding, exit and follow-up

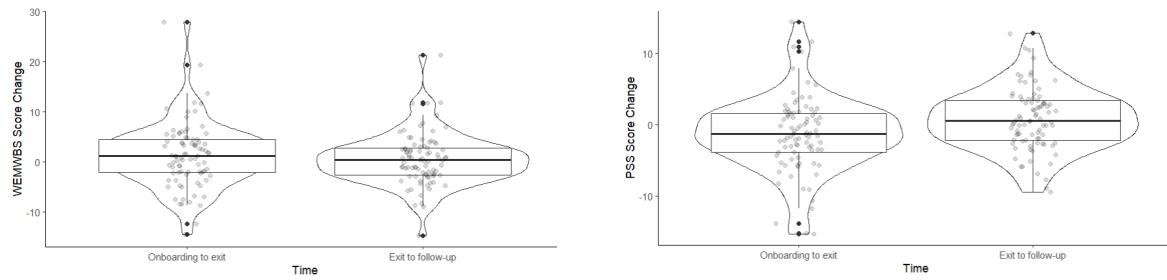


(a) WEMWBS-14 score change from onboarding to exit.



(b) PSS-10 score change from onboarding to exit.

Figure 5.8: Mental health score change from onboarding to exit.



(c) WEMWBS-14 score change from onboarding to exit and exit to follow-up.

(d) PSS-10 score change from onboarding to exit and exit to follow-up.

Figure 5.8: Mental health score change from onboarding to exit and exit to follow-up.

5.3.2 Psychometric analysis of the post-activity ratings

Intraclass Correlation Coefficients

First we explored data from 113 participants who had completed 685 activities between them, with the least popular activity being completed 52 times and the most popular activity being completed 137 times. For each of the 9 activities, and across all activities we computed a pair of Intraclass Correlation Coefficients (ICCs), ICC2 and ICC2k (see table 5.1). Both ICC2 and ICC2k are two-way models with users considered random effects. The difference between the two are whether reliability is estimated based on any single user rating (ICC2) or the average of k user ratings (ICC2k). For this analysis, we also averaged scores for any individual activities they did more than once ($n = 94$), as rater agreement and consistency is more likely to be biased by unequal samples of ratings for each user. The coefficients presented were therefore calculated from a sample of 591 activity ratings from the same sample of 113 participants. For all other analyses which included individual rating parameters, all 685 activity scores were used.

Table 5.1: Intraclass Correlation Coefficients

Activity	ICC2	ICC2k
5 minutes of focus	0.246	0.967
Body reset	0.347	0.980
Letting thoughts come and go	0.081	0.831
Medium unguided deep belly breathing	0.115	0.884
Mindful breathing	0.261	0.958
Progressive muscle relaxation	0.118	0.858
Rise and shine	0.255	0.964
Short unguided deep belly breathing	0.266	0.975
All activities	0.223	0.994

The ICC2k coefficients are all high (> 0.8), indicating that aggregate scores are reliable, however the relatively low ICC2 coefficients (< 0.4) suggest that there is high individual variance for the items within. The low ICC2 scores were no surprise, as the individual ratings were intended to

be capturing different constructs, and highly dependent on the rater. The ICC2k being higher was also expected, as it tends to reduce person-based variance, but the coefficients here were notably high, in some cases coming close to 1. This suggests that the ratings still capture a single underlying construct, or perhaps an overall behavioral pattern relating to how participants engaged with the questions overall. There were also a few activities that stood out with lower reliability, although this was possibly related to the sample sizes, as the lowest 3 coefficient pairs were also for the three least popular activities.

Correlation and PCA

For the PCA on the socioeconomic measures *income*, *education* and *perceived social status*, variables loaded onto two principal components which explained 87% of the variance. As in Chapter 3, income and perceived social status loaded strongly onto a single component (loadings over 0.8 for both) and education loaded strongly onto its own component (0.99). The component scores were used for all subsequent regression analysis. We will refer to these as the "*education component*" and the "*income and status component*" for the remainder of this study.

For the post-activity ratings, all variables showed moderate to strong, positive correlations with Spearman correlation coefficients between 0.45 and 0.75 with the exception of perceived effort, which had negative correlations across the board, with the strongest correlation being -0.25. We once more excluded effort ratings from the PCA (keeping the scores separate for further analysis), including the six other variables. These six ratings loaded onto 3 components which explained 85% of the variance. To aid with unbiased interpretation, we gave ChatGPT the information about the raw ratings and how they loaded on the extracted components, and asked it to name the three components. For the first component which included high loadings from the satisfaction rating (0.85), the cheerfulness rating (0.69) and the wellbeing rating (0.64), with a moderate loading from the coping rating (0.50), it suggested the name "*Positive emotional experience and satisfaction*". For the second component which only the stress rating loaded strongly onto (0.92), it suggested the name "*Stress level*", and for the final component which saw high loadings from the connection rating (0.89) and the coping rating (0.67) with moderate loadings from the cheerfulness rating (0.52) and the wellbeing rating (0.52) it suggested the name "*Coping and social connection*". The components will be referred to by these names for the remainder of the chapter.

Gstudy

For the last step, we ran a *Person x Activity x Rating (PxAxR)* Gstudy, where *P* represented individual variance and our facet of generalization, *A* was the 8 unique activities, and *R* was the 6 different rating parameters (excluding effort to avoid excessive contributions to variance). The same was done where PCA scores replaced the rating component. In each case we calcu-

lated the coefficients relating to the generalizability of each facet, as well as the absolute and relative G coefficients. The variance profiles can be seen in figure 5.8. We found that scores generalized well across ratings (ratings coefficient = 0.79) and moderately well across activities (activity coefficient = 0.59). Both the relative G coefficient (0.32) and absolute G coefficient (0.31) were relatively low, and notably their closeness indicates that most variance stems from individual differences interacting with the included facets, rather than the main effects of the facets themselves.

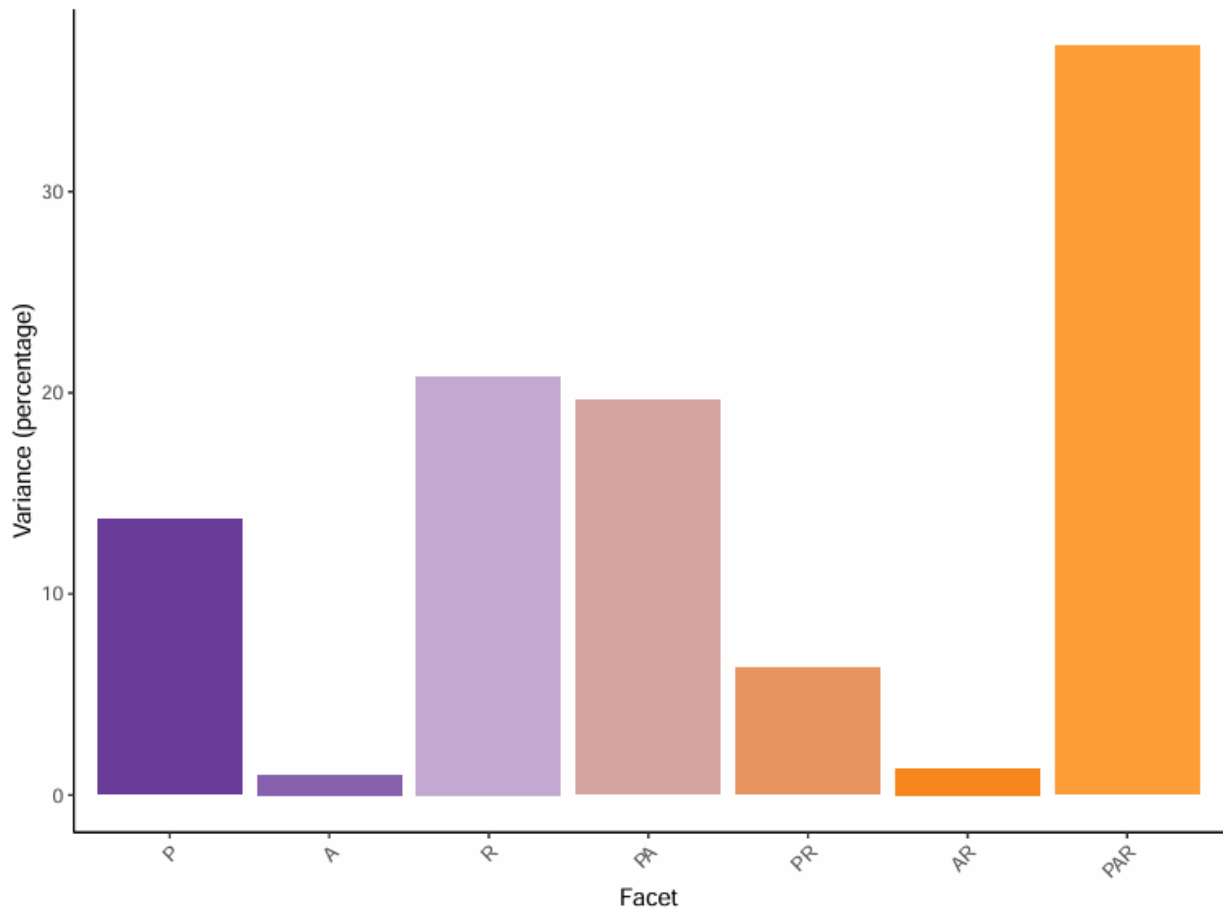


Figure 5.9: Gstudy variance profiles

5.3.3 Analysis of post-activity ratings and their potential predictors.

Individual activity predictors

First, we ran 4 models looking at the individual activities, with raw effort ratings and each of the 3 component scores as our dependent variables. Our independent variables were the 8 different activities with 5 minutes of focus as a baseline, the activity number as a continuous variable (ie. whether it was a participant's first, third or fifth completion), and for the component score models, effort ratings were included as a predictor. Results for all of these models can be found in table 5.2.

Table 5.2: Regression results for predictors of post-activity ratings

a) Predictors of effort ratings (adjusted $R^2 = 0.044$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Body reset	-0.354	0.347	-1.018	0.309
Letting thoughts come and go	1.438	0.398	3.614	< .001
Medium unguided deep belly breathing	0.444	0.399	1.113	0.266
Mindful breathing	0.407	0.385	1.059	0.290
Progressive muscle relaxation	1.210	0.439	2.757	0.006
Rise and shine	0.390	0.363	1.074	0.283
Short unguided deep belly breathing	-0.409	0.322	-1.272	0.204
Activity number	-0.021	0.103	-0.201	0.841

b) Predictors of *Positive emotional experience and satisfaction* component (adjusted $R^2 = 0.068$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Body reset	0.083	0.134	0.619	0.536
Letting thoughts come and go	0.128	0.155	0.823	0.410
Medium unguided deep belly breathing	-0.250	0.154	-1.622	0.105
Mindful breathing	0.065	0.149	0.435	0.664
Progressive muscle relaxation	0.141	0.170	0.829	0.408
Rise and shine	0.326	0.141	2.317	0.021
Short unguided deep belly breathing	-0.282	0.125	-2.263	0.024
Activity number	0.091	0.040	2.295	0.022
Effort rating	-0.195	0.038	-5.141	< .001

c) Predictors of *Stress level* component (adjusted $R^2 = 0.022$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Body reset	-0.087	0.138	-0.631	0.528
Letting thoughts come and go	0.142	0.159	0.890	0.374
Medium unguided deep belly breathing	-0.010	0.158	-0.062	0.951
Mindful breathing	0.086	0.152	0.563	0.574
Progressive muscle relaxation	0.061	0.175	0.349	0.728
Rise and shine	-0.059	0.144	-0.410	0.682
Short unguided deep belly breathing	-0.018	0.128	-0.142	0.887
Activity number	-0.030	0.041	-0.739	0.460
Effort rating	-0.182	0.039	-4.666	< .001

d) Predictors of *Coping and social connection* component (adjusted $R^2 = 0.034$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Body reset	0.058	0.137	0.427	0.669
Letting thoughts come and go	0.045	0.158	0.287	0.774
Medium unguided deep belly breathing	-0.390	0.157	-2.482	0.013
Mindful breathing	-0.084	0.151	-0.555	0.579
Progressive muscle relaxation	-0.104	0.174	-0.599	0.550
Rise and shine	0.051	0.143	0.357	0.721
Short unguided deep belly breathing	-0.092	0.127	-0.725	0.469
Activity number	0.162	0.040	4.008	< .001
Effort rating	0.081	0.039	2.107	0.036

Notably in these models, activity number was positively significant for the *Positive emotional experience and satisfaction* component model and the *Coping and social connection* component model but not the Effort model or the *Stress level* component model. The effort rating was significant in all three component models, positively significant for the *Positive emotional experience and satisfaction* component model and the *Coping and social connection* component model but negatively for the *Stress level* component model. Furthermore, for each of these models, we also ran a separate model excluding the 8 activities as predictors, and compared them via an omnibus test to see if they were significantly different. Including activities significantly improved the Effort model ($p < .001$) and the *Positive emotional experience and satisfaction* component model ($p < .001$), but not the *Stress level* component model ($p = 0.876$) or the *Coping and social connection* component model ($p = 0.130$).

Predictors of mean activity ratings

For this part of the analysis user mean scores were calculated for their effort ratings and each of the principal rating components, leaving the same 4 dependent variables as used in the previous section, yet only one set of scores per user. 13 independent variables were included for the effort model, and 14 for the three component models, as mean effort ratings were included as independent variables. Due to this large number of variables, to avoid over-fitting the models, we utilized the model selection method described in Chapter 3. Results from all four final models can be seen in table 5.3. Original models including all variables can be found in the appendix. The independent variables included were as follows:

1. Whether participants scored above or below the 41-point threshold on the WEMWBS-14 at onboarding, with scores below indicating probable clinical depression.
2. Participant PSS-10 scores at onboarding grouped by less than 14 (indicating low stress), between 14 and 26 (indicating moderate stress) or above 26 (indicating high stress)

3. Honesty subscale score of the Hexaco-60
4. Emotionality subscale score of the Hexaco-60
5. Extraversion subscale score of the Hexaco-60
6. Agreeableness subscale score of the Hexaco-60
7. Conscientiousness subscale score of the Hexaco-60
8. Openness subscale score of the Hexaco-60
9. Participant age
10. Participant gender
11. Whether participants were in the bottom third, middle third or top third of the education principal component scores
12. Whether participants were in the bottom third, middle third or top third of the *Income and status* principal component scores
13. Whether participants had completed only the minimum required activities (4) or more.
14. Effort scores (except for the model predicting effort scores)

Table 5.3: Regression results for final models predicting mean post-activity ratings

(a) Final predictors of mean effort ratings (adjusted $R^2 = 0.038$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Emotionality	0.223	0.190	1.171	0.244
Extraversion	0.240	0.200	1.203	0.232
Agreeableness	0.201	0.194	1.037	0.302
Openness	-0.253	0.192	-1.322	0.189
Middle <i>income and status</i> bracket	-0.484	0.458	-1.058	0.293
Upper <i>income and status</i> bracket	-1.104	0.479	-2.306	0.023

(b) Final predictors of mean *Positive emotional experience and satisfaction* component (adjusted $R^2 = 0.290$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Moderate stress	0.145	0.102	1.431	0.155
Honesty	-0.270	0.060	-4.470	< .001
Agreeableness	0.186	0.055	3.345	0.001
Conscientiousness	0.239	0.057	4.201	< .001
Openness	-0.062	0.053	-1.170	0.245
Activity bucket	0.326	0.111	2.946	0.004
Middle <i>income and status</i> bracket	0.147	0.105	1.400	0.164
Mean effort rating	-0.131	0.050	-2.641	0.010

(c) Final predictors of mean *Stress level* component (adjusted $R^2 = 0.022$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Honesty	-0.165	0.064	-2.575	0.011
Extraversion	-0.058	0.057	-1.021	0.310
Agreeableness	0.226	0.058	3.928	< .001
Conscientiousness	0.084	0.059	1.428	0.156
Activity bucket	0.157	0.115	1.360	0.177
Age	-0.060	0.055	-1.104	0.272
Upper <i>income and status</i> bracket	-0.119	0.114	-1.039	0.301
Mean effort rating	-0.120	0.052	-2.313	0.023

(d) Final predictors of mean *Coping and social connection* component (adjusted $R^2 = 0.083$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Honesty	-0.179	0.079	-2.263	0.026
Emotionality	-0.092	0.068	-1.346	0.181
Agreeableness	0.165	0.074	2.247	0.027
Conscientiousness	0.104	0.074	1.409	0.162
Activity bucket	0.224	0.146	1.529	0.129
Mean effort rating	0.096	0.066	1.453	0.149

5.3.4 Analysis of offboarding data.

Time as a predictor of change

Time did not significantly predict change in either WEMWBS-14 score or PSS-10 scores.

