Communication of Health Messages: A Personalised Chatbot for Weight Loss Maintenance

William Samuel Holmes

Faculty of Arts, Humanities and Social Sciences at the Ulster University

Submitted for a PhD in May 2020

I confirm that the word count of this thesis is less than 100,000 words

Table of Contents

List of Tables

List of Figures

Acknowledgements

This PhD would not have been possible without the assistance and support of the following:

Dr Anne Moorhead, Faculty of Arts, Humanities and Social Sciences
Dr Raymond Bond, Faculty of Computing, Engineering and the Built Environment
Professor Huiru Zheng, Faculty of Computing, Engineering and the Built Environment
Professor Vivien Coates, Faculty of Life & Health Sciences
Professor Michael McTear, Faculty of Computing, Engineering and the Built Environment

I would like to thank my parents, William and Claire Holmes for their willingness to proof-read and for their relentless encouragement.

I would like to thank my wife Céline Holmes for her relentless encouragement, for believing in me, and for persuading me not to give up.

Abstract

Background

Weight loss and maintenance are improved by effective communication. Technology-based interventions may improve short term behaviour change however many have used older technology such as text-messaging. The long-term benefits of technology are unknown, as is the effectiveness of chatbots (conversation-driven user interfaces) for delivering weight loss maintenance interventions. *WeightMentor* is a personalised motivational chatbot for weight loss maintenance.

Aim

The aim of this PhD was to design, develop, and test the usability of *WeightMentor*, a personalised motivational chatbot to assist individuals with weight loss maintenance.

Methods

This PhD was a mixed methods technology design and development project, comprising five individual studies: a review of literature relating to technology for weight loss maintenance, a needs analysis of the *WeightMentor* target users, designing and developing the *WeightMentor* chatbot, testing the usability of the *WeightMentor* chatbot, and validating the Chatbot Usability Questionnaire (CUQ), designed for measuring chatbot usability.

Key Results

The needs analysis identified five key themes: “Weight loss maintenance is a challenge”, “Social contact can be a double-edged sword”, “Apps are popular”, “Personalised messages are more useful” and “Chatbots have potential for weight loss maintenance”. Participants considered the *WeightMentor* prototype to be engaging and simple. *WeightMentor* was designed using contemporary technologies to implement features that would be most useful for target users. *WeightMentor* scored well above conventional usability benchmarks. The minimum number of users required to identify most usability issues was 39, more than conventional wisdom suggests (n=5-8). During validation the CUQ demonstrated construct validity and test-retest reliability.

Discussion/Conclusion

Weight loss maintenance is a challenge that may be supported using technology, and the newly developed chatbot, *WeightMentor* has the potential for weight loss maintenance as it is usability acceptable and is a solution that should be explored further.

Chapter 1: Introduction

1.1 Background

Obesity is a global public health issue (World Health Organisation (WHO) 2018). It may be challenging for individuals to lose weight, but it may also be a greater challenge for them to maintain weight loss. There have been many strategies and approaches to address obesity and one such approach is the use of digital technologies. There are numerous commercial technologies available in the area of obesity management and weight loss maintenance such as mobile applications (Brindal et al. 2016), text-messaging-based systems (Donaldson et al. 2014), websites (Cussler et al. 2008), and e-mail-based systems (Thomas et al. 2012). However, there are limited evidence-based resources especially those which use contemporary technological advances for weight loss maintenance. This PhD considered the research area of digital technologies for weight loss maintenance, specifically chatbots.

1.1.1 Obesity

Globally, obesity is now considered to have reached epidemic status and as such is considered a global challenge (World Health Organisation (WHO), 2018). Overweight is defined as a Body Mass Index (BMI) greater than 25 kg/m^2 (National Health Service (NHS), 2019), while obesity is defined as a BMI greater than 30 kg/m^2 (NHS, 2019). Overweight and obesity are recognised as significant risk factors for several noncommunicable diseases such as type 2 diabetes, stroke, cancer, and cardiovascular diseases (WHO, 2018). On a global scale, more people die because of obesity and overweight than because of underweight (BMI less than 18.5 kg/m^2).

Statistics from the European Union suggest that in 2014 51.6% of the EU population were overweight (Eurostat, 2019). In the United Kingdom, two thirds of the population are reported to be either overweight or obese, and one third of these are obese (Sky News, 2019). It is suggested that 29% of adults in England and Scotland are obese, with women and people over the age of 35 more likely to be obese. Obesity has also been observed to be linked with household income (Sky News, 2019), and obesity rates are reported to be higher in parts of the United Kingdom where income is lower. The obesity rate is slightly lower in Northern Ireland than in Scotland and England, the Health Survey for Northern Ireland (2018) reported that 64% of adults in Northern Ireland were either obese (27%)

or overweight (37%), and that 40% of men and 47% of women had proactively tried changing their behaviour in the past three years in order to try and lose weight (Department of Health, Northern Ireland, 2018). It was reported that men were more likely (73%) to be overweight or obese than women (57%). Although obesity and overweight have not significantly increased in the province since 2015, there has been a noticeable increase in obesity since 2005 (Department of Health Northern Ireland, 2017).

Obesity is having an increasing impact on the UK NHS. Sky News report that hospital admissions where obesity was a primary or secondary diagnosis have quadrupled since 2009/2010, and it is suggested that nearly 9% of the NHS annual budget (approximately £8.8 billion per year) is spent on treating Type 2 diabetes, which is linked with weight. The problem is exacerbated by increased sugar consumption and reduced physical activity (Sky News 2019).

1.1.2 Weight loss maintenance

The UK's National Health Service (NHS) suggested that losing a mere 3% of body weight will significantly improve health outcomes (NHS, 2019), but doing so is a challenge for many individuals. It has been suggested that weight loss is more attractive for individuals than weight loss maintenance because the benefits (increased sense of wellbeing etc) outweigh the costs initially but appear less obvious during maintenance, thus individuals may be more reluctant to continue. Difficulties in sustaining positive behaviour change and effectively communicating obesity and weight loss maintenance contribute to the challenges of weight loss and weight loss maintenance (Moorhead et al. 2013, Wing & Hill, 2001). During the *Weight Care Project*, it was revealed that diet and nutrition information is diverse, confusing and sometimes dubious, and while medical professionals could serve as a useful source of accurate and reliable source of information, many consider obesity and weight loss to be a taboo subject and are reluctant to engage in frank discussion with their patients (Moorhead et al. 2013). Findings from recent studies have suggested that regular self-reporting, motivation and personalised feedback help mitigate the challenge of weight loss maintenance and increase the chances of success, at least in the short term (Donaldson et al. 2014, Fjeldsoe et al 2016). Wing & Hill (2001) proposed that a standardised definition of weight loss maintenance increases the chances of success, and thus defined weight loss maintenance as "intentionally losing at least 10% of initial body weight and keeping it off for at least 1 year". For this thesis,

Wing & Hill's (2001) definition has been modified slightly, and weight loss maintenance will be defined as "intentionally losing at least 10% of initial body weight and actively maintaining it".

1.1.3 Technology for weight loss maintenance

There is evidence from previous research to suggest that behaviour change for weight loss may be stimulated using digital technologies such as websites, video conferencing, text messaging, and mobile phone technologies (mobile web or apps), at least in the short term (Donaldson et al. 2014, Svetkey et al. 2008). These technologies have, in a small number of research studies, also proven somewhat effective for weight loss maintenance. It is worth remembering however that technologies such as text messaging, e-mail, and websites are gradually becoming obsolete in the face of newer smartphone-based apps. A limited number of recent trials have explored the use of dedicated smartphone apps for weight loss maintenance (Brindal et al. 2016). Nutrition apps are popular in Google and Apple stores, however many of these have limited functionality and are not always based on scientific research (Brindal et al. 2016, Franco et al. 2016).

The use of text messaging has declined in recent years (Ofcom, 2018). This is due in no small part to the rise in the use of social media tools such as WhatsApp, Facebook and Snapchat, with Facebook being the most popular of these. This suggests that there is a clear need for further research into the use of smartphone-based apps for weight loss maintenance. However, according to US technology firm *Personetics*, smartphone users are experiencing "app fatigue", growing frustration at the abundance of smartphone apps available today (Personetics, 2016), and around 25% of downloaded smartphone apps are discarded after just one use. This suggests that target users could potentially be put off by "yet another weight loss maintenance app".

1.1.4 Chatbots

Chatbots are intelligent systems that present the user with a lightweight, straightforward interface that is based on human conversation rather than a conventional Graphical User Interface (GUI) (Knowledge@Wharton, 2016). US technology firm Personetics argued that chatbots are a simple and cost-effective way to interface with users without forcing them to download and learn another app (Personetics, 2016). The popularity of Facebook and Facebook messenger (UKOM, 2018) means that integrating a chatbot with these

technologies will potentially create extra convenience for the user, in a self-contained system with an informal, user-friendly interface. Facebook Messenger already supports chatbot development and existing users would not need to learn a new app with an unfamiliar interface. Chatbots may additionally solve limitations of text-message based solutions, such as 24/7 availability and automated responses seeming too "generic" (Donaldson et al. 2014). *WeightMentor*, a chatbot for weight loss maintenance, was developed and tested during this PhD.

1.2 Research Question

"What is the usability of a personalised motivational chatbot for weight loss maintenance?"

1.3 Aim

The aim of this PhD was to design, develop, and test the usability of *WeightMentor*, a personalised motivational chatbot to assist individuals with weight loss maintenance.

1.4 Objectives

The aim of this PhD was achieved through the fulfilment of five main objectives:

1. To review the current literature on digital technologies for weight loss maintenance
2. To design a personalised motivational chatbot for weight loss maintenance
3. To develop a personalised motivational chatbot for weight loss maintenance
4. To test the usability of the personalised chatbot for weight loss maintenance
5. To design and validate the Chatbot Usability Questionnaire for measuring usability of chatbots.

1.5 Chapter Overview

1.5.1 Chapter 2: Literature Review

A narrative literature review was completed on digital technologies for weight loss maintenance to provide a comprehensive overview of the literature in this area using all study designs. The purpose of this literature review was to provide a critique on the

research literature on digital technologies for weight loss maintenance. Three main electronic databases (EMBASE, Ovid/Medline and PubMed) were searched for relevant literature published between 2006 and 2019. In total, 63 research studies were included. In addition, seven papers discussing randomised controlled trials (RCTs, 2006-2018) using technology for weight loss maintenance were reviewed systematically and the findings published in the April 2019 issue of the European Journal of Public Health (Holmes et al. 2019). A systematic review approach was chosen as it is recognised as the gold standard of literature reviews. Systematic reviews provide justification for further research (Moher et al. 2009) and provide clinicians and policy makers with strong evidence for making key decisions and developing practice guidelines (Liberati et al. 2009).

1.5.2 Chapter 3: Needs Analysis

The needs analysis phase of the design process sought to identify the needs of the target demographic, who were adults who had lost weight and wished to maintain the weight lost or had lost weight and since regained it. Structured personal interviews were conducted with individuals who had previously lost weight. During the interviews, participants discussed challenges to weight loss maintenance and solutions, social contact, app use and personalised messaging. Participants were shown a demonstration of the *WeightMentor* prototype chatbot and asked to comment on it. Findings from the needs analysis were used, in conjunction with a recent review of popular nutrition apps (Franco et al. 2016), to inform the development of the *WeightMentor* chatbot.

1.5.3 Chapter 4: Chatbot Design & Chatbot Development

Based on the literature and findings from the needs analysis, a chatbot, known as *WeightMentor*, was developed using Facebook Messenger (currently the most popular smartphone application (UKOM, 2018)) to facilitate easy integration with existing systems that will potentially be already familiar to users, without needing to download and learn to use a new smartphone application. The chatbot's main functions were selected to maximise usefulness for users while providing a positive and engaging user experience (UX). Content analysis was used to compare *WeightMentor* with other chatbots of similar role and function, and to determine the readability of messages.

1.5.4 Chapter 5: Usability Testing

Following completion and initial testing, the *WeightMentor* chatbot was usability tested using 50 human participants. Participants were observed performing common tasks with the chatbot and completed usability questionnaires, one of which, the *Chatbot Usability Questionnaire (CUQ)*, was designed specifically for evaluating chatbot usability, as part of this PhD. Usability test participants were asked to identify potential usability issues during the usability tests.