Predictors of engagement

Here we ran a generalized linear model (family = "binomial") with number of activity completions as a binomial dependent variable (whether users did only the minimum of 4 activities, or above that). The same independent variables as in the previous section (excepting activity bucket) were used, along with the mean component scores for the post-activity ratings, totaling 16. The same model selection process was used. Results from the final model can be seen in table 5.4, and the full original model including all variables can be found in the appendix.

Table 5.4: Final model predicting activity completions

Independent variable	Coefficient	Std. Error	z value	p value
High mental wellbeing	0.957	0.492	1.946	0.052
Mean <i>Positive emotional experience and satisfaction</i>	0.657	0.241	2.723	0.006
Age	0.516	0.253	2.037	0.042

Predictors of change

Here we ran two models, one with WEMWBS-14 and one with PSS-10 as dependent variables. The same independent variables as in the previous section (including activity bucket) were used, along with the mean component scores for the post-activity ratings, totaling 17. The same model selection process was used and results can be seen for the final model in table 5.5. Results from the full original model can be seen in the appendix.

Table 5.5: Regression results for final models predicting change in mental health

(a) Final predictors of change in WEMWBS-14 (adjusted $R^2 = 0.226$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
High mental wellbeing	-6.335	1.288	-4.919	< .001
Mean effort rating	0.575	0.574	1.001	0.319
Mean <i>Positive emotional experience and satisfaction</i>	1.113	0.593	1.877	0.063
Agreeableness	0.732	0.596	1.229	0.222
Conscientiousness	0.835	0.612	1.364	0.175
Openness	0.910	0.594	1.530	0.129
Age	-0.919	0.563	-1.634	0.105
Upper <i>education</i> bracket	-1.556	1.160	-1.341	0.183

(b) Final predictors of change in PSS-10 (adjusted $R^2 = 0.161$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Moderate stress	-2.611	1.060	-2.463	0.015
High stress	-6.496	2.034	-3.194	0.002
Mean effort rating	-0.630	0.466	-1.352	0.179
Mean <i>Positive emotional experience and satisfaction</i>	-1.043	0.486	-2.146	0.034
Extraversion	0.705	0.521	1.354	0.179
Agreeableness	-0.755	0.477	-1.581	0.117
Conscientiousness	-0.593	0.503	-1.180	0.241

5.3.5 Analysis of follow-up data

Time did not significantly predict change in either WEMWBS-14 score or PSS-10 scores between the first and the final assessment. However we also ran models investigating whether change in scores between onboarding and exit for predicted change between exit and follow-up. Initial change was a significant negative predictor for both WEMWBS-14 ($beta = -0.279$, $p < .001$) and for PSS-10 ($beta = -0.303$, $p < .001$)

5.4 Discussion

5.4.1 The immediate effect of mental health activities

The current study showed some potential for utilizing single ratings to determine the perceived effect of activities on different user states. Through the mood and mental health related ratings, we could see that the activities were, for the most part, having their intended effect. The majority of participants either felt no change or positive change in their cheerfulness, stress relief, mental

wellbeing, ability to cope and connection to self and others, a finding which is mirrored in similar research (Lewis et al., 2022). Only a few users perceived a negative effect of the activities. The post-activity ratings also performed more to their intended effect than those described in Chapter 3, as items were more distinct from each other. However, with the exception of effort ratings, they still seemed to tap into the same underlying construct. This was partially expected, as there is a strong link between stress and wellbeing (Coffey, 2004; Teh et al., 2015), and the other items were chosen for their relevance to those two constructs, yet the extent to which participants agreed on average ratings was still surprising.

The relationships between the items was also interesting. We found that stress loaded almost entirely onto its own component, whilst the other ratings overlapped more closely on two different components. This is notable for the satisfaction rating especially, as this was intended to capture information regarding the quality of the activity itself, somewhat similarly to the effort rating. However, it would make sense that these are related if participants are utilizing the app to its intended purpose. If they hope to gain mental health benefits, their momentary satisfaction is quite possibly linked to the perception that an activity performs the function they expect. This has been seen in the longer term where intervention credibility and satisfaction predicts treatment outcomes (Constantino, Coyne, Boswell, Iles, & Vísľá, 2018).

It is also possible that the correlation between variables is explained by situational consistency, or self-report behavioral patterns. Both our ICC calculations and Gstudy showed that individual differences accounted for a large part of the variance in scores, however the Gstudy also showed that generalizability across ratings was quite high, more so than across activities. This may be because there are unmeasured situational factors which contribute to how users feel at any given moment, something which has previously been found to be influential upon people's states (da Fonseca et al., 2023; Lebois et al., 2016). The post-activity ratings were, for the most part, answered within seconds or minutes of each other, where the activities might be completed hours or days apart. The participants' circumstances, their surroundings, the events of the day might all contribute to what they report, and it is possible that scores are less generalizable across activities because they incorporate some degree of temporal or situational variation. In contrast, the post-activity ratings, even when theoretically distinct, are still measured at one point in time, perhaps capturing a snapshot of their overall state and mood. We had hoped to counteract this to an extent by anchoring their states to the perceived change from the activity. Participants were not intended to answer how they felt in that moment, so much as how they perceived the activity's influence on their state, but it is still possible that ratings were influenced by situational factors.

This would also go some of the way in explaining our low overall generalizability of scores (which incorporate the activity and ratings variance). This is not inherently bad for this type of measure, although it is challenging. The ratings were intended to capture the perceived effect and

quality of the activities, but they still concerned themselves with user states, moods and feelings in a single moment, and these are volatile variables. There are several other methods to capture this information, such as implicit user data or wearables, and in each case there are notable fluctuations (Malhi et al., 2017). These fluctuations are human nature, and variable results are to be expected in any attempt to capture an individual's state at any given time. Yet despite the volatile nature of this data it is useful in both research and practice, and several researchers have attempted to leverage it to generate recommendations for long term behavior change and mental health improvement (Beltzer et al., 2022; Lewis et al., 2022; Rohani et al., 2021). By studying the relationships between states and traits and short-term and long-term improvement we may be able to develop more reliable measures for use in more targeted algorithms that are more tailored to their purpose.

5.4.2 Activity effects on short-term improvement

The current study investigated what might influence participant perception of the activities and their effect. Most trial studies of mental health apps evaluate benefits in the longer run and consider only gains in more stable mental health traits and measures (Chandrashekar, 2018; Donker et al., 2013; Khademian et al., 2021; Lecomte et al., 2020; Neary & Schueller, 2018; Wang et al., 2018). This is important, and should be a goal for mental health apps. However, we would argue that if an app offers mental health solutions which provide immediate improvement, or even perceived improvement, this is valuable. The current study therefore examined what predicted these momentary benefits of engaging with mental health apps. First we explored whether the qualities of the individual activities or the context were useful determinants of participant ratings. We found that including the activities themselves improved the models for effort and *Positive emotional experience and satisfaction*, but not for *Stress level* or *Coping and social connection*. This highlights a minor issue in the current methodology that should be addressed in future research. In both cases where the activities contributed significantly to the model, these were outcomes which related, at least in part, to participant perception of the activities, rather than their perception of their own mental health states. We attempted to create ratings that distinctly captured either information regarding mental health related states or subjective engagement, but as previously described there was a large overlap between the two in our data. Therefore, it is hard to tell purely from our results whether the activities influenced participant improvement or whether some were simply perceived as more engaging than others. It is also possible that it is both, that activity satisfaction and state mental health improvement go hand in hand. If so, that would be concurrent with the notion presented by Ruiz de Villa et al. (2023), that there is a positive self-reinforcing cycle between engagement and improvement, but more research is needed to untangle these variables and understand their relationship with each other.

We also investigated whether effort ratings influenced the three principal components, *Positive*

emotional experience and satisfaction, *Stress level* and *Coping and social connection*. Effort rating significantly predicted component scores in each case, however it was negatively associated with *Positive emotional experience and satisfaction* as well as *Stress level*, but was positively associated with *Coping and social connection*. In Chapter 3 we discussed how more effortful tasks can increase the perception of self-efficacy (Locke & Latham, 2006; Rosenstock et al., 1988), and we suspect it is this mechanic that also taps into coping efficacy in this study. The more difficult a task the more a participant may feel equipped to tackle further problems they face. However, as in Chapter 3, there are some indications that this may be a fine line to walk, as in this study higher effort was linked to negative outcomes in the other components. This is consistent with other research which has indicated that increased cognitive effort can be a stressor (Ehrhardt, Fietz, Kopf-Beck, Kappelmann, & Brem, 2022). We therefore hypothesize that mental health activities that challenge users can be beneficial up to a certain point, but go beyond that point and it can become stressful, which in turn is likely to reduce their satisfaction, mental wellbeing and cheerfulness. These nuances are the reason it is important to consider the underlying mechanisms of mental health change, as improving our understanding of them may be critical to optimizing our design of digital therapeutic tools. Human psychology can at times be a fragile thing, and not giving it due consideration may be one of the reasons why several mental health apps work opposite to their intended effect (Anthes, 2016; Torous, 2018). It must also be said, however, that this study only scratches the surface, and more research is necessary, especially given that there were discrepancies between the findings in this study and that described in Chapter 3. As opposed to the previous study, in the current one mean effort ratings predicted an improvement (although non-significantly) in long term mental health gains. Furthermore, the models themselves only explained a small part of the total variance of the scores, and therefore likely only explain a small part of a larger picture. Replicating and refining these findings will be necessary for research in the field to catch up to technology and application.

5.4.3 Predictors of mean activity ratings

The models for predicting mean component scores were relatively good, with the exception of the effort model, which was perhaps to be expected. Our findings indicated that effort ratings varied more, and depended more on the activities, and it would therefore make sense that traits are less useful, especially for predicting mean values where much of the individual information is lost. However, for the three principal components, models performed better, especially component one, *Positive emotional experience and satisfaction*. This could be a useful tool for user modeling, supplementing, for example, the development of user archetypes such as described by Aziz et al. (2023). Their focus was more on how users engaged with mental health apps, whereas our models may inform which individuals are more likely to experience temporary mood boosts, symptom relief or general satisfaction. In fact, our findings highlighted that these might be linked. The number of activities participants did was included in all three final com-

ponent models as a positive predictor, although it was only significant for *Positive emotional experience and satisfaction*. If we were to consider aggregate measures of short-term change as indicative of long-term change, this would also support the premise of some recommender system studies showing that consistently improving mood leads to long term mental health gains (Lewis et al., 2022; Rohani et al., 2021). This would require further testing as this type of assessment is not an established measure of mental health. However there are at least some studies indicating the potential for behavioral and mental health scores to be determined based on the means of single items across a variety of situations (Dutriaux, Clark, Papiés, Scheepers, & Barsalou, 2023; Taylor Browne Lūka, Hendry, Dutriaux, Stevenson, & Barsalou, 2024). The current study also indicated the potential usefulness of composite momentary assessments in assessing mental health, as *Positive emotional experience and satisfaction* scores, although non-significant, positively predicted improvement in both WEMWBS-14 scores and PSS-10 scores.

Some personality traits were also consistent in predicting mean activity scores. Agreeableness, honesty and conscientiousness were present in all three final component models, although conscientiousness was only significant for *Positive emotional experience and satisfaction*. In all cases, higher agreeableness and conscientiousness scores predicted an increase in activity ratings, whereas high honesty scores predicted a decrease. The conscientiousness finding is supported by existing literature, finding that conscientious users tend to prefer and engage more with audio content (Khwaja et al., 2021), which was the only activity format included in the current study. However, the consistency of lower honesty and higher agreeableness predicting higher ratings may indicate a confounding behavioral pattern. Validity issues in self-report data are well known in social sciences (Hussey & Hughes, 2020), and current findings may indicate that some participants are rating activities highly out of kindness rather than true subjective opinions. If this finding can be replicated in future research, it suggests that care should be taken with recommendations based on this type of measure, and perhaps supplemented by implicit data, such as is common in many recommender systems (Madadipouya & Sivananthan Chelliah, 2017).

5.4.4 Long term change

The first thing to note regarding stress and mental wellbeing outcomes was there was no significant difference over time. Trends were positive for both the WEMWBS-14 and the PSS-10, however not significantly so. This contrasts with most findings using Foundations (Catuara-Solarz et al., 2022; Gnanapragasam et al., 2023; Schulze et al., 2024), however there are several possible explanations for this. First, there is precedence for the timeframe making a difference. Participants in the current study only had access to the app for two weeks, which was not enough to induce significant change in the study by Schulze et al. (2024), although four weeks was. In

Chapter 4 we saw that in the wild, there was no significant difference between change in scores at different intervals, so there may be some facet of the controlled settings that means more time is needed to induce mental health change. It could also be the choice of measure, as mental wellbeing in all other studies investigating Foundations evaluated outcomes on the WHO-5 as opposed to the WEMWBS-14. Or, it could be the choice of activities, as we included only a limited sample, and excluded sleep related activities, which were found to be the most effective in improving mental health across multiple constructs by Ruiz de Villa et al. (2023). It could also be that our sample did not have any mental health exclusion criteria as our study in Chapter 3 did, which may be especially likely as we have now consistently seen that change is more pronounced in samples with lower levels of mental health. Again this denotes the importance of considering generalizability of results and the multitude of individual and group level factors that may influence whether or not a mental health app is effective in enabling positive change for its users.

However, the non-significant change only shows that change was not significant at the group level in this study. There was a considerable spread in score changes, with many participants showing mental health progress (and several showing deterioration). This suggests that whilst mental health apps may not always help everyone, there are individuals who benefit from them. This should not come as a surprise. As outlined in Chapter 1, even face-to-face therapies, which have been tried and tested for decades in some cases, are not always effective. We cannot reach everyone. Innovation and refinement of solutions are some ways in which we can attempt to cast a wider net and reach a wider demographic. Personalization is another. If we can understand when and why our solutions work, what character or situational traits drive positive change, then we have the technology to put that knowledge to good use. This was the intent behind our exploration of predictors of change. One of the main findings, which has now been seen in all three studies across Chapter 3, 4 and 5, was that the most effective and consistently significant predictor of mental health improvement in a construct was original mental health levels in the same construct. Low mental wellbeing at the offset predicted positive change in mental wellbeing, and moderate or low stress levels increasingly predicted positive change in stress. This is now consistent evidence that mental health apps are most likely to help those who need it more.

Other variables were notable for their absence, or non-significance. Despite the strong association between different personality traits and mental health seen in the literature (Liu et al., 2021; Luo et al., 2023; Schneider et al., 2024; Sirgy, 2021; Zotova & Karapetyan, 2018), no personality traits were significant in predicting change in the current study. Whilst scores in agreeableness and conscientiousness predicted improvement in both WEMWBS-14 and PSS-10 scores, and the directionality is consistent with the literature, they were not significant. This contrasts with findings in similar research, where the association for all personality traits is stronger (Bucher,

Suzuki, & Samuel, 2019). Of course, these are typically studied in clinical trials, which the current study was not, and the relationships are better researched in traditional therapy types compared to mental health apps. More targeted research may be able to uncover whether there are different associations between settings. It could also be a matter of effect sizes. Differences between scores across timepoints were not significant in this study, and it may be that a larger sample size (or perhaps a sample only from lower mental health brackets) could identify relationships that were hidden here.

Effort was also notable as in both cases higher mean effort scores were positively associated with improvement in mental health outcomes, although in both cases not significantly. This is consistent with other findings (van Agteren et al., 2021), but again, must be taken with a grain of salt. Effort in the current study appeared to be tied to individual activities, and therefore the generalizability of results stemming from aggregated scores is questionable. Despite this, it is worth noting that it has been a consistently influential variable on different mental health outcomes in both this chapter and in the study in Chapter 3. Directionality differs depending on context, and it is not consistently significant, but whether on an individual activity level or as a mean measure, it appears to have a prominent role in mental health change in mental health apps. Targeted research on the specific mechanisms of its interaction with outcomes across time could prove valuable for the field in general.

Perhaps our most critical findings for the field were how the post-activity components influenced change in mental health. We found that in both the WEMWBS-14 and the PSS-10 models, higher mean *Positive emotional experience and satisfaction* predicted an increase in long-term mental health outcomes, although only significantly for PSS-10. This provides some support for algorithms such as those implemented by Lewis et al. (2022) and Rohani et al. (2021), which focus on recommending activities that are most likely to improve user moods. The premise in both these studies was that this would lead to long-term mental health change, but neither directly tested this. However, this was also the only useful component in the present study. Neither *Coping and social connection* nor the *Stress level* components made it into the final models, suggesting that they had very little influence on mental health outcomes. This suggests that some form of distinction is required between state-level assessments. The current study indicated that it can be done, that pursuing long-term mental health goals is feasible through enacting short-term improvement, but that it depends on the short-term construct or measure. It is perhaps especially telling that state-level stress relief did not predict long-term change in stress levels, strongly suggesting that researchers have to be careful in making common-sense assumptions without explicit testing. Understanding this in interdisciplinary research will be critical to avoid the same replication and generalizability crisis currently experienced across social science fields.

Finally, it also merits highlighting that for both WEMWBS-14 and PSS-10, change from on-

boarding to exit scores was significantly negatively associated with change from exit to follow-up. This suggests that when participants experienced a change in mental health whilst they had access to the app, the change was likely to revert after losing access. This was intended to be an early exploration of the retentive benefits of mental health apps. There are several studies on long term effects of different treatments, and several have found enduring benefits for mental wellbeing and stress (Knaevelsrud & Maercker, 2010; Solhaug et al., 2019), but very little research exists on the persistence of improvement from mental health apps. Our results are an early indication that mental wellbeing and stress may stabilize back towards original levels if app use is not consistent. This, however, requires not only replicating, but further research with different intervals. Here and in Schulze et al. (2024), we found that two weeks was not enough time to significantly enact mental health change through Foundations, and it is possible that longer access may lead to more lasting effects. However, even so, it would be important to understand how long, and also whether the duration benefits are retained. Focusing on this, as well as the other temporal and engagement aspects we have explored in this chapter, can help developers, researchers and users understand how best to use mental health apps. Perhaps they can be utilized as treatments such as CBT often are, where people are given tools to take charge of their own mental health even after they no longer receive active counseling. Or, it may be they are most useful as a form of symptom relief, to be picked up on bad days to improve moods, with less focus on long-term effects. Guidelines could be established for daily use, weekly use or monthly use depending on needs and estimated benefits. The research in this chapter is the tip of the iceberg, indicating potential directions future, more specific research could take in understanding mechanisms of mental health change to improve mental health apps.

5.5 Proof of concept recommender system

5.5.1 Rationale and methods

Beyond theoretical understanding of mental health change through app usage, this research was intended to be a stepping stone for practical applications. Specifically, it was meant to provide guidance to future research and development focused on recommendations for mental health outcomes, such as the studies described in Chapter 2. As discussed, there is still a long way to go in understanding mechanisms of change, however the data and findings described here were still enough for preliminary investigations of usefulness in the real world. We therefore evaluated the performance of an offline ratings based recommender system, to investigate the feasibility of recommending activities that were most likely to improve mental health outcomes. To this end we tested a user-based and an item-based collaborative filter based on the standardized component scores for *Positive emotional experience and satisfaction*. We did this in a similar manner to that described by Lewis et al. (2022). Compared to the recommender system described by

Lewis et al. (2022), ours was built with far less data, and requires far more for even robust offline evaluation, however in our case there was an established relationship between the momentary ratings and long-term outcomes for both mental wellbeing and stress within the sample. This is also why we only chose this component, as the other two had no relationship with long-term change.

The recommender system was based on the 685 activity ratings from 113 users. The mean-centered and standardized component scores, ranging from -2.823 to 3.975, a slightly larger scale than the 5-point scales used in similar studies (Lewis et al., 2022; Rohani et al., 2021). We designed it as a matrix completion problem, investigating how well an algorithm could predict missing user ratings. We assessed two personalized models and two baseline models.

1. **Global means baseline:** random samples were drawn from a normal distribution defined by the global means and standard deviations of user ratings.
2. **User average baseline:** user patterns were taken into account as ratings were predicted based on individual means.
3. **User-based k-nearest neighbour:** the algorithm estimates missing ratings by identifying k users with similar preferences and computing their aggregate scores for the activity in question.
4. **Item-based k-nearest neighbour:** the algorithm estimates missing ratings by identifying the k items with highest similarity and computing aggregate scores based on the user's ratings for those similar items.

The *global means baseline* estimates missing ratings based on the overall average ratings of the items. This is based on an expectation of higher uniformity across both users and activities for ratings, as individual differences are not taken into account. The *user average baseline* on the other hand, takes into account these individual differences by predicting ratings based on individual user means. If a user rates activities higher or lower in general compared to the global average, this is thereby represented in the predicted ratings. These models were the same as those described by Lewis et al. (2022).

The personalized models were both k-nearest neighbour (KNN) collaborative filters. These collaborative filters work by grouping either users or items together by similarity, and predicting missing ratings based on aggregate scores. The *user-based KNN* algorithm identifies users with similar rating patterns, based on the assumption that individuals with similar preferences will rate items similarly. The *item-based KNN* identifies item similarities and assumes that similar items will be rated similarly by individual users.

In both cases this was done by computing similarity, finding the nearest neighbours and then

predicting the missing ratings. As there are different methods for doing this, we assessed the optimal choices by tuning our hyperparameters. For each model we evaluated whether they performed better when similarity was computed using the cosine similarity metric or the pearson similarity metric. We determined the optimal number of k nearest neighbours by evaluating performance with k ranging from 3 to 30 in increments of 3.

As in Lewis et al. (2022), we reported Root Mean Square Error (RMSE), Mean Squared Error (MSE) and Mean Absolute Error (MAE), and based our hyperparameter selection on the best RMSE values. For both models we ran a 10-fold cross validation across 20 iterations, and the reported values were the average error values across these 20 iterations. The RMSE and MSE values penalize larger errors to a greater extent, and were chosen to minimize larger prediction mistakes. However, as a result of penalizing larger mistakes, they are both sensitive to outliers, we included MAE values as well, which is based on absolute values and therefore does not share this drawback.

5.5.2 Results and discussion

Table 5.6 shows the evaluation metrics described in the previous section for the different models. The user average baseline performed poorer than the other models, and the user-based collaborative filter performed best across the board. However, differences were marginal, and the global means baseline performed better than the item-based collaborative filter on RMSE and MSE metrics. This is perhaps no great surprise, as the dataset from which the algorithms were built was very small for

Table 5.6: Results of algorithm evaluations

	RMSE	MSE	MAE
Global means baseline	1.060	1.149	0.854
User average baseline	1.330	1.786	1.072
User-based collaborative filter	1.028	1.073	0.830
Item-based collaborative filter	1.086	1.224	0.835

the purpose, and recommender systems can be data hungry. For comparison Rohani et al. (2021) built their recommender system from two datasets, consisting of 1684 entries from 7 patients, and 6344 entries from 134 users. Lewis et al. (2022) had 23,476 ratings from 973 users. In each case, the algorithmic comparisons were the primary focus of the study. In contrast, our focus was more macro-methodological. The algorithms we used were simple and based on too little data to confidently promote it as a useful tool. However, it did perform acceptably, and in our case, was based on a direct relationship with long-term mental wellbeing and stress. Lewis et al. (2022) and Rohani et al. (2021) justified their recommendation inputs and algorithms based on theoretical psychology, and developed algorithms that performed their given task well, but their

influence on long-term mental health gains remains theoretical. However, we would argue that given the state of the research on mental health apps and their recommender systems, this may not be enough. The current study indicated that not all measures of momentary improvement necessarily translate to long-term benefits, and beyond that, findings on the beneficial nature of mental health apps are too inconsistent to be taken for granted in any study (Anthes, 2016; Torous, 2018).

This brings us full-circle to the some of the issues described in the beginning of the thesis. There exists excellent research on mental health recommender systems, and there exists excellent research on the psychological benefits of mental health apps. However, the interdisciplinary fields have not fully merged yet. This was why the current study included a simple proof-of-concept evaluation of recommender systems, despite the limitations of the dataset. Not to sell the idea of a novel algorithm, but to demonstrate one way parallel disciplines can be brought together just a little more. Instead of following the tendency described in Chapter 2 where researchers evaluate algorithms and mental health outcomes in separate studies (or evaluating only one outcome), it is important, at least in this early stage of research, to provide more than a theoretical link between the two. It is possible that eventually the supportive research of mental health apps will be robust enough that researchers can more safely assume benefits or generalizability of results, but that is not currently the case. Studies continue to evaluate new recommender system algorithms rather than replicating existing ones (Portugal et al., 2018) and generalizability within the field suffers (Dacrema et al., 2019). This is compounded with the fact that so many different mental health apps flood the market (Neary & Schueller, 2018) and the field of psychology faces its own replication and generalizability crisis (Oberauer & Lewandowsky, 2019; Yarkoni, 2022). This means, that no matter the quality of an individual study of a recommender system, if it does not include a measure of mental health outcomes, its ability to claim that the algorithm can bring mental health benefits will be tenuous at best.

The methodology of the current study attempted to address this, to merge fields more thoroughly, but this was just the first step. As *Positive emotional experience and satisfaction* was linked to improvements in stress and mental wellbeing within the dataset the algorithms were built from, recommending items with higher predicted ratings should theoretically have a long-term effect. This still requires further testing however, and it is important that these algorithms be tested in live settings. Offline studies are a useful tool for evaluating algorithmic potential, but their usefulness for improving mental health outcomes need to be much better established before we can trust the generalizability of results. Future research should focus on evaluating algorithms in RCT's or published A/B tests, reporting differences in mental health change. Ideally it would also evaluate engagement metrics, both for multi-stakeholder value, but also to investigate how traditional engagement-focused recommender systems compare against recommender systems with a mental health focus on different outcomes.

5.5.3 Limitations

In some ways this study was broad in scope and shallow in detail. That is often the case with exploratory research, but it should still be noted that we cast a wide net and allowed participants freedom in their interactions with the app. These results should be taken as a starting point for future research, especially given that mental health scores did not differ significantly over time. Either a larger sample or a more targeted sample would likely provide more robust results and findings.

Some limitations were also carried over from the study in Chapter 3. On the positive side, the app had been optimized for the current research, and only a few individual participants experienced any lag or technical issues. However, it was still designed primarily as a web app, and although possible to access via mobile, information on medium was not collected. The post-activity ratings had also been refined, and performed more to their intended effect, but we still believe further development and testing would be useful, especially for untangling participant perception of app performance from mental health benefits.

The recommender systems we evaluated were also heavily limited by the dataset as already mentioned. The amount of data compared to similar studies was far smaller, and results should not be taken as more than a proof of concept and food for discussion. The models themselves, if scaled up to larger datasets could also be improved. The baselines, whilst not uncommon in the literature, could be stronger. In fact, given issues with poor baselines in the field leading to inflated results (Rendle, Zhang, & Koren, 2019), we would strongly recommend more focus be devoted to better comparative algorithms with tuned hyperparameters.

5.5.4 Conclusion

The current study investigated what influenced participant mental health change over time, and what the link was between the short-term benefits of activities and long-term improvement. We found that short-term outcomes differed, and may only influence long-term gains depending on the measure, with the most significant being a combination of *Positive emotional experience and satisfaction* with the activities. Once again mental health levels at the outset of the study were strongest and most consistently significant predictors of change, with the participants in the lower brackets of mental wellbeing and stress being those with most to gain. The perceived effort expenditure related to activity completions were also consistently informative, however further research is needed to identify their precise role. Finally, we also demonstrated that our ratings of momentary improvement could be used in algorithms designed to recommend activities that were most likely to influence mental health outcomes, and discussed how future studies could improve methodologically by combining algorithmic and mental health evaluations.

Chapter 6

General discussion

6.1 Summary of previous chapters

It is difficult to maintain an overview when moving through any work of this magnitude. Perhaps especially so in this thesis, as it was not a narrow, streamlined focus, culminating in an opus magnum theoretical framework, or groundbreaking new product. Instead, it moved through several foundational concepts, and considered, on a broader scale, how research can be improved in the field of mental health apps, and the recommender systems that sometimes accompany them. This chapter will bring together the separate threads, explore and advise how researchers might work towards a more unified field. It will first, however, begin by providing a brief summary of the previous chapters, to ensure that the work that was undertaken sits comfortably in the reader's mind, so it can more easily be related to the broader context.

The first chapter set this context. In a broad exploration of multidisciplinary literature, it aimed to clearly define the setting within which all following work took place. The mental health setting. Later we will explore how the field of mental health apps could benefit from more clarity of purpose, and this clarity of purpose comes from acknowledging that mental health apps are primarily designed to improve mental health. I described how they are, in theory, a perfect fit for a current problem, that it is very difficult to provide the support people need. There are many reasons for this. Mental health is an extremely broad topic, and issues are diverse, and do not necessarily overlap. Depression functions differently from anxiety, stress differently from insomnia, and despite some commonalities, unique solutions are required. And whilst many such solutions exist, we saw that they never guaranteed effectiveness, and furthermore that there exist numerous barriers to accessing them in the first place, some structural, some personal, but all reducing our ability to help those who need it. This was where we began our exploration of mental health apps, a ubiquitous solution that is readily, often cheaply or even freely available. Even better, with the existence of recommender systems, the presentation of solutions could

be personalized, allowing for individual support through a widely available medium. However, perhaps because they fit the problem so well, they began to flood the market faster than research could support them. Research highlighted a lack of proper testing of the efficacy of mental health apps, numerous methodological issues in the field, and a sparsity of studies evaluating mechanisms of change. These issues were the ones we explored throughout the rest of the thesis. We examined how studies presented recommender systems, whether they were designed with engagement or mental health outcomes in mind. We evaluated how these variables fit together, and how they might be meaningfully defined, and how generalizability and ecological validity could be improved to allow for more practical research. And we considered which operational definitions, which potential constructs were most useful to include as variables in research or input in algorithms to pursue an understanding of why mental health apps can work, and how we can optimize them to better solve targeted problems.

The first step in achieving this was reviewing existing literature. Chapter 2 consisted of a systematic review designed to understand how recommender systems are applied to mental health settings. This was where we began to firmly establish the joint evaluation of engagement and mental health outcomes, as recommender systems were originally designed for a context where engagement was the primary aim. Through the analysis of ten studies which used recommender systems to personalize mental health solutions, we got an insight into how scattered the field was. Existing research was quite clearly in the early stages, characterized by offline evaluations and novel ideas, or initial testing of a new mental health solution. The quality of research was high individually, but some problems in the field as a whole became apparent. There was a clear separation of disciplines, with some studies following established practices in psychology, some in computer science, but mostly in parallel. Studies with an algorithmic focus mostly did not include mental health variables for study, and studies with a mental health focus were largely RCTs and did not include algorithmic evaluations. Finally, the variables that were included varied widely. Different solutions targeted different mental health issues, inputs varied from geographic locations to mood ratings, and there was no single consistent definition of an optimal engagement outcome, nor any data-based evaluation of the relationship between engagement and mental health. These insights allowed a more targeted focus for the remainder of the thesis.

The first step, executed through the work done in the third chapter, was to investigate the influence of money on research in mental health apps. This had been done in other fields, but not for mental health apps. Yet because of the uncertainty regarding the relationship between engagement and mental health improvement in the literature, we considered that a foundational evaluation of how we incentivize participants was necessary. In an RCT where participants were compensated either for each activity they did, or as a one-time payment, we found that step-wise payment was a strong positive predictor of volume of activity completions. However, it did not

influence mental health outcomes. This also led to the first clear insight into the relationship between engagement and improvement, as a higher volume of activity completions did not lead to greater improvement in mental health. Yet participants in our study *did* see a significant improvement overall, especially those with lower mental health scores originally, and the materials we used had been previously tested with positive results in other trials. This would indicate that some form of engagement was necessary to affect positive change through similar mental health apps, and lead us to believe that more research was necessary on the effect of different engagement parameters. Volume may not be critical to improvement, but other facets, such as subjective engagement, perceived effort, or consistency, might be.

The evaluation of improvement and engagement continued in Chapter 4, but this time with a greater focus on generalizability and ecological validity. Here we analyzed a natural dataset, collected over several years through a commercial app. Here we investigated how activity completions and mental health was related across five different mental health domains included in an in-house wellbeing assessment. We found that users who engaged with activities within the app saw significant improvement in all mental health outcomes, whereas users who only completed assessments and no activities did not see significant improvement. For the group of users with activity completions, we also found that change in any one outcome significantly predicted change in the others, indicating a broad utility of these types of solutions. Once more we also found that improvement was more likely for those who needed it the most, as lower baseline scores within the focus variable significantly predicted positive change. However, *higher* baseline scores across the other variables also predicted positive change, which suggests that higher overall mental health may serve protective or supportive functions for enacting change.

Finally, in Chapter 5 we did a deep-dive into mechanisms of change. In a study which included a broad variety of variables, we explored which were most likely to enable positive mental health change. We were especially focused on investigating the relationship between short-term mental health change and long-term mental health outcomes, as well as the influence of subjective engagement with activities. We described a method for collecting composite snapshots of the immediate effects of activities, and showed that some, but not all, momentary ratings of improvement could be useful in predicting long-term health or even generating recommendations. Once more, lower original scores in a measure was the strongest predictor of positive change, and the volume of activity completions did not influence long-term mental health outcomes. However, the number or order of activities *did* influence some short-term outcomes, indicating more research is necessary to firmly establish the relationship between different types of engagement and different measures of mental health.