1.5.5 Chapter 6: Validation of the Chatbot Usability Questionnaire

The Chatbot Usability Questionnaire was validated by 26 human participants, assessing three chatbots selected by the PhD researcher and classified by a panel of computer scientists as *good*, *average* and *poor quality*. Participants used each of the three chatbots (without knowing which one belonged to each category) during two separate sessions a couple of weeks apart and completed CUQs for each chatbot after each session.

1.5.6 Chapter 7: Discussion, Future Work & Conclusions

It has been concluded that robust clinical trials of technology for weight loss maintenance are limited in number, use older forms of technology and have only been trialled in the short term. The chatbot, *WeightMentor* was determined to be potentially useful for individuals who are maintaining weight loss, provided it supports users when they are most vulnerable, is convenient and requires minimum effort for interaction and permits progress tracking with subtle, appropriate feedback that is not patronising or bossy. *WeightMentor* will provide users with a potentially engaging and interesting tool for weight loss maintenance that users may interact with without fear of being judged. During usability tests it was determined that *WeightMentor* scored more highly on popular usability metrics than conventional systems, suggesting that it is perceived as easier to use. Users were able to become proficient very quickly in use of the chatbot. Suggested future work includes a detailed content analysis of the *WeightMentor* chatbot in comparison to other chatbots, and a study to determine the feasibility of the chatbot for delivering weight loss maintenance interventions.

1.6 Conclusion

Obesity is a significant issue and one which will not be resolved quickly or easily. Research into technology for weight loss maintenance has shown promising results but is limited. Chatbots may present a viable solution for motivating individuals for weight loss maintenance because they can vary their responses and are available 24 hours a day. This PhD reviewed literature relating to weight loss maintenance technology, evaluated potential user needs, designed and developed the *WeightMentor* chatbot in accordance with identified user needs, tested the usability of the *WeightMentor* chatbot and validated a bespoke Chatbot Usability Questionnaire. The outcome of this PhD was to contribute to the research area of digital technologies in weight loss maintenance by designing, developing and usability testing a chatbot, *WeightMentor*, and developing and evaluating a Chatbot Usability Questionnaire for measuring the usability of chatbots.

Chapter 2: Literature Review

2.1 Introduction

Obesity is now considered to be a global epidemic with significant health implications (World Health Organisation (WHO), 2018), and it is not an issue that may easily be resolved in the short term (National Health Service (NHS), 2019, as discussed in Chapter 1). Although there is an abundance of diet plans that promise rapid weight loss with only minimum effort, the techniques encouraged by these plans often pose significant risks to health. The NHS suggests that it is not safe to lose more than 0.5 - 1.0kg per week, as faster weight loss may lead to issues such as malnutrition or gallstones (NHS, 2016). Rapid weight loss is also difficult to sustain in the long term, and may lead to eventual weight regain, and psychological consequences such as disappointment or loss of self-esteem. In the 21st century, digital technologies such as computers and smartphones have become a fundamental part of human life. In the face of a growing obesity crisis and rising healthcare costs, health agencies are now looking to the use of digital technology for health management and health intervention delivery. A range of digital healthcare technologies exists (Mosa et al. 2012), which may be useful in facilitating motivational communication for obesity management and weight loss maintenance, several of which have been tested in clinical trials, with varying successes reported. This literature review discusses the relevant theories of health behaviour change; examines research with respect to weight loss maintenance; explores the effectiveness of digital technologies for weight loss maintenance and identifies the most effective strategies for communicating on weight loss maintenance. The aim of this review is to critique the research literature on health technology for weight loss and weight loss maintenance in order to identify gaps for further research. The review was undertaken using a systematic and narrative approach to capture all relevant research using a range of study designs.

2.2 Literature Search

The key relevant electronic databases, EMBASE, Medline and Pubmed were used to identify relevant literature. Key search terms used included “digital technologies”, “personalised messaging” and “weight loss maintenance”. Searches were conducted for articles published between 2006 and February 2018. Each initial concept was further expanded to include possible keywords (e.g. Digital Technologies: Cell phones, Smartphones, Mobile Applications, Telemedicine) and variations (e.g. Smartphone,

smart phone, smartphones). From these keywords, a list of search terms was developed (e.g. smartphone* OR smart phone*) and these were combined into the final search strings, which are included in **Appendices 1 to 3**. All papers included were written in English, though foreign language papers were considered for inclusion if an English translation was available. Where a paper was referenced by another paper it was considered for inclusion if it met the relevant criteria. Articles which were included in this literature review considered the use of digital technologies for weight loss maintenance where weight loss maintenance was the primary outcome. Studies in which digital technologies were not used were excluded as the focus of this PhD was a digital technology solution and it was desirable to compare interventions of a similar type. Papers where weight loss maintenance was not the primary outcome were excluded. It was observed that interventions focusing on obesity-related disorders such as hypertension, diabetes etc. often included a weight loss or weight maintenance component, however this was not the primary outcome and it was considered necessary to select only those papers where weight loss maintenance was the primary focus of the intervention. The focus of this PhD was otherwise healthy individuals thus studies involving participants who were pregnant or undergoing medical treatments were all excluded. Smoking cessation articles or those relating to diabetes management were also excluded as these were either not totally related to weight loss maintenance (in the case of smoking cessation) or weight loss maintenance was not the primary outcome (as discussed above). If relevant articles were not easily available online through University access to relevant databases, they were retrieved using the university's document delivery system. The inclusion and relevant exclusion criteria are listed in **Table 2.1**.

Table 2.1: Article Screening Criteria

Inclusion	Exclusion
Focusing on weight loss maintenance	Focusing on disorders such as hypertension, diabetes, irritable bowel disorder, or eating disorders.
	Focusing on drug addiction or smoking cessation (even if obesity mentioned)
	Focusing on healthy lifestyles, diet, or exercise as primary outcomes
	Focusing on cancer prevention
Relating to pregnancy if weight loss maintenance is the primary outcome	Relating to pregnancy related disorders such as gestational diabetes mellitus
Digital health technologies only, e.g. text message-based systems, email, websites	Non-digital technologies e.g. drugs, portion control plates

Inclusion	Exclusion
Other relevant developments e.g. development of technologies, participant feedback, suggested app functionality	
Dated 2006 – February 2018	
Written in English	

A total of 6559 results were returned. These were screened for duplicates and the 3600 remaining results were screened using inclusion and exclusion criteria. Screening based on title abstract and criteria excluded a total of 3461 results. A PRISMA diagram of the selection process for the systematic literature review is presented in **Figure 2.1**. Articles selected for this review are categorised into eight areas, which are summarised in **Table 2.2**. This narrative literature review included sixty-two studies, seven of which were Randomised Controlled Trials (RCTs) included in the systematic literature review (Holmes et al. 2018).

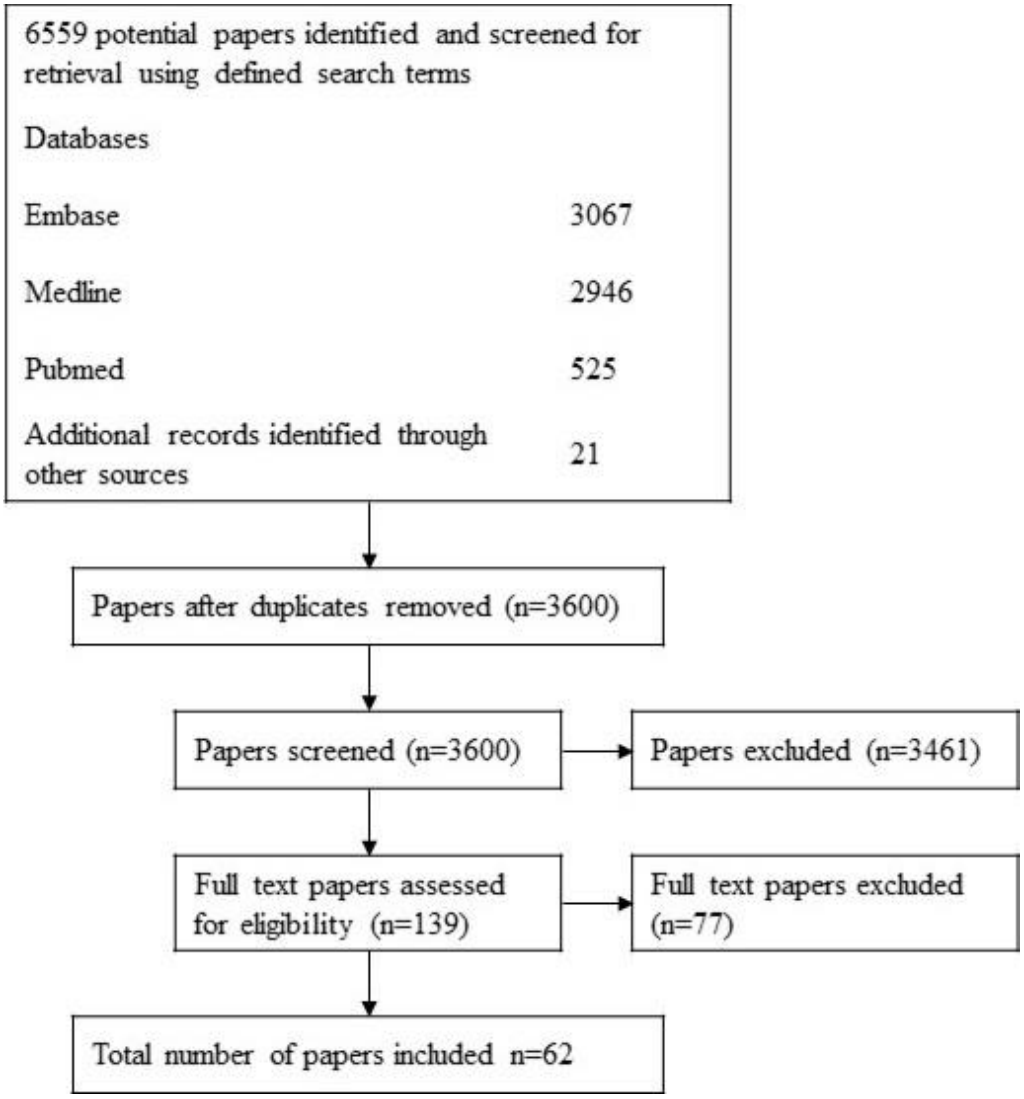


Figure 2.1: PRISMA Diagram

Table 2.2: Selected articles by subject area

Category	n
Mobile phones	27
E-mail	1
Internet	9
Games	3
General behaviour change	1
General technology	4
Tailored messages	12
Comparison of interventions	5
Total n	**62**

2.3 Theories of Behaviour Change

The core principle behind weight loss maintenance is *behaviour change*, which itself is supported by several key theories. Perhaps the most noteworthy of which are Albert Bandura's *Social Cognitive Theory* and *Self-Efficacy Theory*. Social Cognitive Theory, formerly *Social Learning Theory* may be considered a form of *operant conditioning*, which was proposed by psychologist B.F Skinner and based on experiments by Edward Thorndyke at the end of the 19th century. Thorndyke's experiments involved cats trying to escape from a series of boxes. Thorndyke observed that when cats struggled and actively tried to escape from the box, they quickly learnt strategies that led to success (such as activating a lever or switch within the box) (Thorndyke, 1898).

Skinner built on Thorndyke's work by conducting a series of experiments in which a rat in a box was subjected to various stimuli. In one experiment, activation of a lever in the box would dispense a food reward. When initial accidental lever activations resulted in food, the rat learned over time that it could *intentionally* activate the lever to receive a reward. Another experiment subjected the rat to an electric current, causing slight discomfort. Accidental activation of a lever switched off the electric current, thus relieving the discomfort. Eventually, the rat activated the lever as soon as it was placed in the box, to relieve the discomfort as soon as possible (McLeod, 2016). Skinner suggested that any given behaviour may elicit one of three responses: *neutral* responses, which neither encourage nor discourage behaviour, *reinforcers* (either positive or negative), which encourage behaviour and *punishers*, which discourage behaviour, and

that these principles could be used to shape animal behaviour by changing the pattern of rewards and punishments until the desired behaviour is achieved (McLeod, 2016).

Albert Bandura further developed Skinner's theory by suggesting that while humans may primarily learn from our experiences, it is also possible for us to learn by observing the actions of others, and the consequences. In his famous *Bobo doll experiment* (1961), 72 children were organised into 3 groups of 24 children (12 boys and 12 girls) (McLeod, 2014). In the first phase of the experiment, the three groups were exposed to different conditions: the first group were placed in a room with an adult "model" who attacked an inflatable "bobo doll", the second group were placed in a room with an adult "model" who played calmly with a selection of toys, while ignoring the bobo doll, and the third group acted as a control group, thus were not exposed to any adult "model". In phase two of the experiment, all children were exposed to mild "aggression arousal". This was achieved by individually bringing each child into a room with some brightly coloured, attractive toys. After a few minutes of playing with the toys, each child was told by the experimenter that "these were special toys" and that "the experimenter had decided to reserve them for other children" (McLeod, 2014).

In the third and final stage, children were brought into a toy room with "non-aggressive toys" (e.g. a tea set) and "aggressive toys", (e.g. a mallet and a bobo doll). Each child was observed every five seconds for twenty minutes. It was observed that children interacted with the doll based on which model they had been exposed to. Children who had observed the aggressive model were more likely to behave aggressively towards the bobo doll. Bandura concluded that children who observed aggressive adult role models were themselves more likely to behave aggressively (McLeod, 2014). Bandura's *social cognitive theory* built on Skinner's theory of operant conditioning by suggesting that humans actively process information. Rather than mindlessly repeating the actions of others, we carefully consider the implications of the action, and their relevance to us (McLeod, 2016). The principles of social cognitive theory can be applied to Bandura's *Self-Efficacy Theory (1977)*. This theory suggests that self-efficacy comes from four sources. *Mastery experiences* reinforce self-belief through success. *Vicarious experiences* reinforce self-belief through witnessing the successes of people we perceive to be like us. *Verbal persuasion* strengthens belief in ourselves through positive verbal encouragement from people who are close to us. Finally, *Emotional and Physiological States* either boost or dampen self-confidence (Bandura, 1977).

2.3.1 Transtheoretical Model of Health Behaviour Change

The *Transtheoretical Model of Health Behaviour Change* (Prochaska & Velicer, 1997) builds on the work previously discussed by Albert Bandura and proposes seven stages of behaviour change, illustrated in **Figure 2.2**. Individuals at the first stage, *precontemplation*, lack both interest and inclination to effect positive behaviour changes in their lives. Progression through this stage leads to *contemplation* as the individual carefully considers the benefits and challenges of behaviour change. This is consistent with the cognitive element of Bandura's social cognitive theory. During the *preparation* stage, the individual starts actively planning behaviour change, generally within the next month. Preparation leads to *action*, which lasts for at least six months. During this stage, positive, sustainable changes take place, and the benefits start to become visible. These changes will be reinforced during the *maintenance* stage, which should ultimately lead to self-efficacy, described by *Bandura* as the extent to which an individual is able to cope in a specific situation, influenced by personal experience and social learning (*Bandura, 1977*), at which point the individual moves to the *termination* stage. Unsuccessful maintenance may alternatively lead to *relapse*, where the cycle resets.

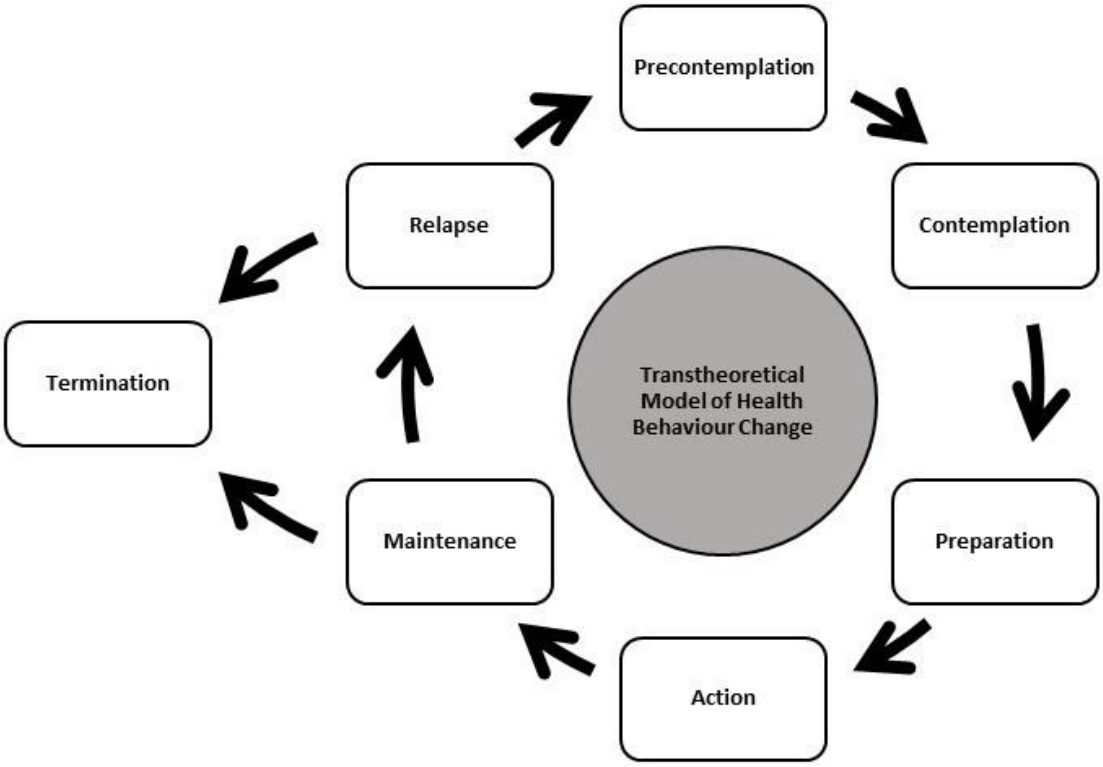


Figure 2.2: Transtheoretical Model of Health Behaviour Change

The *action* stage of the *Transtheoretical Model* is characterised by a sustained period of repeated, intense execution of these ten processes, which continues at reduced intensity in the *maintenance* stage, when strategies learnt during the action stage are reinforced. Successful early stages increase the individual's self-confidence, what Bandura describes as a *mastery experience*. This leads to increased self-belief and improves the ability to maintain change while reducing the likelihood of giving in to temptation. It is suggested that the maintenance stage can last for at least five years, during which time the individual is still at risk of relapse (Prochaska & Velicer, 1997). The transtheoretical model is underpinned by ten processes of behaviour change, which are summarised in **Table 2.3** and illustrated in **Figure 2.3**.

Table 2.3: Processes of Health Behaviour Change

Process	Description
Consciousness raising	Increasing awareness of the cause and impact of behaviour
Dramatic relief	Arousing positive or negative emotions about behaviour
Self-re-evaluation	Re-evaluating self-image to incorporate positive behaviour
Environmental re-evaluation	Awareness of how their behaviour (positive or negative) affects others
Social liberation	Increasing awareness of societal support mechanisms
Self-liberation	Belief that the individual CAN change leads to a commitment TO change
Helping relationships	Building a network of supportive, like-minded individuals who can help encourage change
Counter conditioning	Learning to substitute positive behaviour for negative behaviour
Contingency management	Associating positive and negative behaviours with consequences (emphasising rewards rather than punishments)

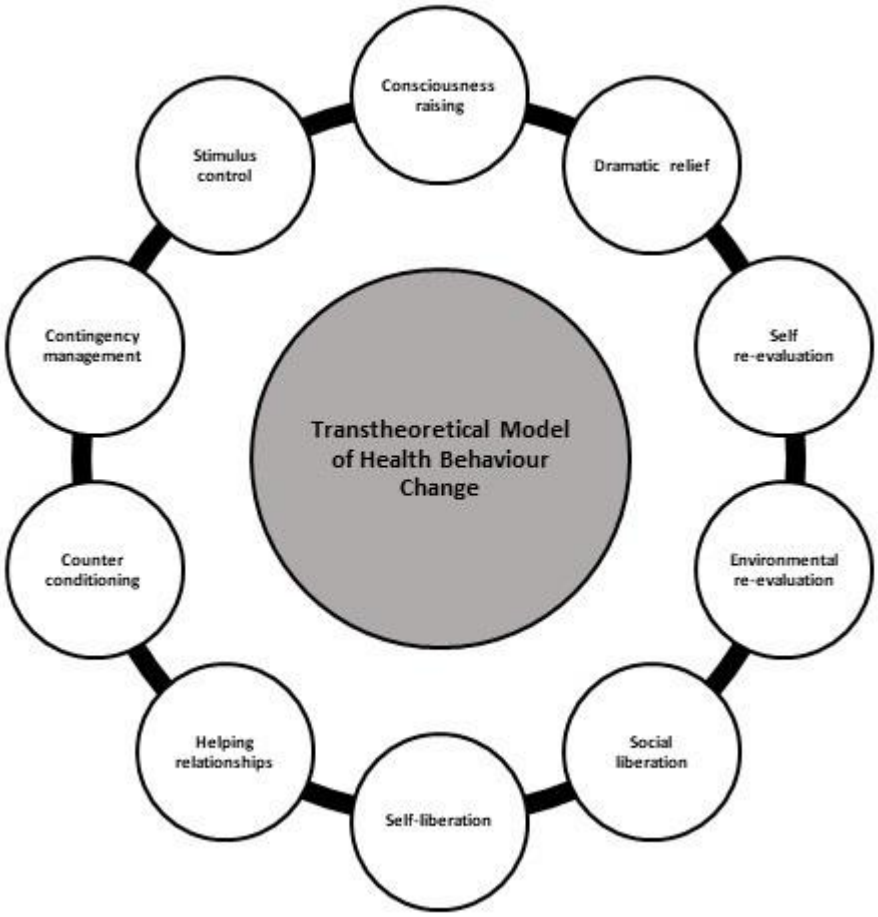


Figure 2.3: Ten processes of health behaviour change

2.3.1.1 Relevance of chatbots to the *maintenance* phase

As may be seen from **Table 2.2** the maintenance phase of the Transtheoretical Model is concerned with increasing awareness, emotional arousal, re-evaluation, reinforcement of positive and reduction of negative behaviours and building support networks. Chatbots may be suited for these purposes as they can respond to human emotion and create empathy (Barak et al. 2009). Chatbots have also demonstrated increased user engagement in health interventions (Fadhil & Gabrielli, 2017). The potential for chatbots in stimulating the maintenance phase of the Transtheoretical Model is discussed further in **chapter 4 (section 4.3.1)** and **chapter 7 (section 7.8)**.

2.4 Weight Loss Maintenance

The NHS recognises that losing as little as 3% body weight is enough to significantly improve health outcomes (NHS, 2018). However, even if this small target amount is reached, maintaining the new weight is a greater challenge, and weight regain is common. The National Institutes of Health suggested that this may be in part due to the perceived

reduction in benefits during maintenance of weight (MacLean et al. 2015). During weight loss, participants observe significant benefits (e.g. feeling better, sense of achievement) relative to the cost of sustaining weight loss. During the maintenance phase, benefits may seem less obvious, and may no longer justify continuing (MacLean et al. 2015). Wing & Hill (2001) refer to a past study in which only around 2% of participants had maintained weight loss of 10kg after 2 years, and another in which it was reported that over 4 years only 0.9% of men and 5.3% of women were consistently able to maintain 100% weight loss.

2.5 Motivation

It is widely accepted that motivation is key to successful weight loss maintenance (Teixeira et al. 2012), and high levels of motivation have been linked with intervention success in recent Randomised Controlled Trials (RCTs) (Donaldson et al. 2014, Fjeldsoe et al. 2016). It has been suggested that motivation may positively influence the processes of behaviour change associated with the *maintenance* phase of the Transtheoretical Model, particularly the process of *self-liberation* (Prochaska & Velicer, 1997), which, as discussed in **section 2.3.1** is the belief that the individual *can* change, leading to commitment to make and sustain positive changes. The *WeightMentor* chatbot will support the *maintenance* phase by reinforcing six of these ten processes (discussed in **section 4.3.1**), one of which is *self-liberation*. Thus, motivation will form an important part of the *WeightMentor* chatbot's built-in functionality.

Motivation may be provided in a variety of ways. The most basic form of motivation is that of the "Motivational Quote". These are a popular phenomenon on the internet today, and although they appeal only to a narrow selection of the population, they are a "powerful incentive to try harder", because they create the impression that someone else believes in you (Moran, 2015). Additionally, it is suggested that motivational quotes reach humans at a primal level, as we are essentially aspirational creatures. We look up to people we consider to be role models and tend to follow their advice. Hence the implication is that motivational quotes inspire us to be more like the people we admire. Another tool is Motivational Interviewing (MI), which was originally developed for treatment of substance abuse and addiction but has since been deployed in a wider healthcare setting. It may be described as "person-centred" and focuses on strengthening an individual's motivation for change through techniques such as reflective listening and

shared decision-making. MI has been observed in some cases to improve intervention outcomes; however, it is most effective when used as part of an intensive intervention (Teixeira et al. 2012).