6.2 Working towards clarity of purpose

The beginning of this thesis described the multitude of paths research could take when studying mental health, mental health apps and personalized mental health solutions. The potential, as with almost any research, is near infinite, and there is not, as such, a wrong way to go. Every study cited has contributed to our knowledge, and the field is saturated with brilliant researchers bringing expertise from a variety of backgrounds, constantly pushing the boundaries of innovation and technology. However, from a birds-eye view it can be difficult to see the individual research efforts as pieces of a greater puzzle. There is a lack of cohesion between findings, very few standard-setting methodological guidelines and different aims and goals that may or may not be aligned. It is quite possible that these things go hand-in-hand. Much of the research that is being done on mental health apps and recommender systems is being done by experts from related fields, with advanced knowledge or capabilities that have been tried and tested in different settings. This can be advantageous, as progress can be fast, but it also means that our understanding is being built from the top down. Assumptions are inherited from past studies in similar fields, countless established treatments and algorithms are being implemented in new ways and innovative solutions pop up at a steady rate. But most studies present new algorithms, or unique variations (Portugal et al., 2018), and a lack of unifying foundational methodologies mean that it is hard to know what is generalizable or in which cases certain treatments or algorithms may shine. The purpose of the research in this thesis was to begin to shore up some of the holes in the foundation of the field. This final chapter will discuss our findings in a larger context, but also highlight how it is only the first step in a long process, and how future research might adapt to fit together within a larger picture.

The first critical step to more cohesive research is a stronger clarity of purpose. Specifically, it is important that the inferences we make and the stated purpose of our research align with our methodologies. The first step comes from acknowledging that mental health apps are built for more than one goal. Certainly there is the hope and the aim that they can help improve people's mental health and give them more autonomous self-care, but it would be a stretch to argue that the huge influx in new commercial apps in recent times stems exclusively from a spike in human altruism. They exist, in part, to make money. Whether it is more for one reason or the other is debatable, but the balance of human nature is a topic for a different researcher to explore. What is important for the current work is that there are multiple stakeholders in research on mental health apps with different goals. How to balance the needs of these stakeholders, and considerations of fairness is an emerging topic, especially within recommender systems (Deldjoo et al., 2024), and to do this we have to understand how their aims align. This was why engagement and improvement were prevalent parallel themes in this thesis. Traditionally, recommender systems were tailored towards increasing user engagement with entertainment items, which was considered beneficial to all parties. Better recommendations lead to higher customer satisfaction and

more revenue for the provider. Health, and mental health are different domains though, and people are not necessarily drawn towards healthy options. In many situations our desire and our wellbeing are in direct competition. We crave food that is bad for us (Stillman & Woolley, 2023), we develop social media addictions that are harmful for our mental health (Sun & Zhang, 2021), we put on sweatpants instead of running shoes, we smoke, we drink, we develop self-destructive behavior. These are not universal rules of course, many people love living healthy lifestyles and feel rewarded by healthy behavior, but it is a common enough trend that we cannot take for granted that what an individual finds engaging is also what is best for them.

In our research, engagement and improvement seemed mostly unrelated, at least when engagement was measured by volume. The number of activities users completed did not predict any of multiple mental health outcomes across three studies in different settings. There were some indications that other engagement metrics may be more important, such as consistency or access interval, and other studies have found that improvement may come down to content or engagement patterns (Ruiz de Villa et al., 2023). On the surface, this could simply be taken as a continuation of established incongruency of findings regarding the effect of engagement on improvement (Lekkas et al., 2021; Linardon & Fuller-Tyszkiewicz, 2020; Molloy & Anderson, 2021), however we would argue that it comes down to compartmentalizing results and creating robust operational definitions. After all, in this thesis our findings were entirely consistent when defining engagement as volume of activity completions. When attempting to understand what drives mental health change, it would be beneficial for future studies to determine which facets of engagement are most important. This can be useful in generating recommendations and guidelines. For example, users could be informed how often they would need to engage with the app to maximize health gain. Or how long they should be using it for at a time. How long they could stop using it for before they would lose the benefits. Or perhaps even many days, weeks, months or years they need to be active before they can expect to plateau, or before temporary improvement becomes permanent. And here again, we must consider how these findings would relate to multi-stakeholder fairness. If future research finds that after a year of consistent usage, a mental health app no longer offers further benefits, users may become inactive or unsubscribe, no longer generating revenue for the provider. Yet they could also advertise apps as lasting solutions and better predict cash-flow. These are, of course, hypothetical scenarios, but they serve as examples for how research might be better framed, establishing clearly who it is meant to benefit, what advantages apps or algorithms have for which outcomes and better highlight multifaceted real world value and opportunity costs.

To achieve this, research also needs clarity of methodological purpose and outcomes. As we established in Chapter 2, there is a tendency for psychological and algorithmic outcomes to be evaluated separately. It is common for early stage research to be hypothesis generating and grounded in theory, and several studies evaluating recommender systems have strong justifica-

tions derived from psychological theory, or established phenomena in traditional psychological research (Lewis et al., 2022; Rohani et al., 2021; Torkamaan & Ziegler, 2022). However, given the inconsistency of the efficacy of mental health apps found in this thesis, as well as in the literature as a whole (Chandrashekar, 2018; Donker et al., 2013; Khademian et al., 2021; Lecomte et al., 2020; Neary & Schueller, 2018; Wang et al., 2018), a theoretical link may not be enough. In order to establish trust in mental health apps as a robust and reliable solution for mental health issues, research has to be clear and precise in what can be understood from its findings. If the goal of a recommender system is to improve mental health, if it is a foundational premise of a study, it has to be measured. This is not to say that research on engagement, and improving engagement is not valuable. It certainly is, but without direct evidence for mental health benefits, the connection is tenuous. The quality of mental health apps varies too much to assume effectiveness without strong empirical support (Anthes, 2016; Torous, 2018). A stronger distinction of purpose should therefore be a focus in future studies. It will be important for research to state clearly whether the focus is engagement, improvement or both, to be precise in their operational definitions of either, and to avoid assumptions of both correlation and causality until the nature of the relationship is more firmly established.

6.3 Establishing methodological guidelines

6.3.1 Defining and selecting engagement and mental health variables

One way cohesion in the field can improve is by attempting to establish more consistent methodological guidelines. This is difficult, because again, the parent fields of psychology and computer science have existing issues which are likely to be inherited. Traditional therapies can, at times lack distinction, with plenty of evidence for efficacy, but less understanding of comparative value (Wampold, 2019). The same can be true for mental health apps, with consumers and clinicians experiencing difficulty in selecting and evaluating appropriate products due to over-attenuation and sparsity of evidence (Neary & Schueller, 2018). Psychology, and several other fields of social science, are undergoing a replication and generalizability crisis, problematic statistical methodologies and inflated results (Yarkoni, 2022). Machine learning algorithms, including recommender systems, are often tested against poor baselines, also leading to overestimation and inflation of effects (Rendle et al., 2019). RCTs, the gold standard for evaluating interventions, often offer little insight on underlying mechanisms or situational differences (Backmann, 2017; Bondemark & Ruf, 2015; Stuart et al., 2015) and heterogeneity and methodological issues are especially common in smartphone-based treatment trials (Tønning et al., 2021). As we have seen throughout this thesis, research on mental health recommender systems suffers from disparate operational definitions and assumptive relationships. These issues are difficult to tackle, and can easily snowball, but we have the tools to improve them. So much individual

research in the field is based on strong methodological ground, and many algorithms deserve more time in the limelight. This section will outline some of the methodological understanding gained from the studies within this thesis, as well as some of the excellent practices of other researchers to establish the beginnings a consistent methodological framework to be developed in future research for more translatable and coherent findings.

First, I have waxed long and lyrical about the need to distinguish between engagement and improvement. But, whilst it should be a higher priority in the field overall, it is not entirely neglected. Several studies have done this at least to some extent, and there are aspects of methodology in most studies which are worth reusing more broadly. One way it is commonly done, and done quite well, is establishing a specific theoretical basis for why an increase in user activity may lead to improved mental health. Torkamaan and Ziegler (2022) strongly justified their recommender system development based on long-standing research on goal-setting theory (Locke & Latham, 2006; Rosenstock et al., 1988). Across two studies, Rohani et al. (2020) and Rohani et al. (2021) established potential benefits from established beneficial results of pleasant event scheduling in the literature (Cuijpers, van Straten, & Warmerdam, 2007; Dimidjian et al., 2006). These are but some examples of how studies with a focus on engagement outcomes still situate their results in mental health contexts. In fact, it could be argued that this was lacking in the current thesis. For the studies described in chapters 3 to 5, the actual theoretical basis for why the activities, or an increase in engagement with activities, might improve mental health was not prioritized highly. As with the studies highlighted for omitting measures of mental health outcomes, we do not believe this reduces the value of our research, as no single study can include everything. But a robust theoretical understanding of the relationship between mental health and engagement is not currently present in the field, and grounding methodologies in psychological theory would be useful for developing our conceptual understanding.

Another key step is to collect data on both engagement and improvement. This has already been discussed multiple times throughout this thesis, and has even been touched upon in this chapter, but is worth framing one final time. If mental health is of interest, to an app or an algorithm, mental health data should be collected. If only engagement data is collected, results should be discussed mostly in terms of engagement outcomes. Similarly, if only mental health data is acquired, few inferences can be made on an app's ability to engage users. As Chapter 2 described, methodologies are often field specific, but discussions extend to outcomes not measured. This was why one thing we consistently included in our studies was at least one measure of engagement and multiple mental health outcomes. We believe this is a necessary step to merge the fields of psychology and computer science in mental health app research.

Beyond the ways we explored that would help deepen our understanding of the variables, one other study deserves spotlight for its methodology in this area. Ruiz de Villa et al. (2023) not only did included measures of both engagement and mental health outcomes, they differentiated

between different types of both, and investigated relationships from different perspectives. They investigated whether the type of activity users engaged with, or the duration of interaction with the app, could predict anxiety, sleep, resilience, mental wellbeing and retention. Their measures and analysis methods differed from ours, but a common factor between their findings and those we presented in Chapter 5 was that outcomes differed substantially depending on the defined variables. We found, for example, that volume of activity completions is unlikely to influence mental health outcomes, however subjective engagement measures, such as effort or satisfaction, might depending on the target mental health variable. They found that the type of activity users engaged with could strongly influence both mental health outcomes and retention (and engagement outcomes), but that duration of interaction was less influential, and only for sleep and mental wellbeing. Other studies (and indications from findings presented in Chapter 5) indicate that long-term access or engagement might be important in facilitating improvement (Schulze et al., 2024), although our findings in Chapter 4 indicated that this may not hold true in a real world setting. Overall, as mentioned in the previous section, this suggests that future studies should be wary in assuming relationships between unmeasured variables, or increase specificity of both methodologies and discussions to targeted variables or phenomena.

This also pertains to the ways we measure and consider mental health outcomes. Several studies have begun to explore the development of recommender systems based on user states or momentary measurements of mood (Beltzer et al., 2022; Lewis et al., 2022; Rohani et al., 2021). These are mostly designed with the purpose of improving engagement or long-term mental health, and the utility of mental health apps as mood boosters or symptom relief tools are relatively unexplored. Of those studies only Beltzer et al. (2022) specifically targeted short-term benefits from their recommendations, where the other two described efforts more designed to affect long-term change. Both approaches, both goals, are viable and valuable. The idea itself, of basing recommendations on predicted mental health outcomes, even if only short-term, may also be key to adapting recommender systems to mental health contexts. However, this again brings us to a point where specificity and clarity are paramount. How we feel in the moment encompasses as many potential options as how our mental health is overall, and is highly prone to situational fluctuation (da Fonseca et al., 2023; Lebois et al., 2016). Currently there is no established gold standard for measuring an individual's mood, state, or the perceived effect of a solution on immediate mental health. Lewis et al. (2022) used data where users rated how they felt relative to before based on a therapy task they just completed, ranging from "worse" to "much better" on a 5 point scale. We utilized a similar rating scale in Chapter 5, but extended it to more specific parameters, such as stress, wellbeing and coping efficacy. Users in the study by Rohani et al. (2021) rated how much they enjoyed the activity, from "not at all" to "a great extent", on a 7 point scale. Whilst mood, state and pleasure related information has been used with promising results to predict future moods (Baumel, Fleming, & Schueller, 2020; Hollis et al., 2017) or even long-term mental health outcomes (Wahle, Kowatsch, Fleisch, Rufer, & Weidt, 2016), general-

izability is questionable when the measures differ. This was a large part of the reasoning behind the design of the study in Chapter 5, to begin to narrow down useful definitions of state and immediate improvement for use in prediction or recommendation. We saw that different perceptions of mental health improvement and activity perception differed in how well they predicted long-term outcomes on more commonly-used measures, and this should be kept in mind for future research. We need to better understand which specific states, situations and perceptions lead to robust gains in which mental health domains, at least if our goal is long-term. Our tools and solutions are more likely to work if their design is streamlined to their purpose. The study by Beltzer et al. (2022) is an excellent example of this principle. Their recommendations were specifically targeted towards short-term emotion regulation solutions for immediate problems. Whilst these would likely carry long-term implications, this was not a primary focus, leading to a clear relationship between their stated purpose and the input and output of their recommender system.

6.3.2 Generalizability and ecological validity

Increasing specificity and clarity of purpose when selecting variables is one aspect of improving generalizability. By understanding nuances between engagement or mental health measurements, and underlying mechanisms of change, we can develop our understanding where solutions are likely to be more effective and where they are not. But, variable selection is only one part of the process. We need to consider the situation, the culture, the individuals, the procedural aspects of our data collection. And, perhaps most importantly, we need to understand whether our results generalize to real world scenarios, that there is ecological validity to what we do. This is an area where the research on recommender systems in mental health apps is quite well situated. Whilst there are certain drawbacks, the simplest way to achieve this is to sample from natural environments, "in the wild", and this is quite commonly done in the field. However the simplest way is not necessarily the easiest, and the entry barrier for this can be high. Not all researchers have the time, resources or access to large natural datasets, and as we demonstrated in Chapter 4 it can also be difficult to isolate variables in a meaningful way through this approach. It is therefore important to understand how to design studies in a more controlled environment that maximize ecological validity and generalizability.

This was the primary purpose of Chapter 3. Whilst there was other information we extracted from the data, the main aim was to investigate whether payment, a common form of participant incentivization in academic studies, influences our results. We found that whilst compensation was the main predictor of number of activities completed, it did not have a significant effect on mental health outcomes. This is not a flashy finding, but it is an example of foundational methodological questions that need to be asked, particularly in a field where engagement outcomes are of considerable importance, especially commercially. And still further research is

required to delineate potential confounds, as the study described in Chapter 3 did compensate all participants. Whilst we could differentiate between levels of compensation, there is still potential that payment of any kind draws in people who would not otherwise participate. Even if individuals motivated by financial incentives do not differ in terms of how much an app or recommender system benefits mental health, there still remains a question of ecological validity regarding engagement. After all, the most effective app in the world would have severely limited usefulness if payment is the only way to encourage users to interact with it.

These considerations were why we wished to compare, as much as possible, our findings in chapters 3 and 5 with findings in a commercial app. And the findings in Chapter 4 did suggest that the mental health benefits of the Foundations app extended beyond controlled environments or paid participants. However, there were other ways in which our findings differed across studies. Perhaps most notably, we did not see a significant improvement in mental wellbeing scores over time in Chapter 5 for users who had access to Foundations activities, despite mental wellbeing consistently improving in our studies in Chapter 3, 4 and other studies using Foundations (Catuara-Solarz et al., 2022; Gnanapragasam et al., 2023; Schulze et al., 2024). This may be because we used a different measure of wellbeing, the WEMWBS-14, as opposed to the WHO-5 used in all the other instances. Or, it could be the different subsample of Foundations activities, as we may have unintentionally removed some of the more impactful ones. It is also possible that it was due to the duration, with users only having access to the app for two weeks rather than the longer durations in other studies. This would be supported by findings by Schulze et al. (2024), indicating two weeks was not enough to enact significant change, although in Chapter 4, there was no significant difference between intervals. This means that although there are directions to explore, there are indications that there are some scenarios in which improvement from Foundations usage isn't generalizable.