Individuals may also be motivated using personalised feedback messages. These messages may be based on previous progress (where the individual is reporting factors such as weight, physical activity or diet on a daily or weekly basis), provide tips or strategies for making or sustaining changes to diet or exercise regime, or share success stories from others who are like-minded. Personalised messages such as these have been shown to be successful for motivating towards maintenance of weight loss (Thomas et al. 2011, Donaldson et al. 2014, Fjeldsoe et al. 2016) and are discussed further in **section 2.6**.

2.6 Communication in Weight Loss Maintenance

Communication is an important factor in weight loss and weight loss maintenance. Moorhead et al. (2013) highlighted the importance of communication in obesity management and pointed to a strong link between effective communication and positive health outcomes but argue that the effectiveness of health communication is reduced by the complexity of diet and nutrition information. Stephens et al. (2015) conducted a focus group to investigate attitudes to technology and weight loss among 18 to 30-year old adults and concluded that individuals in this age group value expert help in relation to diet, physical activity, calorie consumption, setting achievable goals and effective exercise. Participants also stated that information from experts regarding the accuracy of rumours and fad diets was helpful in distinguishing fact from fiction. Clearly there is a significant need for accurate, honest health information, however many health professionals are reluctant to honestly discuss obesity with their patients, and often consider it to be a taboo subject (Moorhead et al. 2013).

2.6.1 Text and Multimedia Messaging

Recent research has investigated the use of technology-based tools as a mechanism for communication for health behaviour change, which is a key part of obesity management. Seven studies in this thesis literature review used SMS (Short Messaging System – text messages) as part of a structured weight loss intervention. Joo & Kim (2007) sent weekly short messages to participants (n=927) giving advice on diet, exercise and behaviour

modification. After 12 weeks, 47% of participants had completed the intervention (n=433) and lost an average of 1.6kg. Waist circumference and BMI were also reduced by 4.3cm and 0.6kg/m^2 respectively. Losses in weight, waist circumference and BMI were significant, and participants reported high levels of satisfaction with the intervention. Results from this study suggest that text messaging is an effective means of delivering community-based weight loss initiatives. Patrick et al. (2009) compared the effectiveness of a combination of daily SMS and MMS (Multimedia Messaging Service) messages, printed materials, and a monthly phone call with that of printed materials only. After four months, weight loss in the intervention group was significant, and greater than the control group. Most (92%) of the twenty-four participants in the intervention group said they would recommend the intervention to others. It was concluded that text messages could be effective for promoting behaviour change in weight loss (Patrick et al. 2009). A 2014 conference presentation by Abraham et al. reported on a pilot study involving obese Chinese teenagers (aged 12 to 18). Participants were asked to set diet and physical activity goals at the start of the intervention and then every month for six months. Weekly text messages were sent to participants, who indicated if they had achieved their goals for that week. This study reported that all participants and 93.8% of parents found the weekly texts and goal setting useful. It was concluded that the text message-based intervention was feasible and highly acceptable for participants (Abraham et al. 2014).

The Short Messaging Service Maintenance Treatment (SMSMT) (Bauer et al. 2010, de Niet et al. 2012) was a text-message-based maintenance intervention designed to improve sustainability of "Big Friends Club" (BFC), a 12-month cognitive behaviour change programme for children aged seven to twelve. After the first three months of BFC, which were highly intensive, participants transitioned into a second, less-intensive phase. SMSMT was initially assessed for feasibility using 40 children who were assigned to the intervention for 36 weeks having completed the 3-month intensive phase of BFC (Bauer et al. 2010). Participants self-monitored eating, exercise and emotions and reported weekly by text using a five-point Likert scale. Weekly tailored response texts were sent based on submitted self-monitoring data. Results from the study suggested that the SMSMT intervention was feasible.

The SMSMT feasibility study was followed by a full randomised controlled trial (de Niet et al. 2012). In this trial, 141 BFC participants who had just completed the initial 3-month

intensive phase were randomised to either the SMSMT intervention group, whose BFC engagement was supplemented by SMS-based support, or the control group, who continued the less-intensive phase as normal. Participants submitted reports on physical activity, diet and mood, using weekly SMS messages (de Niet et al. 2012), and in return received an individually tailored feedback message. Although the SMS intervention had no overall positive effect on BMI-SD, wellbeing and eating behaviour, it was observed that motivation was higher in intervention group participants, who were less likely to withdraw from BFC than those in the control group (de Niet et al. 2012). This is consistent with findings from similar studies, such as Newton et al. (2009), which used pedometers and motivational text messaging to improve physical activity in adolescents with type 1 diabetes. This study determined that motivational reminders had no positive effects on the outcome measures. It is suggested by de Niet et al. (2012) that the self-monitoring procedure could be improved, increasing the frequency of self-monitoring data collection and monitoring other factors such as sedentary behaviour. The use of a Likert scale for measuring health behaviour adherence was identified as a drawback to this study. It is proposed that, in future, more appropriate techniques could be used, such as accelerometers, pedometers and standardised questionnaires (de Niet et al. 2012).

The *Lifestyle Eating and Activity Programme (LEAP) Beep* trial (Donaldson et al. 2014) also used text messaging. Participants (n=34) had successfully completed the *Lifestyle Eating and Activity Programme (LEAP)*, a twelve-week weight management intervention run by Leicestershire Nutrition & Dietetic Services. Participants agreed weekly targets for physical activity, fruit and vegetable consumption and breakfast. Reminder texts were sent twice a week to participants, who sent their own self-monitoring data by text on a weekday and at the weekend. Although both the intervention (n=17) and control (n=17) groups self-reported, only the intervention group received personalised feedback, either congratulatory if goals had been met, or encouragement and tips if they were not. Participants were generally very satisfied with the intervention, and there was a significant difference in waist circumference and BMI between the two groups. It was reported that attendance at follow-up clinics was higher within the intervention group than within the control group. It is unknown whether this was a result of improved motivation due to successful weight loss maintenance, or whether it was simply a case of participants seeking approval or congratulations from the research team. A study conducted by Truby et al. (2006) suggested that merely taking part is enough to motivate participants. This is known as the Hawthorne effect (Sedgewick & Greenwood, 2015).

Donaldson et al. (2014) proposed that if participants are willing to take part then they are already motivated to change behaviour. Successful weight loss maintenance has been shown to have positive effects on health-related quality of life (Donaldson et al. 2014) and it is possible that improved quality of life leads to increased confidence and thus participation in trials becomes easier. The authors acknowledge this as a limitation to the study and suggest that the presence of the Hawthorne effect may introduce bias.

Get Healthy Stay Healthy (GHSH) was a text-message based follow-on from the Australian Government's *Get Healthy Information and Coaching Service (GHS)*, a six-month telephone-based weight loss coaching programme (Fjeldsoe et al. 2016). An initial tailoring telephone call with GHSH participants determined individual goals, target behaviours, barriers to progress and solutions. Participants (n=228) also nominated a "buddy" who could help them achieve their goal. Text messages were highly personalised, based on desired number per week, timing, and type (self-monitoring of weight, goal checking, behavioural prompts, goal reset). Texts were sent automatically and reviewed by research staff if required. The control group were given general feedback after each data collection but did not receive any personalised feedback. Significant benefits for the intervention group were reported in terms of weight change and physical activity and participants were very satisfied with the behaviour reinforcement skills they learnt during the intervention (Fjeldsoe et al. 2016). Several limitations were highlighted by the authors with respect to the size and nature of the population sample. Results were inconclusive for some of the outcome measures including vigorous physical activity, waist circumference and fruit & vegetable intake. This was attributed to the sample size being insufficient for these particular outcome measures. It has also been suggested that the sample may not have accurately represented the whole population.

A feasibility study by Gerber et al. (2009) recruited ninety-five African American women already enrolled in a weight-loss programme. These women chose the content and timing of their own personalised text messages which were delivered to them at least three times per week. Participants were generally very satisfied with the intervention, reporting that they read the text messages regularly and that they helped them to achieve their goals. One participant compared text messages to "a little person sitting on her shoulder helping her make the right choices" (Gerber et al. 2009). Many participants stated that they "looked forward" to receiving the text messages. Despite these positives, the authors

encountered several challenges to the success of the study. Although the majority of over 4,500 sent messages were delivered successfully, 114 were undelivered, either due to the recipient telephone number being deactivated or the user's inbox being full. Delivered text messages were not always read and responded to immediately. While certainly feasible in the short term, it was not possible to determine long-term feasibility (Gerber et al. 2009).

Three literature review papers investigated the effectiveness of text message-based interventions and reported positive benefits of this type of approach. Fjeldsoe et al. (2009) reviewed 14 studies, eight of which observed significant positive behaviour changes, and five observed positive effects but were statistically underpowered. The reviewers concluded that most of the studies focusing on clinical care (e.g. diabetes management) failed to evaluate target behaviours and instead evaluated health outcomes. The authors highlighted the poor evaluation of process outcomes within reviewed studies and wide variability of compliance with interventions and their acceptability among participants.

All the 19 limited contact interventions reviewed by Fry & Neff (2009) made use of "periodic prompts" (defined as messages, reminders, or brief feedback, communicated multiple times) and targeted weight loss, physical activity or diet. Eleven reported positive short-term results, and it was observed that the effectiveness of the intervention was enhanced when prompts were frequent, and interventions included personal contact with a counsellor. The authors identified several study limitations. Firstly, limited long-term follow-up made it difficult to assess long-term effectiveness. Where follow-up data were included, results were inconclusive, and data were often collected using diverse methods that made comparison difficult. Another limitation concerned bias introduced by selection of participants. All except one study included participants who had volunteered to participate in the intervention, and the reviewers suggested that participants who volunteer for such studies are already motivated towards behaviour change. Related to this was the disproportionate numbers of female vs. male participants in the reviewed studies. Women were noted to be more likely to participate in this type of intervention, which creates difficulty in determining how men's behaviour may be changed. Finally, all reviewed studies used in-person data collection methods which may not be easily applied on a larger scale.

A review by Derbyshire & Dancey (2013) investigated the broad role of technology in women's health and reported on two RCTs that suggested positive benefits from the use of SMS messaging for weight loss. Both trials reported very positive results, with good participant adherence and significantly more weight loss in the experiment groups compared to the control groups (Shapiro et al. 2012, Haapala et al. 2009).

It is clear from these seven studies and three reviews that text messaging is a feasible and effective means of delivering weight loss and maintenance interventions, although the benefits of such interventions are largely transient. Further research is warranted, specifically to investigate ways to optimise effectiveness, improve study quality, and explore effectiveness over longer periods. Health messages are most effective when they are sensitive, accurate and honest (Moorhead et al. 2013). This may be facilitated through personalising the message to the individual user and carefully considering the tone of voice of the message.

2.6.2 Personalised Messages

As discussed in **section 2.5**, personalised messages have been linked with increased motivation for weight loss maintenance. Cohen et al. (2017) describe personalised messages as "customised information interventions, which fit the characteristics, beliefs and needs of particular target populations or individuals". Although early personalised messaging systems were primarily text-based, recent research into the use of video-based systems has been conducted (Baranowski & Frankel, 2012). Many contemporary personalised messaging systems require the targeted individual to personalise the system themselves before first use, usually by indicating the unhealthy behaviours they wish to target, or by setting goals. Messages will then be personalised based on the user's preferences, supporting motivation and providing feedback on previous successes or failings. Feedback may be provided by a human operator or by a computerised expert system (Baranowski & Frankel, 2012). This type of system is believed to be popular and effective for health behaviour change because the information will be perceived as more personally relevant, thus the individual is more likely to pay greater attention to received messages.

While generic communication offers greater benefits than no contact at all, message personalisation positively affects outcome measures (Fry & Neff, 2009) and reduces

participant attrition (Fjeldsoe et al, 2009). Two personalised message interventions observed positive, if transient benefits for health behaviours. "FATaintPHAT" ("VETisnietVET" in Dutch) involved 883 participants aged 12 – 13 from across 20 schools, and measured diet, physical activity, and sedentary behaviour along with Body Mass Index (BMI), Waist Circumference (WC) and fitness. Although no effects were observed on physical activity or sedentary behaviour, there were short term benefits observed for dietary behaviours (Ezendam et al. 2013). The "Framed Interactive Theory-driven Texting" (FITT) study (Cohen et al. 2017) recruited 89 obese adults from African American churches and used gain/loss framing to tailor messages for motivating weight loss. Gain/loss framed messages focus on either the benefits (gain) of maintaining positive behaviours, or the costs (loss) of abstaining from positive behaviours. Participants were assessed using the Regulatory Focus Questionnaire (Higgins et al. 2001) to determine their motivational orientation as either promotion (gain) focused or prevention (loss) focused. Based on this information, participants were assigned to one of four groups, receiving gain or loss messages that either "matched" their motivational orientation, or "mismatched". It was observed that where participants received "matched" text messages their motivation to change was increased (Cohen et al, 2017). Findings from these two studies confirm what has already been suggested by Fry & Neff (2009), that personalisation positively impacts intervention outcomes.