This is no surprise, as no app can help everyone in every situation. With enough testing there are bound to be samples of users who do not experience improvement, from any intervention. And the purpose of understanding generalizability is not necessarily to develop apps or research that is generalizable to the broadest population. It is likely more worthwhile to understand in which cases, and across which parameters, mental health improvement is consistent. For example, one finding that repeated itself across all of our studies was that the strongest predictor of positive mental health change was lower mental health scores on the first measurement occasion. This could suggest that the greatest potential benefit exists for those who need it the most. Foundations has shown efficacy in both general samples (Schulze et al., 2024) and samples consisting only of participants who scored below certain mental health thresholds (Catuara-Solarz et al., 2022), the effect may be stronger, more robust, more generalizable across low mental health subsamples. Understanding this has practical value and opens up further research areas. Practically, Foundations, and apps with similar findings can market themselves more efficiently to

users likely to benefit. Recommender systems and prediction algorithms may find it worthwhile to include mental health information in user modeling to strengthen their utility. Future research can also further establish differences. Are differences in generalizability within subsamples due to individual differences, situational factors, content or some other factor? If this can be delineated, development and outreach can be further improved and personalized.

6.3.3 Algorithmic considerations

This thesis has not focused on the development of algorithms. It has been focused on creating a space in which those who build recommender systems have all the information they need to do so optimally. This section will discuss how some of our findings can be directly included in personalization techniques, and which avenues of future research might improve upon this even further. First, there has been a consistent theme of understanding the mechanisms of change and understanding the users we design the activities for. This is valuable for theoretical progress, but it also has a direct bearing on how we recommend items. It is common to design recommendations based on user behavior or patterns, but any information about users can be implemented in models. For example, personality has been used to generate appropriate music recommendations (Ferwerda & Schedl, 2016). Torkamaan and Ziegler (2022) defined user ability in a mental health recommender based on algorithms more commonly used in chess and gaming. The earliest recommender system recommended books based on user stereotypes (Rich, 1979). Similarly, the mechanisms of change evaluated in the current research could be included for modeling purposes. Low mental health was consistently the strongest predictor of improvement, so if future research establishes whether items or activities differentiate in their effectiveness based on mental health levels, this information could be included in models to further tailor recommendations. Personality could be used in a similar way, perhaps by the classification of users into archetypes such as demonstrated by Aziz et al. (2023), although the findings described in Chapter 5 indicated further exploration may be prudent first. Arguments have also been made for avoiding "greedy algorithms" (Larson, 2017) by reducing variables and data requirements, which again highlights the importance of further understanding mechanisms of change. This thesis, and related work, can in some ways be considered as part of a selection process, exploring and identifying the most useful variables to be further tested, to identify the most important factors in enabling mental health improvement.

In this chapter we have also discussed clarity of purpose from a conceptual and methodological standpoint. Yet this is another area which could be directly explored via recommender systems. In Chapter 5 we built a proof-of-concept recommender system, intended to recommend activities with the highest momentary benefits for user mental health. This state-level improvement has been pursued in other research (Beltzer et al., 2022), but we also selected our rating parameters based on the relationship with long-term improvement in the same dataset. However, even this is

a somewhat indirect approach if long-term improvement is the goal. With current technological capabilities, there is no reason we could not estimate which items or solutions have the greatest direct effect on long-term gains. Machine learning has been used before to predict treatment outcomes (Greenberg et al., 2024), and research using the Foundations app has explored the effect of individual activity modules on mental health outcomes (Ruiz De villa et al., 2024; Ruiz de Villa et al., 2023). It would be possible to re-purpose aspects of their methodology to recommendation purposes, and investigate the effect in RCTs. Given the concerns discussed here, it would also be particularly interesting to compare these directly with recommender systems focused on engagement, and measuring the effect of each on user retention and improvement. This would go a long way to contextualizing future research in the field.

6.4 Conclusion

It is difficult to write a concise conclusion to a work of this length and this magnitude. Especially because, for a PhD thesis, I suspect it is relatively scattered in terms of results. The work done was exploratory throughout and included investigation of many variables, and few certain results. However, in some ways, this is a reflection of the field. The way mental health apps and recommender systems are studied, from a birds-eye view is extremely varied. It is a competitive market, both commercially and academically. Very few studies are replicated, very few algorithms are reused between studies and more and more apps are becoming available. It is a field categorized by exceptional individual efforts and technological innovations, but low coherence and robustness overall. We continue to see new solutions that work, but there is less time and effort devoted to understanding why, which reduces our ability to choose between them and understand their relative strengths in different situations. This needs to be remedied. New solutions are valuable, and efforts to develop them should not grind to a halt. But they are being disproportionately favored. With more options without a solid understanding of when and how they are most effective (or in many cases without evidence of efficacy at all), companies, consumers and clinicians are left with choice blindness and little ability to differentiate between apps or algorithms.

The current work offered some insight into how to better do this. We can increase clarity in our research by refining our operational definitions, and use this in our selection of variables for research and recommendation. "Increasing engagement" is a vague concept, and depending on how engagement is defined it may carry little value in improving mental health outcomes. Our most consistent finding was that increased volume of activity completion had little impact on any mental health measures. However, the types of activity might, depending on the specific outcome, or the targeted timeframe. Length of access, consistency, satisfaction, perceived effort, duration of interaction, these are some engagement metrics that may, pending replication and further study, prove very useful in helping users. This is important for personalization. Rec-

ommender systems that focus on increasing activity completions are unlikely to achieve mental health goals directly. They may be indirectly useful, by increasing the likelihood that users return or are more consistent, but it is also possible that they lead to boredom or burnout, achieving the opposite effect. This, of course, does not speak to commercial use, where increasing activity completions may be exactly what a provider aims for, and we should not dismiss the needs of the developers and owners. Mental health apps that do not make money likely will not survive and receive continued support, and no mental health activity is effective if nobody engages with it.

The work done in this thesis was only the first step. The field of research is still in the early stages overall, and it will take time and effort to advance it. But we need to begin to establish robust methodological frameworks, understand generalizability of findings and merge the expertise of various interdisciplinary fields. It is not uncommon for research to be scattered and experimental in early stages, but there are plenty of examples of fields which have progressed to advanced levels that suffer from a lack of ability to replicate early findings, and where methodologies are defined by old habits. It is now, whilst the field is young, that we can best address this. The strength of research is almost always in communal advancement, and that is more easily achieved with coherent findings, repeated analysis and clarity of purpose.

References

- Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., . . . Pizzato, L. (2020, 3). Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30(1), 127–158. doi: 10.1007/s11257-019-09256-1
- Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000, 11). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, White women. *Health Psychology*, 19(6), 586–592. doi: 10.1037/0278-6133.19.6.586
- Ali, S., Rhodes, L., Moreea, O., McMillan, D., Gilbody, S., Leach, C., . . . Delgadillo, J. (2017, 7). How durable is the effect of low intensity CBT for depression and anxiety? Remission and relapse in a longitudinal cohort study. *Behaviour Research and Therapy*, 94, 1–8. doi: 10.1016/j.brat.2017.04.006
- Almeida, R. F. S. D. E., Sousa, T. J., Couto, A. S., Marques, A. J., Queirós, C. M., & Martins, C. L. (2019, 2). Development of weCope, a mobile app for illness self-management in schizophrenia. *Archives of Clinical Psychiatry (São Paulo)*, 46(1), 1–4. doi: 10.1590/0101-60830000000182
- Alqahtani, F., Meier, S., & Orji, R. (2022, 7). Personality-based approach for tailoring persuasive mental health applications. *User Modeling and User-Adapted Interaction*, 32(3), 253–295. doi: 10.1007/s11257-021-09289-5
- Ameko, M. K., Beltzer, M. L., Cai, L., Boukhechba, M., Teachman, B. A., & Barnes, L. E. (2020, 9). Offline Contextual Multi-armed Bandits for Mobile Health Interventions: A Case Study on Emotion Regulation. In *Fourteenth acm conference on recommender systems* (pp. 249–258). New York, NY, USA: ACM. doi: 10.1145/3383313.3412244
- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association. doi: 10.1176/appi.books.9780890425596
- Andrade, L. H., Alonso, J., Mneimneh, Z., Wells, J. E., Al-Hamzawi, A., Borges, G., . . . Kessler, R. C. (2014). Barriers to mental health treatment: Results from the WHO World Mental Health surveys. *Psychological Medicine*, 44(6). doi: 10.1017/S0033291713001943
- Anthes, E. (2016). Pocket psychiatry: mobile mental-health apps have exploded onto the market, but few have been thoroughly tested. *Nature*, 532(7597).

- Arango, C., Dragioti, E., Solmi, M., Cortese, S., Domschke, K., Murray, R. M., . . . Fusar-Poli, P. (2021, 10). Risk and protective factors for mental disorders beyond genetics: an evidence-based atlas. *World Psychiatry, 20*(3), 417–436. doi: 10.1002/wps.20894
- Ashton, M. C., & Lee, K. (2009, 1). An Investigation of Personality Types within the HEXACO Personality Framework. *Journal of Individual Differences, 30*(4), 181–187. doi: 10.1027/1614-0001.30.4.181
- Auerbach, R. P., Mortier, P., Bruffaerts, R., Alonso, J., Benjet, C., Cuijpers, P., . . . Kessler, R. C. (2018, 10). WHO World Mental Health Surveys International College Student Project: Prevalence and distribution of mental disorders. *Journal of Abnormal Psychology, 127*(7), 623–638. doi: 10.1037/abn0000362
- Aziz, M., Erbad, A., Belhaouari, S. B., Almourad, M. B., Altuwairiqi, M., & Ali, R. (2023, 1). Who Uses Mhealth? User Archetypes for Physical and Mental Health Apps. *DIGITAL HEALTH, 9*, 205520762311521. doi: 10.1177/20552076231152175
- Backmann, M. (2017, 12). What's in a gold standard? In defence of randomised controlled trials. *Medicine, Health Care and Philosophy, 20*(4), 513–523. doi: 10.1007/s11019-017-9773-2
- Barsalou, L. W. (2019, 10). Establishing Generalizable Mechanisms. *Psychological Inquiry, 30*(4), 220–230. doi: 10.1080/1047840X.2019.1693857
- Baumel, A., Fleming, T., & Schueller, S. M. (2020, 10). Digital Micro Interventions for Behavioral and Mental Health Gains: Core Components and Conceptualization of Digital Micro Intervention Care. *Journal of Medical Internet Research, 22*(10), e20631. doi: 10.2196/20631
- Becker, D., van Breda, W., Funk, B., Hoogendoorn, M., Ruwaard, J., & Riper, H. (2018, 6). Predictive modeling in e-mental health: A common language framework. *Internet Interventions, 12*, 57–67. doi: 10.1016/j.invent.2018.03.002
- Beel, J., Genzmehr, M., Langer, S., Nürnberger, A., & Gipp, B. (2013). A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Acm international conference proceeding series*. doi: 10.1145/2532508.2532511
- Beel, J., & Langer, S. (2015). A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 9316). doi: 10.1007/978-3-319-24592-8_12
- Beltzer, M. L., Ameko, M. K., Daniel, K. E., Daros, A. R., Boukhechba, M., Barnes, L. E., & Teachman, B. A. (2022, 1). Building an emotion regulation recommender algorithm for socially anxious individuals using contextual bandits. *British Journal of Clinical Psychology, 61*(S1), 51–72. doi: 10.1111/bjc.12282
- Bentley, J. P., & Thacker, P. G. (2004, 6). The influence of risk and monetary payment on

- the research participation decision making process. *Journal of Medical Ethics*, 30(3), 293–298. doi: 10.1136/jme.2002.001594
- Bondemark, L., & Ruf, S. (2015, 10). Randomized controlled trial: the gold standard or an unobtainable fallacy? *The European Journal of Orthodontics*, 37(5), 457–461. doi: 10.1093/ejo/cjv046
- Bostock, S., Crosswell, A. D., Prather, A. A., & Steptoe, A. (2019, 2). 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. doi: 10.1037/ocp0000118
- Bot, S. D. M., Terwee, C. B., van der Windt, D. A. W. M., Bouter, L. M., Dekker, J. M., & de Vet, H. C. W. (2003). Psychometric evaluation of self-report questionnaires: the development of a checklist. *Proceedings of the second workshop on research methodology*, 161–168.
- Boyd, A., Van de Velde, S., Vilagut, G., de Graaf, R., O'Neill, S., Florescu, S., ... Kovess-Masfety, V. (2015, 3). Gender differences in mental disorders and suicidality in Europe: Results from a large cross-sectional population-based study. *Journal of Affective Disorders*, 173, 245–254. doi: 10.1016/j.jad.2014.11.002
- Brennan, R. L. (2001). *Generalizability Theory*. New York, NY: Springer New York. doi: 10.1007/978-1-4757-3456-0
- Broman, J.-E., Smedje, H., Mallon, L., & Hetta, J. (2008, 1). The Minimal Insomnia Symptom Scale (MISS). *Uppsala Journal of Medical Sciences*, 113(2), 131–142. doi: 10.3109/2000-1967-221
- Bucher, M. A., Suzuki, T., & Samuel, D. B. (2019, 6). A meta-analytic review of personality traits and their associations with mental health treatment outcomes. *Clinical Psychology Review*, 70, 51–63. doi: 10.1016/j.cpr.2019.04.002
- Butryn, T., Bryant, L., Marchionni, C., & Sholevar, F. (2017). The shortage of psychiatrists and other mental health providers: Causes, current state, and potential solutions. *International Journal of Academic Medicine*, 3(1), 5. doi: 10.4103/IJAM.IJAM_49_17
- Catuara-Solarz, S., Skorulski, B., Estella-Aguerri, I., Avella-Garcia, C. B., Shepherd, S., Stott, E., ... Dix, S. (2022, 7). The Efficacy of “Foundations,” a Digital Mental Health App to Improve Mental Well-being During COVID-19: Proof-of-Principle Randomized Controlled Trial. *JMIR mHealth and uHealth*, 10(7), e30976. doi: 10.2196/30976
- Chandrashekar, P. (2018). Do mental health mobile apps work: evidence and recommendations for designing high-efficacy mental health mobile apps. *mHealth*, 4. doi: 10.21037/mhealth.2018.03.02
- Chen, X., Zhang, Y., & Wen, J.-R. (2022, 2). Measuring "Why" in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation.
- Clement, S., Schauman, O., Graham, T., Maggioni, F., Evans-Lacko, S., Bezborodovs, N., ... Thornicroft, G. (2015, 1). What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological*

- Medicine*, 45(1), 11–27. doi: 10.1017/S0033291714000129
- Coffey, M. (2004, 7). Stress in Social Services: Mental Wellbeing, Constraints and Job Satisfaction. *British Journal of Social Work*, 34(5), 735–746. doi: 10.1093/bjsw/bch088
- Cohen, S., Kamarck, T., & Mermelstein, R. (1983, 12). A Global Measure of Perceived Stress. *Journal of Health and Social Behavior*, 24(4), 385. doi: 10.2307/2136404
- Constantino, M. J., Coyne, A. E., Boswell, J. F., Iles, B. R., & Višlā, A. (2018, 12). A meta-analysis of the association between patients' early perception of treatment credibility and their posttreatment outcomes. *Psychotherapy*, 55(4), 486–495. doi: 10.1037/pst0000168
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963, 11). Theory of generalizability: A liberalization of reliability theory†. *British Journal of Statistical Psychology*, 16(2), 137–163. doi: 10.1111/j.2044-8317.1963.tb00206.x
- Cruwys, T., Haslam, S. A., Dingle, G. A., Haslam, C., & Jetten, J. (2014, 8). Depression and Social Identity. *Personality and Social Psychology Review*, 18(3), 215–238. doi: 10.1177/1088868314523839
- Cuijpers, P., Marks, I. M., van Straten, A., Cavanagh, K., Gega, L., & Andersson, G. (2009, 6). Computer-Aided Psychotherapy for Anxiety Disorders: A Meta-Analytic Review. *Cognitive Behaviour Therapy*, 38(2), 66–82. doi: 10.1080/16506070802694776
- Cuijpers, P., van Straten, A., & Warmerdam, L. (2007, 4). Behavioral activation treatments of depression: A meta-analysis. *Clinical Psychology Review*, 27(3), 318–326. doi: 10.1016/j.cpr.2006.11.001
- Czaja, S. J., & Sharit, J. (2003, 3). Practically Relevant Research: Capturing Real World Tasks, Environments, and Outcomes. *The Gerontologist*, 43(suppl_1), 9–18. doi: 10.1093/geront/43.suppl_1.9
- Dacrema, M. F., Cremonesi, P., & Jannach, D. (2019). Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Recsys 2019 - 13th acm conference on recommender systems*. doi: 10.1145/3298689.3347058
- da Fonseca, M., Maffei, G., Moreno-Bote, R., & Hyafil, A. (2023, 2). Mood and implicit confidence independently fluctuate at different time scales. *Cognitive, Affective, & Behavioral Neuroscience*, 23(1), 142–161. doi: 10.3758/s13415-022-01038-4
- De Croon, R., Van Houdt, L., Htun, N. N., Štiglic, G., Abeele, V. V., & Verbert, K. (2021). *Health recommender systems: Systematic review* (Vol. 23) (No. 6). doi: 10.2196/18035
- Deldjoo, Y., Jannach, D., Bellogin, A., Difonzo, A., & Zanzonelli, D. (2024, 3). Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction*, 34(1), 59–108. doi: 10.1007/s11257-023-09364-z
- Denecke, K., Schmid, N., & Nüssli, S. (2022). *Implementation of Cognitive Behavioral Therapy in e-Mental Health Apps: Literature Review* (Vol. 24) (No. 3). doi: 10.2196/27791
- Dimidjian, S., Hollon, S. D., Dobson, K. S., Schmaling, K. B., Kohlenberg, R. J., Addis, M. E., ... Jacobson, N. S. (2006). Randomized trial of behavioral activation, cognitive therapy,

- and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology*, 74(4), 658–670. doi: 10.1037/0022-006X.74.4.658
- 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* (Vol. 15) (No. 11). doi: 10.2196/jmir.2791
- Dutriaux, L., Clark, N. E., Papiés, E. K., Scheepers, C., & Barsalou, L. W. (2023, 6). The Situated Assessment Method (SAM2): Establishing individual differences in habitual behavior. *PLOS ONE*, 18(6), e0286954. doi: 10.1371/journal.pone.0286954
- Edbrooke-Childs, J., & Deighton, J. (2020, 11). Problem severity and waiting times for young people accessing mental health services. *BJPsych Open*, 6(6), e118. doi: 10.1192/bjo.2020.103
- Ehrhardt, N. M., Fietz, J., Kopf-Beck, J., Kappelmann, N., & Brem, A. (2022, 5). Separating EEG correlates of stress: Cognitive effort, time pressure, and social-evaluative threat. *European Journal of Neuroscience*, 55(9-10), 2464–2473. doi: 10.1111/ejn.15211
- Eisenberg, D., Golberstein, E., & Gollust, S. E. (2007, 7). Help-Seeking and Access to Mental Health Care in a University Student Population. *Medical Care*, 45(7), 594–601. doi: 10.1097/MLR.0b013e31803bb4c1
- Ekstrand, M. D., Das, A., Burke, R., & Diaz, F. (2022). Fairness in Information Access Systems. *Foundations and Trends® in Information Retrieval*, 16(1-2), 1–177. doi: 10.1561/15000000079
- Fan, P. J. . H. Z., P. (2013). The discussion of current mental health status and mental health management strategy in China. *Journal of Practical Medical Techniques*, 20(8), 911–912.
- Ferré-Bergadà, M., Valls, A., Raigal-Aran, L., Lorca-Cabrera, J., Albacar-Riobóo, N., Lluch-Canut, T., & Ferré-Grau, C. (2021, 12). A method to determine a personalized set of online exercises for improving the positive mental health of a caregiver of a chronically ill patient. *BMC Medical Informatics and Decision Making*, 21(1), 74. doi: 10.1186/s12911-021-01445-6
- Ferwerda, B., & Schedl, M. (2016). Personality-Based User Modeling for Music Recommender Systems. In (pp. 254–257). doi: 10.1007/978-3-319-46131-1_29
- Ghazisaeedi, M., Mahmoodi, H., Arpaci, I., Mehrdar, S., & Barzegari, S. (2022, 6). Validity, Reliability, and Optimal Cut-off Scores of the WHO-5, PHQ-9, and PHQ-2 to Screen Depression Among University Students in Iran. *International Journal of Mental Health and Addiction*, 20(3), 1824–1833. doi: 10.1007/s11469-021-00483-5
- Gnanapragasam, S. N., Tinch-Taylor, R., Scott, H. R., Hegarty, S., Souliou, E., Bhundia, R., ... Wessely, S. (2023, 2). Multicentre, England-wide randomised controlled trial of the ‘Foundations’ smartphone application in improving mental health and well-being in a healthcare worker population. *The British Journal of Psychiatry*, 222(2), 58–66. doi:

10.1192/bjp.2022.103

- Grady, C. (2019, 9). The Continued Complexities of Paying Research Participants. *The American Journal of Bioethics*, 19(9), 5–7. doi: 10.1080/15265161.2019.1643654
- Graham, A. K., Kwasny, M. J., Lattie, E. G., Greene, C. J., Gupta, N. V., Reddy, M., & Mohr, D. C. (2021, 9). Targeting subjective engagement in experimental therapeutics for digital mental health interventions. *Internet Interventions*, 25, 100403. doi: 10.1016/j.invent.2021.100403
- Greenberg, J. L., Weingarden, H., Hoepfner, S. S., Berger-Gutierrez, R. M., Klare, D., Snorrason, I., . . . Wilhelm, S. (2024, 6). Predicting response to a smartphone-based cognitive-behavioral therapy for body dysmorphic disorder. *Journal of Affective Disorders*, 355, 106–114. doi: 10.1016/j.jad.2024.03.044
- Hedin, G., Garmy, P., Norell-Clarke, A., Tønnesen, H., Hagell, P., & Westergren, A. (2022, 12). Measurement properties of the minimal insomnia symptom scale (MISS) in adolescents. *Sleep Science and Practice*, 6(1), 5. doi: 10.1186/s41606-022-00075-9
- Hellström, A., Hagell, P., Fagerström, C., & Willman, A. (2010, 12). Measurement properties of the Minimal Insomnia Symptom Scale (MISS) in an elderly population in Sweden. *BMC Geriatrics*, 10(1), 84. doi: 10.1186/1471-2318-10-84
- Henckens, M. J. A. G., Klumpers, F., Everaerd, D., Kooijman, S. C., van Wingen, G. A., & Fernández, G. (2016, 4). Interindividual differences in stress sensitivity: basal and stress-induced cortisol levels differentially predict neural vigilance processing under stress. *Social Cognitive and Affective Neuroscience*, 11(4), 663–673. doi: 10.1093/scan/nsv149
- Henry, S., Thielmann, I., Booth, T., & Möttus, R. (2022, 1). Test-retest reliability of the HEXACO-100—And the value of multiple measurements for assessing reliability. *PLOS ONE*, 17(1), e0262465. doi: 10.1371/journal.pone.0262465
- Henwood, A., Guerreiro, J., Matic, A., & Dolan, P. (2022, 1). The duration of daily activities has no impact on measures of overall wellbeing. *Scientific Reports*, 12(1), 514. doi: 10.1038/s41598-021-04606-9
- Hollis, V., Konrad, A., Springer, A., Antoun, M., Antoun, C., Martin, R., & Whittaker, S. (2017, 11). What Does All This Data Mean for My Future Mood? Actionable Analytics and Targeted Reflection for Emotional Well-Being. *Human-Computer Interaction*, 32(5-6), 208–267. doi: 10.1080/07370024.2016.1277724
- Holtz, B. E., Kanthawala, S., Martin, K., Nelson, V., & Parrott, S. (2023, 7). Young adults' adoption and use of mental health apps: efficient, effective, but no replacement for in-person care. *Journal of American College Health*, 1–9. doi: 10.1080/07448481.2023.2227727
- Hsieh, G., & Kocielnik, R. (2016, 2). You Get Who You Pay for. In *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing* (pp. 823–835). New York, NY, USA: ACM. doi: 10.1145/2818048.2819936
- Hussey, I., & Hughes, S. (2020, 6). Hidden Invalidity Among 15 Commonly Used Measures in

- Social and Personality Psychology. *Advances in Methods and Practices in Psychological Science*, 3(2), 166–184. doi: 10.1177/2515245919882903
- Inkster, B., Sarda, S., & Subramanian, V. (2018, 11). An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth*, 6(11), e12106. doi: 10.2196/12106
- Jannach, D., & Jugovac, M. (2019, 12). Measuring the Business Value of Recommender Systems. *ACM Transactions on Management Information Systems*, 10(4), 1–23. doi: 10.1145/3370082
- Jauhar, S., Laws, K. R., & McKenna, P. J. (2019, 6). CBT for schizophrenia: a critical viewpoint. *Psychological Medicine*, 49(8), 1233–1236. doi: 10.1017/S0033291718004166
- Khademian, F., Aslani, A., & Bastani, P. (2021, 12). The effects of mobile apps on stress, anxiety, and depression: overview of systematic reviews. *International Journal of Technology Assessment in Health Care*, 37(1), e4. doi: 10.1017/S0266462320002093
- Khwaja, M., Pieritz, S., Faisal, A. A., & Matic, A. (2021, 6). Personality and Engagement with Digital Mental Health Interventions. In *Proceedings of the 29th acm conference on user modeling, adaptation and personalization* (pp. 235–239). New York, NY, USA: ACM. doi: 10.1145/3450613.3456823
- Knaevelsrud, C., & Maercker, A. (2010, 3). Long-Term Effects of an Internet-Based Treatment for Posttraumatic Stress. *Cognitive Behaviour Therapy*, 39(1), 72–77. doi: 10.1080/16506070902999935
- Knight, R. C., & Emery, L. J. (2022, 7). Immediate and delayed effects of suppression and mindfulness as emotion regulation strategies*. *Anxiety, Stress, & Coping*, 35(4), 474–487. doi: 10.1080/10615806.2021.1978430
- Kopelovich, S. L., Monroe-DeVita, M., Buck, B. E., Brenner, C., Moser, L., Jarskog, L. F., . . . Chwastiak, L. A. (2021). Community Mental Health Care Delivery During the COVID-19 Pandemic: Practical Strategies for Improving Care for People with Serious Mental Illness. *Community Mental Health Journal*, 57(3). doi: 10.1007/s10597-020-00662-z
- Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001, 9). The PHQ-9. *Journal of General Internal Medicine*, 16(9), 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x
- Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2003, 11). The Patient Health Questionnaire-2. *Medical Care*, 41(11), 1284–1292. doi: 10.1097/01.MLR.0000093487.78664.3C
- Kuehn, B. M. (2022, 6). Clinician Shortage Exacerbates Pandemic-Fueled “Mental Health Crisis”. *JAMA*, 327(22), 2179. doi: 10.1001/jama.2022.8661
- Lara-Cabrera, M., Betancort, M., Muñoz-Rubilar, A., Rodríguez-Novo, N., Bjerkeset, O., & De las Cuevas, C. (2022, 8). Psychometric Properties of the WHO-5 Well-Being Index among Nurses during the COVID-19 Pandemic: A Cross-Sectional Study in Three Countries. *International Journal of Environmental Research and Public Health*, 19(16), 10106.

doi: 10.3390/ijerph191610106

- Lara-Cabrera, M. L., Betancort, M., Muñoz-Rubilar, C. A., Rodríguez Novo, N., & De las Cuevas, C. (2021, 9). The Mediating Role of Resilience in the Relationship between Perceived Stress and Mental Health. *International Journal of Environmental Research and Public Health*, 18(18), 9762. doi: 10.3390/ijerph18189762
- Largent, E. A., & Lynch, H. F. (2017). Paying Research Participants: The Outsized Influence of "Undue Influence". *IRB*, 39(4), 1–9.
- Larson, Z. A. L. B. . C. P., M. (2017). Towards minimal necessary data: The case for analyzing training data requirements of recommender algorithms.
- Lattie, E. G., Schueller, S. M., Sargent, E., Stiles-Shields, C., Tomasino, K. N., Corden, M. E., ... Mohr, D. C. (2016, 5). Uptake and usage of IntelliCare: A publicly available suite of mental health and well-being apps. *Internet Interventions*, 4, 152–158. doi: 10.1016/j.invent.2016.06.003
- Lebois, L. A., Hertzog, C., Slavich, G. M., Barrett, L. F., & Barsalou, L. W. (2016, 9). Establishing the situated features associated with perceived stress. *Acta Psychologica*, 169, 119–132. doi: 10.1016/j.actpsy.2016.05.012
- Lecomte, T., Potvin, S., Corbière, M., Guay, S., Samson, C., Cloutier, B., ... Khazaal, Y. (2020, 5). Mobile Apps for Mental Health Issues: Meta-Review of Meta-Analyses. *JMIR mHealth and uHealth*, 8(5), e17458. doi: 10.2196/17458
- Lee, E.-H. (2012, 12). Review of the Psychometric Evidence of the Perceived Stress Scale. *Asian Nursing Research*, 6(4), 121–127. doi: 10.1016/j.anr.2012.08.004
- Lee, S. Y., & Lee, S. W. (2023, 4). Normative or Effective? The Role of News Diversity and Trust in News Recommendation Services. *International Journal of Human–Computer Interaction*, 39(6), 1216–1229. doi: 10.1080/10447318.2022.2057116
- Leech, T., Dorstyn, D., Taylor, A., & Li, W. (2021, 8). Mental health apps for adolescents and young adults: A systematic review of randomised controlled trials. *Children and Youth Services Review*, 127, 106073. doi: 10.1016/j.childyouth.2021.106073
- Lekkas, D., Price, G., McFadden, J., & Jacobson, N. C. (2021, 10). The Application of Machine Learning to Online Mindfulness Intervention Data: a Primer and Empirical Example in Compliance Assessment. *Mindfulness*, 12(10), 2519–2534. doi: 10.1007/s12671-021-01723-4
- Levis, B., Sun, Y., He, C., Wu, Y., Krishnan, A., Bhandari, P. M., ... Thombs, B. D. (2020, 6). Accuracy of the PHQ-2 Alone and in Combination With the PHQ-9 for Screening to Detect Major Depression. *JAMA*, 323(22), 2290. doi: 10.1001/jama.2020.6504
- Levy, H. C., O’Bryan, E. M., & Tolin, D. F. (2021, 6). A meta-analysis of relapse rates in cognitive-behavioral therapy for anxiety disorders. *Journal of Anxiety Disorders*, 81, 102407. doi: 10.1016/j.janxdis.2021.102407
- Lewis, R., Ferguson, C., Wilks, C., Jones, N., & Picard, R. W. (2022, 4). Can a Recommender

- System Support Treatment Personalisation in Digital Mental Health Therapy? A Quantitative Feasibility Assessment Using Data from a Behavioural Activation Therapy App. In *Chi conference on human factors in computing systems extended abstracts* (pp. 1–8). New York, NY, USA: ACM. doi: 10.1145/3491101.3519840
- Linardon, J., & Fuller-Tyszkiewicz, M. (2020, 1). 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. doi: 10.1037/ccp0000459
- Liu, S., Lithopoulos, A., Zhang, C.-Q., Garcia-Barrera, M. A., & Rhodes, R. E. (2021, 1). Personality and perceived stress during COVID-19 pandemic: Testing the mediating role of perceived threat and efficacy. *Personality and Individual Differences, 168*, 110351. doi: 10.1016/j.paid.2020.110351
- Locke, E. A., & Latham, G. P. (2006, 10). New Directions in Goal-Setting Theory. *Current Directions in Psychological Science, 15*(5), 265–268. doi: 10.1111/j.1467-8721.2006.00449.x
- López-López, J. A., Davies, S. R., Caldwell, D. M., Churchill, R., Peters, T. J., Tallon, D., ... Welton, N. J. (2019, 9). The process and delivery of CBT for depression in adults: a systematic review and network meta-analysis. *Psychological Medicine, 49*(12), 1937–1947. doi: 10.1017/S003329171900120X
- Luo, J., Zhang, B., Cao, M., & Roberts, B. W. (2023, 5). The Stressful Personality: A Meta-Analytical Review of the Relation Between Personality and Stress. *Personality and Social Psychology Review, 27*(2), 128–194. doi: 10.1177/10888683221104002
- Macaskill, A. (2013, 8). The mental health of university students in the United Kingdom. *British Journal of Guidance & Counselling, 41*(4), 426–441. doi: 10.1080/03069885.2012.743110
- Madadipouya, K., & Sivananthan Chelliah. (2017). A Literature Review on Recommender Systems Algorithms, Techniques and Evaluations. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 8*(2).
- Malhi, G. S., Hamilton, A., Morris, G., Mannie, Z., Das, P., & Outhred, T. (2017, 11). The promise of digital mood tracking technologies: are we heading on the right track? *Evidence Based Mental Health, 20*(4), 102–107. doi: 10.1136/eb-2017-102757
- Malik, A. O., Peri-Okonny, P., Gosch, K., Thomas, M., Mena, C., Hiatt, W. R., ... Smolderen, K. G. (2020, 6). Association of Perceived Stress Levels With Long-term Mortality in Patients With Peripheral Artery Disease. *JAMA Network Open, 3*(6), e208741. doi: 10.1001/jamanetworkopen.2020.8741
- Marshall, J. M., Dunstan, D. A., & Bartik, W. (2020). *Clinical or gimmickal: The use and effectiveness of mobile mental health apps for treating anxiety and depression* (Vol. 54) (No. 1). doi: 10.1177/0004867419876700