According to four recent focus group studies, individuals are more likely to engage with messages that are personally relevant to them (Pollard et al. 2016). Adolescent participants in a study by Smith et al. (2014), stated that personalised messages felt more human and less automated, and that overly formal messages were just another form of nagging. Young adults (aged 18 – 25) who took part in a focus group study by Stephens et al. (2015) said that if messages were not personalised there was a high chance they would be ignored. Stephens et al. (2015) found that young adults who use smartphones are used to receiving large numbers of generic notifications, whether from their bank or from university. These are obviously sent in bulk to a distribution group and contain very little personalisation. For messages to be considered important they will need to be personalised. Tailored messages were the topic of a focus group study conducted by Woolford et al. (2011), which included adolescents recruited from the *Michigan Pediatric Outpatient Weight Evaluation and Reduction (MPOWER)* program. Participants in this study said they found personalised messages very relevant. Where messages were designed to specifically address participants' target behaviours, these were popular if

participants could see how these tips might form part of their daily routine (Woolford et al. 2011).

It is evident from current research on message personalisation that health messages are most effective when relevant to the recipient. Yap & Davis (2008) suggested that personalised messages attract more attention (as discussed by Kreuter et al. 2000) and "enable the participant to say, 'this applies to me'" (Yap & Davis, 2008). The focus group conducted by Smith et al. (2014) recommended that messages should not be too formal and should read as though written by a human rather than appearing to come from a machine (Smith et al. 2014).

2.6.3 Auditory Messages

Most of the research into message personalisation has focused primarily on text-based messaging systems. However, a recent study by Elbert et al. (2017) explored the potential of computer-tailored auditory messages, a still largely unexplored area. The authors described specific characteristics unique to auditory messages. The presence of an actual human voice helps to emphasise the point of the message, which becomes more prominent in the mind of the listener and generates a sense of social linkage. Additionally, auditory information creates a level of convenience that would be unavailable with text-based information, as auditory messages may be listened to and processed by the brain while performing other tasks, such as driving, walking, or housework (Elbert et al. 2017). The three core components, or "ingredients" of a message, as described by Elbert et al. (2017) are *personalisation*, in which specific characteristics of the target individual (e.g. the person's name) are added to an otherwise generic message, *feedback*, which acknowledges and comments on an individual's previous experiences or achievements, and *adaptation* or *content matching*, which selects the tone and content of a message based on the recipient's demographic or social group etc. For example, a message intended for an adolescent girl would be worded differently from one intended for a middle-aged man.

Study participants were assessed for fruit and vegetable intake and self-efficacy, the ability to perform specific behaviours (Elbert et al. 2017). Participants were exposed to four types of auditory message – generic messages, and those personalised using each of the three components discussed above. It was concluded that in cases where participant

self-efficacy was high, none of the messages made any difference to fruit and vegetable intake, whereas participants experiencing low self-efficacy benefited most significantly from listening to personalised messages (using the recipient's name), but only in relation to vegetable intake. Findings suggest that message tailoring is feasible using auditory messages, but further research is required into this area, specifically comparing auditory messages with text-based messages, and investigating if the effectiveness of auditory messages may be enhanced through the addition of visual media.

2.6.4 Message Tones

Messages may be designed to use different tones of voice. Three recent focus group studies explored attitudes to message tones of voice and concluded that message tone influences message acceptability for the recipient. Pollard et al. (2016) presented participants with five different message tones: *authoritative*, *empathetic*, *generation Y engaging*, *solutions* and *substitutions* and sought to determine which were most effective for persuading behaviour change. Authoritative messages, telling the individual what to do and why, received mixed responses. While some participants (mostly male) said they would "sit up and listen" to this type of message, others said they felt these messages were condescending or offensive. In some cases, they led to feelings of guilt or defensiveness. Substitution messages were "helpful and practical" (Pollard et al. 2016). Messages suggesting "tasty" food alternatives were appreciated as they reminded participants that "they could still enjoy food" (Pollard et al. 2016). Empathy was considered to be important, and participants appreciated being encouraged and acknowledged, particularly where the acknowledgement was relevant (for example emphasising "feeling good", and not just "being healthy" (Pollard et al. 2016)) or when it recognised previous achievements. Generation Y engaging messages (targeting young people aged between 18 and 34) were either "casual and humorous", or "belittling and patronising". Some participants said that these messages were "trying too hard", and one participant reported "being made to feel like they were five years old" (Pollard et al. 2016). Participants did however report that these messages were least likely to come across as "preachy".

Smith et al. (2014) conducted a focus group study with adolescents and their parents who had just completed the Curtin University Activity, Food and Attitudes Program (CAFAP). This study found that both parents and adolescents preferred casual, more personal

messages. Feedback from participants suggested that messages needed to "sound more human" and that in some cases messages "sounded like they were coming from an adult". Consistent with Pollard et al. (2016), Smith et al. (2014) found that adolescents wanted messages to create empathy, encouraging them to think about long term goals and why they chose to participate in CAFAP. Messages of support were preferable to those related to behaviour change (Smith et al. 2014). Practical tips were valued as they provided concrete examples and were not seen as simply "telling someone what to do". The study found that although, in general, adolescents appreciated the usefulness of text messages as reminders or to raise awareness, they were quick to dismiss the motivational potential of the messages. This feeling was echoed by the parents, who said they felt that lack of motivation was a significant barrier that could not be overcome by text messages.

Woolford et al. (2011) identified six types of message for focus group testing. Testimonials suggested things that had worked for other people. Meal/recipe ideas proposed healthy options and substitutions. Targeted tips offered suggestions based on individual participants' identified target behaviours. Reflective questions encouraged participants to think about ideas and the consequences of their actions. Feedback questions requested a response from the participant. Tailored messages were tailored specifically to each individual participant. Participants voted "yes" or "no" to each message they viewed (around 70 in total). Participants in this study reacted positively to messages that directly told them what to do. Recipe ideas were popular because they provided concrete examples of things to try, and testimonials provided hope. It was noted however that testimonials were most effective if they came from other teenagers. Reflective questions were unpopular, but feedback questions were popular. In keeping with findings by Smith et al. (2014) and Pollard et al. (2016), participants in this study felt that positive, encouraging, messages were most helpful, and that messages should be casual and natural but should avoid trying to sound like adolescents.

Results from these studies clearly show the importance of message tone, which should be considered carefully when designing messages for communication of weight loss maintenance.

2.6.5 Effect of Threat Information on Message Attention

The focus group studies discussed in section 4.2 suggest that message recipients are more likely to pay attention to messages that are personally relevant to them. While it has long been proposed that fear may also increase message attention (Leventhal, 1971), it has been found in some cases that presentation-of-threat information in fact results in reduced attention to personally relevant messages (Liberman & Chaiken, 1992). This may be a result of cognitive dissonance (Festinger, 1957). A 2010 study by Kessels et al. investigated the effects of both message tailoring and threat information on message attention. Participants (34 undergraduate students) were presented with text-based, tailored health information, which also presented as either high threat or low threat. While reading this information, participants heard either a high or low audio tone and were required to either respond by pressing a button (in the case of a high tone) or ignoring the tone (in the case of the low tone). Messages were tailored based on the individual participant's fat, fruit and vegetable intake and attitudes to behaviour change (Kessels et al. 2011). Participant reaction times were recorded and analysed. Results showed faster reactions to the audio tone when participants were reading non-tailored or low-threat messages, than when reading tailored or high-threat messages. These results concur with what has already been suggested by Liberman & Chaiken (1992), that message personalisation is significant for increasing message attention, but high threat information may be detrimental to message attention (Kessels et al, 2011).

Results from these studies are consistent with findings from Smith et al. (2014) and Pollard et al. (2016) that individuals are likely to pay more attention to messages that they perceive as personal to them. However, if a message contains high levels of threat information, attention may be reduced. As some types of "authoritative" message may be seen as condescending or patronising (Pollard et al. 2016), it is hardly surprising that high levels of threat information may result in reduced attention, as individuals may find the messages too "preachy".

2.7 Digital Technologies in Weight Loss Maintenance

Smartphones have become a prominent part of human life. According to a survey by Petrie (2013), 94% of pregnant women felt that regular smartphone use had improved their lives. Of the women surveyed, 65% reported downloading and using pregnancy apps. A survey in the US reported that 93% of women "keep their smartphone at arm's

length" (Arbitron & Edison Research, 2013). Smartphone applications are also seeing increasing use in healthcare. In their review paper, Derbyshire & Dancey (2013) concluded that there is a clear need for further research and development of mobile phone apps for women's health, particularly in relation to their use in healthcare settings. Feedback suggests that apps should stimulate motivation and should be based on trustworthy evidence. Ten of the sixteen papers in the Derbyshire & Dancey review reported on the use of smartphone technologies for weight loss. Positive benefits were observed within intervention groups in all but one trial. Participants (n=170) in the largest trial were overweight and obese adults. It was observed that high participant engagement resulted in greater weight loss (Shapiro et al. 2012). Mosa et al. (2012) identified a range of smartphone applications for healthcare professionals, medical and nursing students and patients, with a range of functions from disease diagnosis and drug reference libraries to chronic illness management tools. The authors concluded that smartphone-based applications play an important role in patient education, remote management and self-monitoring (Mosa et al. 2012).

The high cost and personnel requirements of face-to-face weight loss interventions (Abraham et al. 2014) present a common rationale for the use of digital technology. A quasi-experimental study by Haugen et al. (2007) determined that telehealth systems were useful for supporting weight loss maintenance as an alternative to participation in structured face-to-face programmes. Participants (n=87, 14 men, 73 women) had all successfully completed *Colorado Weigh (CW)*, a six-month structured behavioural weight loss programme. Post-CW participants were allocated to one of three groups. Group 1 participants (n=31) had opted to participate in *Colorado Weigh Graduate (CWG)*, a classroom-based weight loss maintenance programme targeting physical activity, behaviour modification and nutrition, delivered by a registered dietitian. Physical activity, diet and body weight were monitored and recorded using paper-based logs. Group 2 participants (n=31) took part in *Colorado Weight High-Tech (CWHT)*, in which CWG content was delivered electronically. Participants communicated with their "Healthy Coach", a registered dietitian, online on a fortnightly basis (Haugen et al. 2007). A computer program, *BalanceLog*, was used to monitor dietary behaviour and physical activity. Group 3 participants (n=25) had declined participation in both programmes and received no support but were observed for six months, including two follow-up clinic visits at three and six months (Haugen et al. 2007).

Results at study completion showed that, whereas participants in groups 1 and 2 (CWG and CWHT) experienced further weight loss (0.5±4.3kg and 0.6±2.5kg respectively), those in group 3 (no programme) regained weight (1.7±3.0kg). Participants in the first two groups were asked to rate the respective programmes in terms of satisfaction and convenience. Both programmes were rated equally highly for satisfaction, although the CWHT programme was rated significantly higher than CWG for convenience (Haugen et al. 2007). It was concluded that support, monitoring and feedback are essential for further weight loss or maintenance (consistent with findings from *LEAP Beep* (Donaldson et al. 2014) and *GHSH* (Fjeldsoe et al. 2016)). The observation that there is no significant difference in effectiveness between telehealth and classroom-based programmes suggests that telehealth programmes may be a suitable alternative for individuals who are not keen to undertake further classroom-based programmes. Telehealth programmes also offer the benefits of reduced time commitment and longer sustainability (Haugen et al. 2007).

2.7.1 Electronic Diaries

Burke et al. (2012) discussed the SMART Trial, a 24-week intervention which evaluated the effectiveness of paper-based self-monitoring against two PDA (Personal Digital Assistant) based systems, one which sent daily tailored feedback messages, and one which did not. A variety of treatment components were used: group sessions, daily (energy and fat) and weekly (exercise) goals and self-monitoring (exercise and diet). It was observed that adherence to the programme was higher in the two PDA groups than in the paper diary group. Higher weight loss was associated with better treatment adherence. The researchers concluded that electronic diaries for self-monitoring can improve adherence to weight loss treatments. These findings are consistent with what has been suggested by Donaldson et al. (2014) and Fjeldsoe et al. (2016), that self-monitoring is key to successful weight loss maintenance.

2.7.2 Mobile Web and Phone Apps

A systematic literature review by Aguilar-Martinez et al. (2014) identified eight studies that delivered weight loss interventions using mobile phone apps or web-based systems designed to be accessed from a mobile phone. Two of these studies were conducted in order to assess feasibility of the chosen intervention delivery methods (Aguilar-Martinez et al. 2014). From these two studies it was determined that mobile phone based tools were useful as part of weight loss interventions (Mattila et al. 2011), acceptable to

participants (Morak et al. 2008), and were most effective when used regularly (Mattila et al. 2011). The remaining six studies in the Aguilar-Martinez review assessed the effectiveness of mobile phone-based interventions (Aguilar-Martinez et al. 2014). All of these except two reported a reduction in outcome measures. One study observed no difference between groups, while another reported in-group weight reduction at 3 and 6 months (Aguilar-Martinez et al. 2014). These findings suggest strong evidence in favour of mobile web and phone-based interventions for weight loss maintenance.

Two of the studies discussed by Derbyshire & Dancey (2013) investigated the use of smartphone apps for weight loss. A specialised app was developed for a meal replacement programme trial by Brindal et al. (2013), which took place over 8 weeks. Although differences in weight loss were not significant, intervention group participants did report improvements in mood compared to those in the control group. It is suggested that findings may have been more significant if the study duration had been longer.

A weight loss app was compared with a website and paper-based diaries in a six-month trial by Carter et al. (2013). Participants were overweight adults, randomised into a smartphone group (SG) (n=43), website group (WG) (n=42) and diary group (DG) (n=43). SG participants used a specially designed app called “My Meal Mate” (MMM) to set goals and self-monitor diet and physical activity and received weekly motivational feedback text messages. This group was compared to both the WG, whose participants used a slimming website, “Weight Loss Resources” to self-monitor, and also to the DG, whose participants used a paper diary and calorie-counting book, provided by “Weight Loss Resources” (Carter et al. 2013). Participant retention was highest in the SG (93%) and lowest in the DG (53%). Adherence to the intervention was statistically significantly higher in the SG, compared to WG and DG. The study was not sufficiently powered to detect weight changes in study groups, however intention-to-treat analysis was used to estimate weight change in each group. This analysis determined that weight change was the largest in the SG, and the difference between groups was statistically significant at six months (Carter et al. 2013). It was concluded that the MMM app was acceptable and feasible for delivering weight loss interventions, but a full RCT was warranted.

Health-E-Call” was a 12-week weight loss programme incorporating a smartphone-based intervention with weekly in-person weigh-ins and paper-based lessons. The smartphone-based component involved self-monitoring, feedback and training videos, using both the

"Health-E-Call" app developed by the research team, and a commercially available app called "DailyBurn". Twenty overweight/obese participants self-monitored weight, food intake and exercise, monitored and acted on their own problem behaviours, and received feedback. Participant satisfaction was high, and self-monitoring, real-time feedback and accountability were viewed as very helpful (Thomas & Wing, 2013). This was one of the first studies to test smartphone-based systems for delivering behavioural weight loss interventions and showed favourable outcomes. However, the lack of a control group made it impossible to compare results. The small sample size was also a limitation. The authors recommend a further randomised controlled trial of the intervention (Thomas & Wing 2013).

Diet-A was a mobile phone-based application supporting voice- or text-based entry of food intake, real-time feedback and advice on disease prevention based on user input. A three-month pre-post intervention study was conducted to determine the feasibility of the application for monitoring dietary intake (Lee et al. 2017). Participants (n=33, 9 male, 24 female) were 16-18-year-old high-school students, and used the system to self-report food and beverage consumption using the app. Nutrient intake was assessed using two 24h recalls at the start and end of the intervention, and compared with nutrient intake recorded using the app. Based on the two 24h recalls, it was observed that sodium and calcium intake significantly decreased between the start and end of the intervention. It was suggested that participants tended to underreport their nutrient intake during daily app use, because the difference between nutrient intake estimated by the app and nutrient intake estimated in the 24h recalls was significant for energy, carbohydrates, protein, fat, sodium and calcium. Most (61.9%) of the 24 participants who completed questionnaires reported satisfaction with the Diet-A application, and 47.7% liked the personalised feedback aspect. Despite this, more than 70% of participants reported difficulty remembering to use the app or found it time-consuming to use. This study suggested that, while it may be possible to feasibly use smartphone applications for monitoring dietary intake, accuracy may be limited due to frustration with using the app, and participant tendency to under-report daily nutrient intake. The sample size was judged to be too small and under-representative of the general population. In order to improve the study strength, it will be prudent to conduct further research, and to improve app functionality.

From the studies discussed above, there is some evidence to suggest that mobile web and phone apps may be feasible for promoting weight loss, however many of the studies

reviewed require further research in order to be considered conclusive. Despite a strong trend towards the use of mobile applications, to date there has been limited research into the use of smartphone apps for weight loss maintenance, although nutrition apps seem to be popular on Google and Apple stores. A 2016 review of nutrition apps found that the thirteen most popular apps had been downloaded at least one million times (Franco et al. 2016). This review found that most of these apps functioned as a food diary, and generally targeted weight loss and calorie counting. The lack of input from health professionals was identified as a serious limitation which could mislead users – it is suggested that blindly following recommendations made by such an app could trigger eating disorders or harmful dieting (Franco et al. 2016).

Brindal et al. (2016) identify this lack of scientific basis as a major shortcoming of weight management apps and have proposed a protocol for a science-based, purpose-built smartphone app for weight loss maintenance. The MotiMate app is designed to support behaviour change for weight loss maintenance by helping the user to develop coping strategies and improve self-awareness. Key features include push notifications to remind the user to view feedback and submit data, tools for monitoring diet and exercise, weight and wellbeing, and tools for summarising and reviewing previous submitted data. These include a weekly summary of progress, and the capacity to identify areas for improvement. Coping strategy development is based on social support (for example encouraging the user to talk about their problem with someone who can help), personal strengths (encouraging the user to identify how their own strengths can help them to cope), distraction (finding ways to avoid the problem), emotional cognitive strategies (identifying ways to change their thinking) and active strategies (making an action plan).

A bespoke mobile-web based app was just one of many tools used in the *NULevel Trial* by Evans et al. (2015), which acknowledges the limitations of previous internet-based interventions (Evans et al. 2015). Results of this study were recently published (Sniehotta et al. 2019), and suggest that there was no significant difference in weight after 12 months between the intervention group (n=131) and control group (n=133), however intervention group participants were more physically active, weighed themselves more frequently, and reported greater satisfaction with outcomes, better planning, and increased confidence (Sniehotta et al. 2019).

2.7.3 Gamification in Weight Loss Maintenance

Technologies for weight control and healthy living are not solely limited to text/multimedia messaging or smartphone apps. Baranowski & Frankel (2012) discussed how gaming technology may be used to promote healthy lifestyles and stimulate physical activity changes, particularly in children. The authors identified children as some of the heaviest users of electronic media (Baranowski & Frankel, 2012) and suggested that much of the electronic media favoured by children do little to promote physical activity and can lead to unhealthy lifestyles and dietary practices. It is reported that some forms of digital media have been shown to increase knowledge and awareness of obesity, but these still do not do enough to help children apply this knowledge to their lives. This paper identifies five main categories of technology that may be helpful in promoting behaviour change.

2.7.3.1 Active Video Games

Active Video Games (AVG) are, first and foremost, video games, but they differ from traditional video games in that they require some form of physical activity from the player. The most well-known examples of AVG are the Nintendo Wii and XBOX Kinect. Although the presence of such systems in the family home has been shown in some cases to increase physical activity in the short term (Barnett et al. 2011), a recent randomised controlled trial suggested that, in the long term, there are no benefits to be gained from AVG over non-AVG games (Baranowski & Frankel 2012). Further research will perhaps determine how AVG use may increase physical activity and maintain this in the long term.

2.7.3.2 Interactive Multimedia Games

Interactive Multimedia Games (IMG) are traditional video games that involve some sort of multimedia element, which are designed to stimulate behaviour change. A recent paper by Baranowski et al. (2009) reviewed twenty-seven articles relating to the use of IMG technology for behaviour change and reported that all but one of the articles showed positive outcomes for behaviour change as a result of playing the games (Baranowski et al. 2009). While this research is promising for the technology, further research is required into how video games may be designed to maximise their potential for health behaviour change (Baranowski & Frankel 2012).

2.7.4 E-mail

Email support was the basis of a trial by Thomas et al. (2011). Participants (n=55) were recruited from a twice-weekly weight loss clinic run by Portsmouth Hospitals Trust. Intervention participants (n=28) received a weekly tip email from a trust dietitian and reported their current weight monthly. Control group participants (n=27) had no further contact but returned to the clinic after six months for a follow up. The intervention group maintained a greater weight loss than the control group and reported that the weekly tip served to "jog their memory" effectively. Attrition rates in the study were very low (8%) compared to similar studies such as that of Cussler et al. (2008), who reported 21.2% and 14.5% for their two intervention groups. Thomas et al. reported that participant ages ranged from 17 – 71, suggesting that older participants were not put off using technology. Despite these strengths, it was suggested that the study is potentially biased, in the sense that participant email addresses were required, eliminating the possibility of study blinding. Additionally, participants were at different stages of weight loss at the start of the study, some had lost as little as 5% of their body weight while others had lost more than 20%. Another weakness was the short duration of the study. The authors suggest that such a trial should run for longer than six months (Thomas et al. 2011).

2.7.5 Internet-based Systems

2.7.5.1 Web-based technology

Although web-based technologies are gradually being superseded by newer technologies, several recent studies have investigated the use of bespoke web-based systems for behaviour change. Healthy Weight Assistant (HWA) was a multifunctional web-based system developed by the Netherlands Nutrition Centre (Kelders et al. 2010), designed to help users maintain a healthy body weight. Target users were either normal body weight or slightly overweight. On first use, users self-reported body weight, diet, physical activity and emotions. These data were used to make personalised recommendations for goal setting and motivation. Users also identified times when they were most at risk of unhealthy behaviours (difficult moments) and were coached to develop solutions. HWA was evaluated by Kelders et al. (2010), and it was found that evaluation participants were very enthusiastic about the system. Participants suggested desirable features for inclusion in the HWA interface; these included re-using previous data (to prevent users needing to re-enter it), combining user entered data with other sources, and more feedback on entered

data. It was felt by participants that progress should be represented visually for clarity. Overall, the system was received very positively by the participants, and the researchers concluded that improving the HWA based on participant feedback would improve its effectiveness (Kelders et al. 2010).

PREDIRCAM was designed to improve communication for behaviour change (González et al. 2013). The project made use of a bespoke website combined with a heart rate monitor. Website functionality included an electronic record module for clinical data recording and scheduling clinical visits, modules for monitoring and reporting exercise data and diet, modules for communication and notifications, questionnaires, and a library of relevant articles and guidelines for healthy living. A 15-day feasibility study of the system was conducted using 15 participants, who were asked to regularly visit the website and use the various features. The website was visited an average of 1.1 times per day per participant, and the website was evaluated favourably by 11 of the 15 participants, with 10 considering the platform to be feasible in the long-term (González et al. 2013). In general, the system was deemed feasible and ready for further clinical evaluation. A potential shortcoming was identified in the food intake monitoring process, which was perceived as "laborious" by participants. The authors concluded that this procedure would require prior training which was not accounted for during the feasibility trial (González et al. 2013). The small sample size and short duration make it difficult to draw definitive conclusions from this study.