- May, J. M., Richardi, T. M., & Barth, K. S. (2016, 3). Dialectical behavior therapy as treatment for borderline personality disorder. *Mental Health Clinician*, *6*(2), 62–67. doi: 10.9740/mhc.2016.03.62
- Mayer, G., Hummel, S., Oetjen, N., Gronewold, N., Bubolz, S., Blankenhagel, K., ... Schultz, J.-H. (2022, 1). User experience and acceptance of patients and healthy adults testing a personalized self-management app for depression: A non-randomized mixed-methods feasibility study. *DIGITAL HEALTH*, *8*, 205520762210913. doi: 10.1177/20552076221091353
- McDermott, M. B. A., Wang, S., Marinsek, N., Ranganath, R., Ghassemi, M., & Foschini, L. (2019, 7). Reproducibility in Machine Learning for Health.
- McGrath, J. J., Lim, C. C. W., Plana-Ripoll, O., Holtz, Y., Agerbo, E., Momen, N. C., ... de Jonge, P. (2020, 8). Comorbidity within mental disorders: a comprehensive analysis based on 145 990 survey respondents from 27 countries. *Epidemiology and Psychiatric Sciences*, *29*, e153. doi: 10.1017/S2045796020000633
- McLafferty, M., Lapsley, C. R., Ennis, E., Armour, C., Murphy, S., Bunting, B. P., ... O'Neill, S. M. (2017, 12). Mental health, behavioural problems and treatment seeking among students commencing university in Northern Ireland. *PLOS ONE*, *12*(12), e0188785. doi: 10.1371/journal.pone.0188785
- Meldrum, M. L. (2000, 8). A BRIEF HISTORY OF THE RANDOMIZED CONTROLLED TRIAL. *Hematology/Oncology Clinics of North America*, *14*(4), 745–760. doi: 10.1016/S0889-8588(05)70309-9
- Miller, L. C., Shaikh, S. J., Jeong, D. C., Wang, L., Gillig, T. K., Godoy, C. G., ... Read, S. J. (2019, 10). Causal Inference in Generalizable Environments: Systematic Representative Design. *Psychological Inquiry*, *30*(4), 173–202. doi: 10.1080/1047840X.2019.1693866
- Mitchell, A. J., Yadegarfar, M., Gill, J., & Stubbs, B. (2016, 3). Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. *BJPsych Open*, *2*(2), 127–138. doi: 10.1192/bjpo.bp.115.001685
- Mitchell, M. D., Gehrman, P., Perlis, M., & Umscheid, C. A. (2012, 12). Comparative effectiveness of cognitive behavioral therapy for insomnia: a systematic review. *BMC Family Practice*, *13*(1), 40. doi: 10.1186/1471-2296-13-40
- Molloy, A., & Anderson, P. L. (2021, 12). Engagement with mobile health interventions for depression: A systematic review. *Internet Interventions*, *26*, 100454. doi: 10.1016/j.invent.2021.100454
- More, B. K. A. M. C. . P. L. A., K. R. (2022). Paying participants: The impact of compensation on data quality. *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, *29*(4), 403–417.
- Neary, M., & Schueller, S. M. (2018). State of the Field of Mental Health Apps. *Cognitive and*

- Behavioral Practice*, 25(4). doi: 10.1016/j.cbpra.2018.01.002
- Nechushtai, E., & Lewis, S. C. (2019, 1). What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior*, 90, 298–307. doi: 10.1016/j.chb.2018.07.043
- Ng, M. M., Firth, J., Minen, M., & Torous, J. (2019, 7). User Engagement in Mental Health Apps: A Review of Measurement, Reporting, and Validity. *Psychiatric Services*, 70(7), 538–544. doi: 10.1176/appi.ps.201800519
- Niles, A. N., & O'Donovan, A. (2018, 12). Personalizing Affective Stimuli Using a Recommender Algorithm: An Example with Threatening Words for Trauma Exposed Populations. *Cognitive Therapy and Research*, 42(6), 747–757. doi: 10.1007/s10608-018-9923-8
- Niles, A. N., Woolley, J. D., Tripp, P., Pesquita, A., Vinogradov, S., Neylan, T. C., & O'Donovan, A. (2020). Randomized Controlled Trial Testing Mobile-Based Attention-Bias Modification for Posttraumatic Stress Using Personalized Word Stimuli. *Clinical Psychological Science*, 8(4). doi: 10.1177/2167702620902119
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., . . . Vazire, S. (2022, 1). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, 73(1), 719–748. doi: 10.1146/annurev-psych-020821-114157
- Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdil, W., Vidal, M., . . . Staab, S. (2020, 5). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3). doi: 10.1002/widm.1356
- Oberauer, K., & Lewandowsky, S. (2019). *Addressing the theory crisis in psychology* (Vol. 26) (No. 5). doi: 10.3758/s13423-019-01645-2
- Operario, D., Adler, N. E., & Williams, D. R. (2004, 4). Subjective social status: reliability and predictive utility for global health. *Psychology & Health*, 19(2), 237–246. doi: 10.1080/08870440310001638098
- Osborn, T. G., Li, S., Saunders, R., & Fonagy, P. (2022, 12). University students' use of mental health services: a systematic review and meta-analysis. *International Journal of Mental Health Systems*, 16(1), 57. doi: 10.1186/s13033-022-00569-0
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . Moher, D. (2021, 3). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. doi: 10.1136/bmj.n71
- Pashazadeh Kan, F., Raoofi, S., Rafiei, S., Khani, S., Hosseinifard, H., Tajik, F., . . . Ghashghae, A. (2021, 10). A systematic review of the prevalence of anxiety among the general population during the COVID-19 pandemic. *Journal of Affective Disorders*, 293, 391–398. doi: 10.1016/j.jad.2021.06.073

- Patel, B., Desai, P., & Panchal, U. (2018). Methods of recommender system: A review. In *Proceedings of 2017 international conference on innovations in information, embedded and communication systems, iciiecs 2017* (Vol. 2018-January). doi: 10.1109/ICIIECS.2017.8275856
- Pincay, J., Teran, L., & Portmann, E. (2019). Health recommender systems: A state-of-the-art review. In *2019 6th international conference on edemocracy and egovernment, icedeg 2019*. doi: 10.1109/ICEDEG.2019.8734362
- Plummer, F., Manea, L., Trepel, D., & McMillan, D. (2016, 3). Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. *General Hospital Psychiatry, 39*, 24–31. doi: 10.1016/j.genhosppsych.2015.11.005
- Portugal, I., Alencar, P., & Cowan, D. (2018). *The use of machine learning algorithms in recommender systems: A systematic review* (Vol. 97). doi: 10.1016/j.eswa.2017.12.020
- Prasad, R., & Kumari, V. V. (2012). A Categorical Review of Recommender Systems. *International Journal of Distributed and Parallel systems, 3*(5). doi: 10.5121/ijdps.2012.3507
- Punton, G., Dodd, A. L., & McNeill, A. (2022, 3). ‘You’re on the waiting list’: An interpretive phenomenological analysis of young adults’ experiences of waiting lists within mental health services in the UK. *PLOS ONE, 17*(3), e0265542. doi: 10.1371/journal.pone.0265542
- Qin, X., Wang, S., & Hsieh, C.-R. (2018, 10). The prevalence of depression and depressive symptoms among adults in China: Estimation based on a National Household Survey. *China Economic Review, 51*, 271–282. doi: 10.1016/j.chieco.2016.04.001
- Rendle, S., Zhang, L., & Koren, Y. (2019, 5). On the Difficulty of Evaluating Baselines: A Study on Recommender Systems.
- Ricci, F., Rokach, L., & Shapira, B. (2022). *Recommender Systems Handbook* (F. Ricci, L. Rokach, & B. Shapira, Eds.). New York, NY: Springer US. doi: 10.1007/978-1-0716-2197-4
- Rich, E. (1979, 10). User modeling via stereotypes. *Cognitive Science, 3*(4), 329–354. doi: 10.1016/S0364-0213(79)80012-9
- Rodriguez-Paras, C., Tippey, K., Brown, E., Sasangohar, F., Creech, S., Kum, H.-C., . . . Benzer, J. K. (2017, 10). Posttraumatic Stress Disorder and Mobile Health: App Investigation and Scoping Literature Review. *JMIR mHealth and uHealth, 5*(10), e156. doi: 10.2196/mhealth.7318
- Rohani, D. A., Quemada Lopategui, A., Tuxen, N., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2020, 4). MUBS: A Personalized Recommender System for Behavioral Activation in Mental Health. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–13). New York, NY, USA: ACM. doi: 10.1145/3313831.3376879
- Rohani, D. A., Springer, A., Hollis, V., Bardram, J. E., & Whittaker, S. (2021). Recommending

- Activities for Mental Health and Well-Being: Insights from Two User Studies. *IEEE Transactions on Emerging Topics in Computing*, 9(3). doi: 10.1109/TETC.2020.2972007
- Rosendahl, D. M. M. S. R. R. J. E. I. K. K. C. A. I. . E. O., H. (2022). *Danskernes sundhed: Den nationale sundhedsprofil 2021* (Tech. Rep.).
- Rosenstock, I. M., Strecher, V. J., & Becker, M. H. (1988, 6). Social Learning Theory and the Health Belief Model. *Health Education Quarterly*, 15(2), 175–183. doi: 10.1177/109019818801500203
- Rossetti, M., Stella, F., & Zanker, M. (2016). Contrasting offline and online results when evaluating recommendation algorithms. In *Recsys 2016 - proceedings of the 10th acm conference on recommender systems*. doi: 10.1145/2959100.2959176
- Ruggeri, K., Garcia-Garzon, E., Maguire, , Matz, S., & Huppert, F. A. (2020, 12). Well-being is more than happiness and life satisfaction: a multidimensional analysis of 21 countries. *Health and Quality of Life Outcomes*, 18(1), 192. doi: 10.1186/s12955-020-01423-y
- Ruiz De villa, A., Sottocornola, G., Coba, L., Lucchesi, F., & Skorulski, B. (2024, 6). Ranking the causal impact of recommendations under collider bias in k-spots recommender systems. *ACM Transactions on Recommender Systems*, 2(2), 1–29. doi: 10.1145/3643139
- Ruiz de Villa, A., Sottocornola, G., Coba, L., Maffei, G., Lucchesi, F., Guerreiro, J., & Skorulski, B. (2023, 6). Leveraging Causal Inference to Measure the Impact of a Mental Health App on Users’ Well-being. In *Proceedings of the 31st acm conference on user modeling, adaptation and personalization* (pp. 228–237). New York, NY, USA: ACM. doi: 10.1145/3565472.3592967
- Schlosser, D. A., Campellone, T. R., Truong, B., Anguera, J. A., Vergani, S., Vinogradov, S., & Arean, P. (2017, 6). The feasibility, acceptability, and outcomes of PRIME-D: A novel mobile intervention treatment for depression. *Depression and Anxiety*, 34(6), 546–554. doi: 10.1002/da.22624
- Schlosser, D. A., Campellone, T. R., Truong, B., Etter, K., Vergani, S., Komaiko, K., & Vinogradov, S. (2018, 8). Efficacy of PRIME, a Mobile App Intervention Designed to Improve Motivation in Young People With Schizophrenia. *Schizophrenia Bulletin*, 44(5), 1010–1020. doi: 10.1093/schbul/sby078
- Schneider, L., Matic, A., Buda, T. S., & Dolan, P. (2024, 5). Me, my thoughts and I – Personality as a moderator of the effect of thoughts on subjective well-being. *Personality and Individual Differences*, 222, 112584. doi: 10.1016/j.paid.2024.112584
- Schulze, L., Henwood, A., Matic, A., Skorulski, B., Schneider, L., Dix, S., . . . Dolan, P. (2024, 6). Efficacy of the “Foundations” Smartphone Application in Improving Mental Well-Being in Students: A Randomized Controlled Trial. *Journal of Technology in Behavioral Science*. doi: 10.1007/s41347-024-00419-5
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, 1). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the conference on*

- fairness, accountability, and transparency* (pp. 59–68). New York, NY, USA: ACM. doi: 10.1145/3287560.3287598
- Sigg, S., Lagerspetz, E., Peltonen, E., Nurmi, P., & Tarkoma, S. (2016, 11). Sovereignty of the Apps: There's more to Relevance than Downloads.
- Sirgy, M. J. (2021). Effects of Personality on Wellbeing. In (pp. 207–221). doi: 10.1007/978-3-030-71888-6_9
- Sischka, P. E., Costa, A. P., Steffgen, G., & Schmidt, A. F. (2020, 12). The WHO-5 well-being index – validation based on item response theory and the analysis of measurement invariance across 35 countries. *Journal of Affective Disorders Reports, 1*, 100020. doi: 10.1016/j.jadr.2020.100020
- Solhaug, I., de Vibe, M., Friberg, O., Sørli, T., Tyssen, R., Bjørndal, A., & Rosenvinge, J. H. (2019, 8). Long-term Mental Health Effects of Mindfulness Training: a 4-Year Follow-up Study. *Mindfulness, 10*(8), 1661–1672. doi: 10.1007/s12671-019-01100-2
- Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006, 5). A Brief Measure for Assessing Generalized Anxiety Disorder. *Archives of Internal Medicine, 166*(10), 1092. doi: 10.1001/archinte.166.10.1092
- Stallman, H. M. (2010, 12). Psychological distress in university students: A comparison with general population data. *Australian Psychologist, 45*(4), 249–257. doi: 10.1080/00050067.2010.482109
- Stewart-Brown, S. (2015, 8). Measuring wellbeing: What does the Warwick-Edinburgh Mental Well-being Scale have to offer integrated care? *European Journal of Integrative Medicine, 7*(4), 384–388. doi: 10.1016/j.eujim.2014.08.004
- Stillman, P. E., & Woolley, K. (2023, 9). Undermining Desire: Reducing Unhealthy Choices by Highlighting Short-Term (vs. Long-Term) Costs. *Journal of Consumer Research, 50*(3), 554–575. doi: 10.1093/jcr/ucad004
- Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015, 4). Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science, 16*(3), 475–485. doi: 10.1007/s11121-014-0513-z
- Substance Abuse and Mental Health Services Administration. (2020). *Key substance use and mental health indicators in the United States: results from the 2019 National Survey on Drug Use and Health* (Tech. Rep.). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration.
- Sun, Y., & Zhang, Y. (2021, 3). A review of theories and models applied in studies of social media addiction and implications for future research. *Addictive Behaviors, 114*, 106699. doi: 10.1016/j.addbeh.2020.106699
- Suzuki, K., Siriaraya, P., She, W. J., Tanaka, R., Li, D., & Nakajima, S. (2021). HappyRec: Evaluation of a "happy Spot" Recommendation System Aimed at Improving Mental Well-Being. In *Ieee international conference on data mining workshops, icdmw* (Vol. 2021-

- December). doi: 10.1109/ICDMW53433.2021.00116
- Taylor Browne Lūka, C., Hendry, K., Dutriaux, L., Stevenson, J. L., & Barsalou, L. W. (2024, 7). Developing and Evaluating a Situated Assessment Instrument for Trichotillomania: The SAM ² TAI. *Assessment*. doi: 10.1177/10731911241262140
- Teh, H. C., Archer, J. A., Chang, W., & Chen, S. A. (2015, 2). Mental Well-Being Mediates the Relationship between Perceived Stress and Perceived Health. *Stress and Health, 31*(1), 71–77. doi: 10.1002/smi.2510
- Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., ... Stewart-Brown, S. (2007, 12). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health and Quality of Life Outcomes, 5*(1), 63. doi: 10.1186/1477-7525-5-63
- The WHO World Mental Health Survey Consortium. (2004, 6). Prevalence, Severity, and Unmet Need for Treatment of Mental Disorders in the World Health Organization World Mental Health Surveys. *JAMA, 291*(21), 2581. doi: 10.1001/jama.291.21.2581
- Timulak, L., Keogh, D., Chigwedere, C., Wilson, C., Ward, F., Hevey, D., ... Mahon, S. (2022, 3). A comparison of emotion-focused therapy and cognitive-behavioral therapy in the treatment of generalized anxiety disorder: Results of a feasibility randomized controlled trial. *Psychotherapy, 59*(1), 84–95. doi: 10.1037/pst0000427
- Tønning, M. L., Faurholt-Jepsen, M., Frost, M., Martiny, K., Tuxen, N., Rosenberg, N., ... Kessing, L. V. (2021). The effect of smartphone-based monitoring and treatment on the rate and duration of psychiatric readmission in patients with unipolar depressive disorder: The RADMIS randomized controlled trial. *Journal of Affective Disorders, 282*. doi: 10.1016/j.jad.2020.12.141
- Tønning, M. L., Kessing, L. V., Bardram, J. E., & Faurholt-Jepsen, M. (2019). *Methodological challenges in randomized controlled trials on smartphone-based treatment in psychiatry: Systematic review* (Vol. 21) (No. 10). doi: 10.2196/15362
- Topp, C. W., Østergaard, S. D., Søndergaard, S., & Bech, P. (2015). The WHO-5 Well-Being Index: A Systematic Review of the Literature. *Psychotherapy and Psychosomatics, 84*(3), 167–176. doi: 10.1159/000376585
- Torkamaan, H., & Ziegler, J. (2022, 3). Recommendations as Challenges: Estimating Required Effort and User Ability for Health Behavior Change Recommendations. In *27th international conference on intelligent user interfaces* (pp. 106–119). New York, NY, USA: ACM. doi: 10.1145/3490099.3511118
- Torous, L. J. C. S. R., J. (2018). Mental health apps: what to tell patients. *Current Psychiatry, 17*(3), 21–25.
- Tran, T. N. T., Felfernig, A., Trattner, C., & Holzinger, A. (2021). Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems, 57*(1). doi: 10.1007/s10844-020-00633-6