Project YEAH (Young adults Eating & Active for Health) was designed by researchers from 14 US universities to prevent weight gain in young adults (Kattelmann et al. 2013). It was based on *Project WebHealth*, a previous intervention by Greene et al. (2011 & 2012). This intervention successfully improved fruit and vegetable intake and physical activity outcomes but did not reduce weight gain. *Project WebHealth* participants also identified shortcomings with the lessons, most notably the lesson length and a need for greater personalisation and enhanced technology use (Kattelmann et al. 2013). *Project YEAH* was developed using the Community Based Participatory Research (CBPR) process. CBPR was selected as it involves the target audience in the development phases of the intervention and ensures that their health problems and challenges are suitably addressed. Over the eight phases of development of the project, social, environmental and educational factors were assessed and used to design the web-based system, which was then pilot tested (Kattelmann et al. 2013). A full randomised controlled trial followed

pilot testing (Kattelmann et al. 2014); this was a fifteen-month intervention involving 1,639 students across thirteen university campuses. The intervention made use of web-based lessons and email messages targeting eating behaviour, physical activity, stress and weight management. Results from *Project YEAH* suggest that, although there were no significant benefits for weight change, BMI, or waist circumference, the intervention brought about positive changes in other areas such as fruit and vegetable intake and self-regulation (Kattelmann et al. 2014). It is believed that these positive behaviour changes could reduce weight gain over time. Physical activity was also not significantly affected by the intervention, but this is believed to be partly because most of the participants were already active at the start of the intervention (Kattelmann et al. 2014).

Results from these studies suggest that web-based systems are generally received positively by potential users, who tend to engage well with this type of system. However, in order to be most effective this type of system requires careful design to ensure that it is user-friendly and provides features that will be most useful.

2.7.5.2 Web-based Educational/Therapeutic programs

Web-based educational/therapeutic programs are essentially digital or on-line versions of face-to-face group education or therapy programmes. Such a system was used in the SB2-BED randomised controlled trial (Jones et al. 2008). Participants were high school students (105 total) at risk of becoming overweight. The 16-week intervention was web-based, and used cognitive behaviour therapy, educational weight loss and self-control techniques. Outcomes were modest, and it was suggested that participant motivation was not maintained during the study.

A report by Bennett & Glasgow (2009) revealed that interventions for weight management are the most extensively studied of internet-based health interventions. The report identifies several RCTs where internet technologies are used for weight loss and suggests that internet interventions are generally very effective in promoting short term weight loss, but there is limited evidence of their effectiveness over longer periods, or for effectively supporting weight loss maintenance (Bennett & Glasgow, 2009). The authors conclude that the accuracy of recent internet-based trials is challenged by the wide variety of components used and high participant attrition rates (Bennett & Glasgow, 2009).

Healthy Weight for Life (Cussler et al. 2008) investigated the use of internet technology for weight loss maintenance in perimenopausal women who had completed an initial four-month weight loss programme. The intervention gave participants access to a website with communication tools (message boards, chat rooms, and personal or group messaging tools), self-monitoring tools (for body weight and physical activity recording etc.), diet and exercise information and links to other online resources. Participants reported their weight once a week, diet four times a week, and exercise three times a week, and received personalised progress reports. Chat rooms were used to organise support groups. The control group had no access to the website but organised their own in-person support groups to practice techniques they had learnt during the weight loss intervention. There was no difference in weight regain between the two groups, and it has been suggested that this may be due to the impact of the social support groups which those in the control group organised. Additionally, the control group demonstrated a strong sense of competition with the intervention group, a principle known as the Avis effect (Cussler et al. 2008).

Two similar trials, the *Weight Loss Maintenance* randomised controlled trial (WLM) (Svetkey et al. 2008) and *Study to Prevent Regain (STOP Regain)* (Wing et al. 2006) evaluated the effectiveness of web-based tools compared with face-to-face contact. Participants in the technology arm of the WLM trial self-reported their weight via a web portal which also gave them access to social support tools (bulletin board and chat room). Users received personalised reminders to log in to the website and could view personalised progress reports but had no access to the intensive face-to-face coaching. Participants in the face-to-face intervention attended regular coaching sessions where they reviewed progress and goals. These two interventions were compared with a self-directed group who were given general guidance at the start of the trial and attended a follow-up session after 12 months. It was reported that the technology-based intervention showed early transient benefits for preventing weight regain, while the face to face intervention showed more modest long-term benefits. The web portal was specifically designed for the WLM trial by an interdisciplinary team of designers, behaviour change experts and public health researchers (Stevens et al. 2008). It included several tools intended to emulate the features and functionality of a face-to-face intervention and to help support participants in weight loss maintenance (Funk et al. 2010). Social support, self-monitoring, guidance, tailored messages and problem solving were among the key functions of the website (Stevens et al. 2008).

One feature of the website was the Tailored Self-Assessment Tool (TSA), designed for intervention group participants to review progress and plan future weight management. Participants could complete the tool at any time, but email reminders were sent to those participants who had not only not completed it, but had also regained weight (Funk et al. 2011). TSA coached users towards behaviour change, but this was a user-led process. Funk et al. (2011) reported good engagement with the TSA tool. Fifty-three percent of participants used the tool, with seventy-two percent of these using it more than once (Funk et al. 2011). During the trial, it was observed that almost two-thirds (65%) of participants were continuing to use the site after 28 months (Funk et al. 2010). The WLM trial was similar in design to STOP Regain (Wing et al. 2006). The technology-based intervention was identical in content and design to the face-to-face intervention, although weight reporting was done via a website, and weekly meetings conducted via chat room. Intervention group participants also had access to weekly tips and information via the site. The face-to-face group had no access to the website and all meetings were conducted in person. This study concluded that the two interventions were both more effective for weight loss maintenance than self-directed maintenance, although the face-to-face intervention was slightly more effective.

All three of the internet-based RCTs reviewed here suggest that internet-based interventions may be effective for weight loss maintenance in the short term, there have been significant advances in the internet in the past ten years, and it could be of interest to replicate the studies using contemporary internet technologies.

2.7.6 Electronic Health Record (EHR) Tools

An innovative ongoing trial by Conroy et al (2017) investigates the use of EHR tools for weight loss maintenance in primary care patients. Intervention participants self-report through an EHR web-based portal and receive two years of personalised coaching. Participants report on factors such as diet, physical activity and stress by completing online questionnaires which are reviewed by a specialist team and each participant receives personalised feedback based on their responses. The control group use the same portal and complete the same questionnaires but receive only general feedback (Conroy et al. 2017). Results from this study have not yet been published.

2.7.7 Use of Multiple Technology Types

Some contemporary studies have used more than one type of technology for delivery of healthy living and weight management interventions. TXT2Bfit was a 2013 trial by Hebden et al. (2013) which sought to improve weight management and promote healthy behaviour among young adults. It is reported that young adults are particularly at risk from unhealthy weight gain. Problems with obesity as a young adult can lead to chronic disease later in life (Hebden et al. 2013). Participants were recruited through Australian General Practice clinics. The programme itself used a combination of text messages, email, smartphone apps, a website and coaching calls. All participants had access to the website element of the programme, though resources available differed for intervention and control group participants. The website gave intervention participants access to information about meal planning and physical activity, a weight tracker, smartphone applications for self-monitoring and a community blog. Text messages were tailored based on the Transtheoretical Model of Health Behaviour Change (Prochaska et al. 1992) and gender. Coaching calls were used to set goals for participants, and text messages were also personalised based on these goals. Smartphone apps were specifically designed for the intervention for assisting with behaviour change, targeting physical activity, fast food, sugary beverage and fruit & vegetable consumption (Hebden et al. 2012). Emails were used as reminders and as a source of information on physical activity and self-monitoring etc. A 2016 evaluation of the TXT2BFit intervention reported that coaching calls, text messages and emails were very popular with participants, who found them helpful. Conversely, participants disliked the website and smartphone apps, and reported that these would have been more effective if they were incorporated into one single application that could be personalised (Partridge et al. 2016).

2.7.8 Chatbots

Chatbots are often described as “intelligent” systems, on the basis that they are capable of simulating conversation and can react to human emotions. They may be enabled for *Machine learning (ML)*, the ability of a computer system to learn from its experiences with no additional programming. The world’s first true chatbot was called ELIZA, developed in 1966, by Joseph Weizenbaum, a German American computer scientist at Massachusetts Institute of Technology. ELIZA was programmed to respond to user input through stock phrases and repetition. For example, a statement such as “My head hurts” would generate the response “Why do you say your head hurts?” Although primitive,

ELIZA paved the way for what today is known as conversational AI (Knowledge@Wharton, 2016). Wharton University of Pennsylvania described 2016 as "the rise of the chatbots", and revealed that Microsoft, Google, and Facebook are actively embracing this technology. Perhaps the most well-known chatbots are Apple's *Siri*, Microsoft's *Cortana* and *Google Home*. At the 2016 Google Developer conference the company unveiled *Google Allo*, a smart messaging app. The goal of chatbot developers is to appear as human as possible, so the user feels at ease, as though talking to a human, rather than interacting with intelligent software spouting pre-programmed responses (Knowledge@Wharton, 2016).

Versatile chatbots have the potential to solve the problem of what can be accurately described as "app fatigue". Technology firm *Personetics* published a 2016 white paper, *Bot Brief, How Chatbots fit into your omnichannel strategy*, which suggests that most smartphone owners use just a handful of apps on a regular basis and estimates that around 25% of apps are downloaded, used once, and then discarded. Personetics found that the use of instant messaging apps is soaring, with approximately 2.5 billion people worldwide using at least one instant messaging app, a figure that is predicted to rise to more than 3 billion in the next few years (Personetics, 2016).

2.7.8.1 Chatbot Intelligence

2.7.8.1.1 What is intelligence?

British Psychologist Charles Spearman first discussed intelligence in 1904 in his theory of *General Intelligence* (Cherry, 2019b). Spearman suggested that while intelligent individuals who excelled in one area (for example mathematics or science), were likely to do well in other areas and he proposed the notion that intelligence could be measured using a single scale known as the *g-factor*. Spearman's theory has been criticised by psychologists such as American Howard Gardner, who argued that there are several *types* of intelligence and that it is thus not possible to measure intelligence using a single factor (Cherry, 2019a). His1983 *Theory of Multiple Intelligences* proposed that there are eight main types of intelligence: *Visual-Spatial (visual and spatial judgement e.g. reading maps, giving directions), Linguistic-Verbal (words and language e.g. reading, writing), Logical-Mathematical (analytical/mathematical skills e.g. problem solving), Bodily-Kinaesthetic (physical movement/coordination e.g. dancing, sports), Musical (sense of rhythm and music e.g. playing/singing), Interpersonal (understanding/relating to other*

people e.g. conflict resolution, communicating), Intrapersonal (self-reflection skills), Naturalistic (relating to nature e.g. botany, exploring the outdoors) (Cherry, 2019a). Gardner's theory showed strong similarity to Louis L. Thurstone's *Theory of Primary Metal Abilities*, which suggested that intelligence is based on seven primary mental abilities: *Verbal Comprehension, Reasoning, Perceptual Speed, Numerical Ability, Word Fluency, Associative Memory and Spatial Visualisation* (Cherry, 2019c).

2.7.8.2 How are chatbots intelligent?

The idea of "intelligent" computer systems was first proposed in Alan Turing's 1950 paper *Computing Machinery and Intelligence*, in which he discussed "the imitation game". This game involves three "players", a man, a woman and a judge of either gender, who are placed in separate rooms and able to communicate only through a text-based computer interface. The judge's role in the game is to ask specific questions of the other two participants, to try and determine which participant is male, and which is female. The male participant must help the judge, by answering honestly, however the female can attempt to deceive the judge by giving false or ambiguous answers (Turing, 1950). Turing expanded this game further by replacing one of the participants with a computer. He stated that a computer can be considered "intelligent" if the judge is convinced that the computer is human more than 50% of the time. Today this is known as the "Turing test", and it is used to determine how close artificial intelligence comes to actual human intelligence.

However, based on the theories discussed in section 2.6.8.1 above, there is more to intelligence than a mere ability to engage in conversation (like the ELIZA bot) or convince someone that you are human! In 1980, the philosopher John Searle used the *Chinese Room Argument* to propose that computers are incapable of true "understanding", regardless of how human they may appear to be (Searle, 1980). His argument proposes a scenario in which he is locked in a sealed room and is passed messages written in Chinese. Searle asserts that he can neither speak nor read Chinese, but suggests that, if provided with clear instructions, in English, for interpreting messages written in Chinese, he should in theory be able to devise responses to these messages which are indistinguishable from those written by a native Chinese speaker (Searle, 1980). Searle suggests that although he has been provided with all the necessary tools to interpret and respond to messages written in Chinese, his actual *understanding* of the language has not changed. Searle's argument is that while it is possible for computers to understand the

rules for interacting with human operators, they lack the *semantic* knowledge of the human mind and thus can never truly "understand" the way a human can (Searle, 1980).