- Twomey, C., O'Reilly, G., & Byrne, M. (2015, 2). Effectiveness of cognitive behavioural therapy for anxiety and depression in primary care: a meta-analysis. *Family Practice*, 32(1), 3–15. doi: 10.1093/fampra/cmu060
- Valentine, L., D'Alfonso, S., & Lederman, R. (2023, 8). Recommender systems for mental health apps: advantages and ethical challenges. *AI & SOCIETY*, 38(4), 1627–1638. doi: 10.1007/s00146-021-01322-w
- van Agteren, J., Iasiello, M., Lo, L., Bartholomaeus, J., Kopsaftis, Z., Carey, M., & Kyrios, M. (2021, 4). A systematic review and meta-analysis of psychological interventions to improve mental wellbeing. *Nature Human Behaviour*, 5(5), 631–652. doi: 10.1038/s41562-021-01093-w
- van Breda, W., Hoogendoorn, M., Eiben, A., Andersson, G., Riper, H., Ruwaard, J., & Vernmark, K. (2016, 12). A feature representation learning method for temporal datasets. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). IEEE. doi: 10.1109/SSCI.2016.7849890
- Vultureanu-Albisi, A., & Badica, C. (2021, 8). Recommender Systems: An Explainable AI Perspective. In *2021 International Conference on Innovations in Intelligent Systems and Applications (INISTA)* (pp. 1–6). IEEE. doi: 10.1109/INISTA52262.2021.9548125
- Wahle, F., Kowatsch, T., Fleisch, E., Rufer, M., & Weidt, S. (2016, 9). Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild. *JMIR mHealth and uHealth*, 4(3), e111. doi: 10.2196/mhealth.5960
- Wampold, B. E. (2019, 3). A smorgasbord of PTSD treatments: What does this say about integration? *Journal of Psychotherapy Integration*, 29(1), 65–71. doi: 10.1037/int0000137
- Wang, K., Varma, D. S., & Prospero, M. (2018). *A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders* (Vol. 107). doi: 10.1016/j.jpsychires.2018.10.006
- Weingarden, H., Matic, A., Calleja, R. G., Greenberg, J. L., Harrison, O., & Wilhelm, S. (2020, 6). Optimizing Smartphone-Delivered Cognitive Behavioral Therapy for Body Dysmorphic Disorder Using Passive Smartphone Data: Initial Insights From an Open Pilot Trial. *JMIR mHealth and uHealth*, 8(6), e16350. doi: 10.2196/16350
- World Health Organization. (2022). *World mental health report: Transforming mental health for all*. World Health Organization.
- Xu, R., Frey, R. M., Fleisch, E., & Ilic, A. (2016, 9). Understanding the impact of personality traits on mobile app adoption – Insights from a large-scale field study. *Computers in Human Behavior*, 62, 244–256. doi: 10.1016/j.chb.2016.04.011
- Yamaguchi, S., Ling, H., Kim, K., & Mino, Y. (2014, 7). Stigmatisation towards people with mental health problems in secondary school students: an international cross-sectional study between three cities in Japan, China and South-Korea. *International Journal of Culture and Mental Health*, 7(3), 273–283. doi: 10.1080/17542863.2013.786108

- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45. doi: 10.1017/S0140525X20001685
- Yu, B., & Hu, X. (2019, 11). Toward Training and Assessing Reproducible Data Analysis in Data Science Education. *Data Intelligence*, 1(4), 381–392. doi: 10.1162/dint_a_00053
- Zollars, I., Poirier, T. I., & Palden, J. (2019, 10). Effects of mindfulness meditation on mindfulness, mental well-being, and perceived stress. *Currents in Pharmacy Teaching and Learning*, 11(10), 1022–1028. doi: 10.1016/j.cptl.2019.06.005
- Zotova, O., & Karapetyan, L. (2018, 11). Psychological Wellbeing and Personality Mental Health. *KnE Life Sciences*. doi: 10.18502/cls.v4i8.3353

Appendix Chapter 3

Table 1: Model predicting effort ratings including all variables (adjusted $R^2 = 0.049$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
WHO-5 screener scores	0.045	0.250	0.178	0.859
PSS-10 screener scores	0.351	0.255	1.376	0.175
Education component	-0.221	0.237	-0.933	0.355
Income and social status component	0.116	0.254	0.457	0.650
Age	-0.066	0.256	-0.259	0.797
Gender	0.693	0.494	1.403	0.166
Group condition	0.838	0.475	1.764	0.083

Table 2: Model predicting usefulness and satisfaction component including all variables (adjusted $R^2 = 0.028$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
WHO-5 screener scores	0.017	0.099	0.168	0.867
PSS-10 screener scores	0.185	0.101	1.838	0.071
Education component	0.078	0.093	0.837	0.406
Income and social status component	0.024	0.100	0.238	0.813
Age	0.025	0.101	0.248	0.805
Gender	0.138	0.194	0.708	0.482
Group condition	0.316	0.187	1.689	0.097

Table 3: Model predicting activity completions including all variables

Independent variable	Coefficient	Std. Error	z value	p value
Mean effort scores	0.205	0.052	3.926	< 0.001
Mean usefulness and satisfaction scores	-0.091	0.054	-1.702	0.089
WHO-5 screener scores	0.017	0.054	0.319	0.750
PSS-10 screener scores	-0.131	0.056	-2.347	0.019
Education component	-0.101	0.055	-1.822	0.068
Income and social status component	0.004	0.058	0.071	0.943
Age	0.166	0.054	3.090	0.002
Gender	0.010	0.106	0.093	0.926
Group condition	0.701	0.120	5.848	< 0.001

Table 4: Model predicting WHO-5 change including all variables (adjusted $R^2 = 0.173$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Mean effort scores	-1.009	0.421	-2.399	0.020
Mean usefulness and satisfaction scores	0.608	0.397	1.533	0.131
Number of completions	0.384	0.459	0.835	0.407
WHO-5 screener scores	-0.803	0.399	-2.010	0.051
PSS-10 screener scores	-0.584	0.436	-1.339	0.186
Education component	-0.170	0.387	-0.438	0.663
Income and social status component	-0.533	0.405	-1.314	0.195
Age	-0.474	0.420	-1.129	0.264
Gender	-1.383	0.806	-1.716	0.092
Group condition	-0.362	0.881	-0.411	0.683

Table 5: Model predicting PSS-10 change including all variables (adjusted $R^2 = 0.211$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
Mean effort scores	0.335	0.461	0.727	0.470
Mean usefulness and satisfaction scores	0.005	0.434	0.012	0.990
Number of completions	0.109	0.503	0.217	0.829
WHO-5 screener scores	-0.984	0.438	-2.248	0.029
PSS-10 screener scores	-1.846	0.477	-3.871	< 0.001
Education component	-0.258	0.423	-0.609	0.545
Income and social status component	-0.308	0.444	-0.693	0.491
Age	0.012	0.460	0.026	0.979
Gender	-0.695	0.882	-0.788	0.434
Group condition	0.801	0.964	0.831	0.410

Appendix Chapter 5

Table 6: Models predicting mean post-activity ratings including all variables

a) Predictors of mean effort ratings (adjusted $R^2 = -0.043$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
High mental wellbeing	0.240	0.528	0.454	0.651
Moderate stress	-0.112	0.503	-0.223	0.824
High stress	-0.204	0.940	-0.217	0.829
Honesty	-0.098	0.257	-0.380	0.705
Emotionality	0.221	0.238	0.927	0.356
Extraversion	0.199	0.248	0.803	0.424
Agreeableness	0.221	0.224	0.989	0.325
Conscientiousness	0.027	0.234	0.116	0.908
Openness	-0.274	0.217	-1.267	0.208
Activity bucket	-0.364	0.451	-0.807	0.422
Gender	0.106	0.466	0.228	0.820
Age	0.037	0.214	0.175	0.861
Middle <i>income and social status</i> bracket	-0.500	0.496	-1.009	0.316
Upper <i>income and social status</i> bracket	-1.222	0.523	-2.335	0.022
Middle <i>education</i> bracket	0.325	0.558	0.582	0.562
Upper <i>education</i> bracket	0.186	0.492	0.379	0.706

b) Predictors of mean *Positive emotional experience and satisfaction* component (adjusted $R^2 = 0.241$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
High mental wellbeing	-0.134	0.141	-0.957	0.341
Moderate stress	0.108	0.134	0.807	0.422
High stress	-0.089	0.250	-0.356	0.723
Honesty	-0.285	0.068	-4.158	< 0.001
Emotionality	-0.005	0.064	-0.080	0.936
Extraversion	0.003	0.066	0.051	0.960
Agreeableness	0.194	0.060	3.246	0.002
Conscientiousness	0.235	0.062	3.774	< 0.001
Openness	-0.062	0.058	-1.072	0.286
Activity bucket	0.352	0.120	2.922	0.004
Gender	-0.093	0.124	-0.750	0.455
Age	0.011	0.057	0.202	0.841
Middle <i>income and social status</i> bracket	0.163	0.133	1.229	0.222
Upper <i>income and social status</i> bracket	0.051	0.143	0.357	0.722
Middle <i>education</i> bracket	0.090	0.149	0.605	0.547
Upper <i>education</i> bracket	-0.020	0.131	-0.154	0.878
Mean effort rating	-0.129	0.053	-2.415	0.018

c) Predictors of mean *Stress level* component (adjusted $R^2 = 0.022$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
High mental wellbeing	-0.020	0.146	-0.139	0.890
Moderate stress	-0.077	0.139	-0.555	0.580
High stress	-0.096	0.259	-0.368	0.714
Honesty	-0.152	0.071	-2.145	0.035
Emotionality	0.046	0.066	0.691	0.491
Extraversion	-0.070	0.069	-1.015	0.313
Agreeableness	0.229	0.062	3.696	< 0.001
Conscientiousness	0.078	0.065	1.207	0.230
Openness	0.010	0.060	0.162	0.872
Activity bucket	0.158	0.125	1.268	0.208
Gender	-0.004	0.129	-0.032	0.975
Age	-0.065	0.059	-0.276	0.276
Middle <i>income and social status</i> bracket	0.034	0.138	0.249	0.804
Upper <i>income and social status</i> bracket	-0.096	0.148	-0.650	0.517
Middle <i>education</i> bracket	-0.044	0.154	-0.283	0.778
Upper <i>education</i> bracket	-0.010	0.136	-0.070	0.944
Mean effort rating	-0.120	0.055	-2.181	0.032

d) Predictors of mean *Coping and social connection* component (adjusted $R^2 = 0.083$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
High mental wellbeing	0.054	0.190	0.286	0.775
Moderate stress	0.150	0.181	0.831	0.408
High stress	0.080	0.338	0.238	0.812
Honesty	-0.201	0.092	-2.175	0.032
Emotionality	-0.116	0.086	-1.346	0.181
Extraversion	-0.002	0.089	-0.024	0.981
Agreeableness	0.167	0.081	2.066	0.042
Conscientiousness	0.109	0.084	1.296	0.198
Openness	0.005	0.078	0.069	0.945
Activity bucket	0.197	0.163	1.213	0.228
Gender	0.012	0.168	0.073	0.942
Age	0.028	0.077	0.367	0.714
Middle <i>income and social status</i> bracket	0.076	0.179	0.426	0.671
Upper <i>income and social status</i> bracket	-0.039	0.193	-0.203	0.839
Middle <i>education</i> bracket	0.079	0.201	0.395	0.694
Upper <i>education</i> bracket	0.106	0.177	0.597	0.552
Mean effort rating	0.092	0.072	1.276	0.205

Table 7: Model predicting activity completions including all variables

Independent variable	Coefficient	Std. Error	z value	p value
High mental wellbeing	1.580	0.670	2.358	0.018
Moderate stress	0.200	0.718	0.278	0.781
High stress	0.667	1.202	0.555	0.579
Mean effort rating	-0.013	0.291	-0.045	0.964
Mean <i>Positive emotional experience and satisfaction</i>	0.834	0.323	2.582	0.010
Mean <i>Stress level</i>	0.325	0.287	1.131	0.258
Mean <i>Coping and social connection</i>	0.246	0.306	0.803	0.422
Honesty	0.428	0.384	1.114	0.265
Emotionality	-0.209	0.312	-0.672	0.502
Extraversion	-0.175	0.334	-0.522	0.601
Agreeableness	-0.397	0.334	-1.189	0.234
Conscientiousness	-0.318	0.341	-0.931	0.352
Openness	0.388	0.284	1.366	0.172
Gender	-0.481	0.652	-0.737	0.461
Age	0.539	0.303	1.777	0.076
Middle <i>income and social status</i> bracket	0.336	0.671	0.501	0.617
Upper <i>income and social status</i> bracket	-0.231	0.675	-0.342	0.732
Middle <i>education</i> bracket	-0.989	0.729	-1.356	0.175
Upper <i>education</i> bracket	-0.182	0.665	-0.273	0.785

Table 8: Models predicting mean post-activity ratings including all variables

a) Predictors of mean WHO-5 change (adjusted $R^2 = 0.080$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
High mental wellbeing	-6.389	1.596	-4.004	< 0.001
Moderate stress	-0.797	1.522	-0.524	0.602
High stress	0.298	2.828	0.105	0.916
Mean effort rating	0.664	0.641	1.036	0.303
Mean <i>Positive emotional experience and satisfaction</i>	0.863	0.719	1.201	0.233
Mean <i>Stress level</i>	-0.050	0.646	-0.078	0.938
Mean <i>Coping and social connection</i>	-0.407	0.623	-0.653	0.516
Honesty	-0.613	0.863	-0.710	0.479
Emotionality	0.401	0.727	0.552	0.582
Extraversion	-0.946	0.751	-1.259	0.211
Agreeableness	1.248	0.761	1.640	0.104
Conscientiousness	1.296	0.760	1.705	0.092
Openness	0.758	0.660	1.149	0.254
Activity bucket	0.925	1.432	0.646	0.520
Gender	-0.747	1.405	-0.532	0.596
Age	-0.678	0.647	-1.048	0.297
Middle <i>income and social status</i> bracket	1.566	1.509	1.037	0.302
Upper <i>income and social status</i> bracket	2.019	1.620	1.247	0.216
Middle <i>education</i> bracket	0.561	1.685	0.333	0.740
Upper <i>education</i> bracket	-0.966	1.482	-0.652	0.516

b) Predictors of mean *Coping and social connection* component (adjusted $R^2 = 0.083$)

Independent variable	Coefficient	Std. Error	t-statistic	p value
High mental wellbeing	0.545	1.335	0.408	0.684
Moderate stress	-3.041	1.273	-2.389	0.019
High stress	-6.773	2.365	-2.863	0.005
Mean effort rating	-0.836	0.536	-1.559	0.122
Mean <i>Positive emotional experience and satisfaction</i>	-0.826	0.601	-1.374	0.173
Mean <i>Stress level</i>	-0.281	0.541	-0.520	0.604
Mean <i>Coping and social connection</i>	-0.083	0.521	-0.160	0.873
Honesty	0.370	0.722	0.512	0.610
Emotionality	0.629	0.608	1.034	0.304
Extraversion	0.804	0.628	1.280	0.204
Agreeableness	-0.680	0.636	-1.068	0.288
Conscientiousness	-0.731	0.636	-1.149	0.253
Openness	-0.274	0.552	-0.497	0.620
Activity bucket	-0.183	1.198	-0.153	0.879
Gender	1.087	1.175	0.925	0.357
Age	0.051	0.541	0.095	0.925
Middle <i>income and social status</i> bracket	-0.548	1.262	-0.434	0.665
Upper <i>income and social status</i> bracket	-1.555	1.355	-1.147	0.254
Middle <i>education</i> bracket	-0.133	1.409	-0.095	0.925
Upper <i>education</i> bracket	0.481	1.240	0.388	0.699