In March 2016, Microsoft released Tay, a Twitter-based chatbot. Tay was intended to engage with users through tweets, in the style of a teenage girl (Schwartz, 2019). Within 24 hours of "her" launch, however, Tay had advanced from tweeting harmless banter and lame jokes to posting content that was abusive, racist and misogynistic. It emerged that online "trolls" had exploited Tay's Machine Learning (MI) capabilities by asking the bot to repeat offensive words and phrases (Schwartz, 2019). Tay was designed to learn from everything "she" was asked (or told to repeat), and so in the end learnt to make statements which could be considered offensive. Tay was "intelligent" enough to learn new words and phrases from interactions with humans but lacked enough *understanding* to recognise when something was likely to be considered offensive (Schwartz, 2019).

Chatbots are indeed "intelligent" systems, but only insofar as they can engage in conversation with their users. They are not yet able to truly understand human emotions, the condition of being "alive", or the complex nature of human relationships. It is potentially unethical for chatbots to "pretend" or "convince" the user that they have these capabilities, and this will be discussed further in **section 2.6.8.4.1** below.

2.7.8.2 Chatbots in Healthcare

There is evidence to suggest that chatbots may be of use in therapeutic interventions. A recent paper by Barak et al. (2009) reported an association between chatbots or "chatterbots" and higher compliance with therapeutic treatments, where the chatbot has been programmed to detect and react to emotions and create a sense of empathy with the user (Barak et al. 2009). A 2017 paper by Fadhil & Gabrielli proposed a chatbot system for weight gain prevention in primary care adults. The authors identified clear benefits of chatbots for this purpose – increased user engagement, automated personalised feedback that may otherwise need to come from a human respondent, and high levels of usability for different target demographics (Fadhil & Gabrielli, 2017). Chatbots may be of interest to those individuals wishing to use existing technologies for weight loss maintenance but are reluctant to download or install new apps for the purpose (a stance that is consistent with the *Personetics* discussion on app fatigue).

Chatbots have shown potential for mental health management and coaching. Fitzpatrick et al. (2017) reported that almost three quarters of all mental health problems in the US are diagnosed before the age of 24, and university students are most at risk, yet most of those students affected do not attempt to access mental health services (Fitzpatrick et al. 2017). It is suggested that the stigma associated with mental health problems may discourage individuals from accessing services they really need (Fitzpatrick et al. 2017). Internet and smartphone-based mental health services have emerged as a popular means of overcoming stigma, and these systems have been seen to be effective for managing mental health and have enjoyed high demand from patients (Fitzpatrick et al. 2017). In spite of this however, there has been generally poor adherence to interventions of this type, and it is believed that this is due to the fact that these systems do not offer the level of human contact offered by traditional in-person services (consistent with Donaldson et al. 2014). The *Woebot* chatbot was developed as a conversation-driven agent for Cognitive Behavioural Therapy (CBT), and was deemed potentially effective for delivering mental health interventions.

A similar intervention was developed by Cameron et al (2018) in partnership with Inspire Wellbeing Northern Ireland. iHelpr is a mental health chatbot that facilitates self-help through provision of self-assessment tools and advice (Cameron et al. 2018). iHelpr has been designed to work in partnership with the Inspire Workplaces programme which seeks to promote wellbeing and build employee resilience. Cameron et al. suggest that a chatbot will be a suitable platform for delivering this type of intervention because users will spend less time accessing services and functions (Cameron et al. 2018).

Based on this research there is evidence to suggest that chatbots are effective for delivering healthcare interventions as they can improve participant engagement, providing 24/7 automated feedback, and simplifying the process of accessing services. Chatbots present specific limitations however, not least in terms of ethics. Limitations are discussed in **section 2.7.8.4**.

2.7.8.3 Chatbots & Weight Loss Maintenance

Given the popularity of social media and messaging apps, and advances in conversational AI, it seems logical to investigate the effectiveness of conversational bots for supporting weight loss maintenance. It has been shown that there is limited research into the use of

digital technologies for weight loss maintenance, and that most of the research to date uses older systems, such as internet sites, email, or text messaging. Considering the recent statistics from Ofcom and UKOM showing a trend towards greater use of social media, it makes sense to design digital health technologies to take advantage of this trend (Ofcom 2018, UKOM, 2018).

Given the increase in "app fatigue", it would be wise to use systems that can be built on existing technologies (such as Facebook Messenger). Findings from the Lifestyle Eating and Activity Programme (LEAP) Beep trial (Donaldson et al. 2014) suggest that simple SMS-based systems are popular with users, since the system is easily available to them and does not require installation of any new apps or logging into a website. Chatbots are casual and conversational in nature, which makes them attractive for use in situations where the system needs to interact with a human. It is, however, important to ensure that responses are carefully designed so they do not feel too "generic" or impersonal, as this may be off-putting for the user (Personetics, 2016). When designing the chatbot conversation dialogues, it will be prudent to consider attitudes to message tones previously explored in focus group studies (Pollard et al. 2016, Stephens et al. Smith et al. Woolford et al.).

2.7.8.4 Limitations of Chatbots

Although chatbots present a natural, lightweight interface to the user, eliminate the need to learn a complex user interface, and permit 24-hour access to services, they are not without their limitations. Chatbots are, first and foremost, a form of artificial intelligence (AI), and, while they can effectively emulate basic human conversations, AI is still not yet advanced enough to accurately model more complex conversations. Chatbots will likely struggle to accurately convey characteristically human concepts such as empathy or compassion (Cameron et al. 2018). As will be seen later in this thesis, chatbot messages intended to convey empathy or encouragement need to be subtle, in order to avoid overdoing sentiment, and thus becoming less acceptable to the user. Chatbots may also be unable to identify and effectively respond to risk. In 2018, BBC news reported that two mental health chatbots (*Wysa* and *Woebot*) were unable to spot signs of child sexual abuse and failed to advise victims to seek help in this situation. Both chatbots also struggled to correctly identify drug abuse and eating disorders and offer appropriate

advice (White, 2018). It is important to remember that, regardless of how capable, no chatbot can ever replace a qualified human specialist (Cameron et al. 2018).

2.7.8.4.1 Chatbots and Trust

The use of chatbots creates specific ethical issues, particularly regarding trust. For example, it is potentially unethical to attempt to convince a user that they are in fact talking to a human, or to replace a human specialist. Users may feel deceived or cheated if they discover that they thought they were talking to a human when in fact they were talking to a chatbot (Needleman, 2017), as was the case for American psychologist Dr Robert Epstein. As reported by the BBC's Tim Harford, in 2006 Dr Epstein began an email exchange with a Russian woman, who appeared to be friendly and eventually confessed to developing feelings for him. Epstein observed that the woman, who claimed her name was Ivana, never directly answered his questions, and he eventually determined that "Ivana" was in fact a chatbot (Harford, 2019). There have been reported cases in which chatbots have appeared to encourage users to do things they would not normally do (Cameron et al. 2018). Chatbot design should always adhere to established ethical standards. Mulvenna et al. (2017) outline an "Ethical by Design" manifesto, which recommends creating empathy, permitting informed, shared decision making, respecting the right to choice, balancing privacy and security with global access, integrating and supporting policy, complementing a wide range of needs and abilities while actively challenging inherent biases and values, striving for sustainable designs, adequately handling failures, being realistic and supporting the system or service (Mulvenna et al. 2017).

A 2018 study explored the risks posed by using natural language agents (NLAs) as a source of medical information. The authors state that confidence in this type of system is increasing and suggest that Amazon alone offers 78 medical add-ons for the Alexa platform (Bickmore et al. 2018). NLAs are not flawless, nor will they be in the recent future, and users may overestimate their capabilities. Over-reliance on NLAs to access medical information without guidance from medical professionals may pose risks to health (Bickmore et al. 2018). Participants in this study (n=54) used three of the most popular conversational assistants, Amazon Alexa, Siri and Google Assistant to request answers to three types of medical question. Questions included one query about any medical subject of the participant's choosing, one medication query, and one emergency

query. It was observed that 57% (n=226) of all tasks resulted in failure, and of the completed tasks (n=168), 49 resulted in harmful advice, and 27 resulted in advice that could potentially have led to death. The authors concluded that conversational assistants should not be used for medical advice without consulting a medical professional (Bickmore et al. 2018).

2.7.8.4.2 Chatbot Abandonment

Chatbots are becoming a popular tool for customer service and in healthcare, however they are at high risk of user abandonment (Kyselova, 2018). In 2017, Facebook suggested that the failure rate of Messenger-based chatbots was 70%. Companies using Facebook Messenger bots have reported that Messenger based chatbots are complicated to use, and often offer insufficient personalisation. If users are to engage with chatbots then it is vital that they be understood, and if users become frustrated by chatbots which seem unintelligent they will often discard them after one use. Regardless of how popular chatbots may be, to a degree they are still not considered suitable replacements for humans, and limitations of chatbots may lead to frustration. It is vital therefore to carefully design the WeightMentor chatbot to use a platform that will be user friendly, consider and design carefully the functions that the chatbot should provide to its users so they are relevant, useful and appropriate, and to ensure that the interactions with the user are highly personalised.

2.7.9 Limitations of Technology

Digital technologies have enjoyed varying degrees of success as weight loss interventions, but they are not without their limitations. A review by Gasch et al. (2016) suggested that such studies are limited by the sample size, and further research is needed to determine how interventions of this type may be scaled to suit larger populations. This suggestion has been echoed by Evans et al. (2015). Gasch et al. (2016) raised concerns regarding confidentiality of personal data and accuracy of information. Donaldson et al. (2014) stated that personalised feedback often requires analysis and input from a human operator. This limits times when personalised feedback may be available such as during the night and at weekends. It is proposed by the authors that research into automated feedback would be warranted, considering techniques that may be used to automate feedback without losing the human touch (Donaldson et al. 2014). All the trials identified above use older forms of technology. However, Ofcom’s 2018 market report revealed

that “traditional” mobile messaging use has declined thanks to smarter devices and an increase in the use of social media platforms such as Facebook, WhatsApp and Snapchat (Ofcom, 2018). Text message sending was highest in 2012 (an average of 162 messages sent per contract per month), but by 2017 had roughly halved (82 messages sent per contract per month). The same report suggests that Facebook is the most visited social media platform in the United Kingdom (Ofcom 2018). The UKOM 2018 UK Digital Market Overview suggests that smartphone usage accounts for 78% of all online usage, with highest usage reported among 18-24-year olds (UKOM, 2018).

2.8 Usability

2.8.1 What is Usability?

Usability may be defined as *appropriateness to a purpose*, and the usability of a system is often relative to the context in which it is being used (Brooke 1996). A *usable* system will therefore be appropriate to the purpose for which it is designed and will facilitate users in completing required tasks. All electronic systems, whether smartphone-, web- or computer-based, require some sort of user interface. As technology becomes more affordable and accessible to wider sections of the community, user interfaces must also maintain levels of simplicity and intuitiveness that match users’ expertise. Human Computer Interaction (HCI) studies the relationship between humans and computers and suggests strategies that may be implemented to facilitate easy interaction. HCI best practices often define clear rules that should be followed when designing a user interface, such as Ben Shneiderman’s Eight Golden Rules (Shneiderman et al. 2016) or Nielsen’s Usability Heuristics (Nielsen, 1994), which are both discussed in chapter 4.

2.8.2 How can usability be measured?

The usability of a system may be measured in several ways, the simplest of which is *user testing* which involves observing human users conducting representative tasks (Nielsen, 2012) and recording areas where they succeed and where they fail. Usability testing provides insight into what users think of the system, whether the main system functions are usable, and the extent to which the system will be suitable for its intended use (Foggia, 2018). There are two main types of usability testing, formative and summative.

2.8.2.1 Formative Usability Testing

Formative usability testing usually happens as part of an iterative design process. It identifies aspects of the system that will not work (Sauro, 2016), and seeks to improve design by considering ways in which it can be improved (Travis, 2006). Formative testing involves observation of users in a real-time scenario, to record how the user interacts with the system and any difficulties they encounter (UX24/7, 2017). This type of testing generally collects qualitative data, such as thought processes of participants, opinions on the user interface, and so on.

2.8.2.2 Summative Usability Testing

Summative Usability Testing is concerned solely with quantifying the usability of a system (Sauro, 2016). It usually takes place at the end of a design process, and generally collects quantitative data, as it reflects what has happened during the test (UX24/7, 2017). Users will be observed in real time, and usability will be assessed using metrics such as time taken to complete tasks and number of tasks completed. Summative usability testing will also attempt to quantify usability with some form of rating scale, such as the System Usability Scale (SUS), designed by British computer consultant John Brooke in 1986 as a "quick and dirty" means of assessing usability (Brooke, 1996). SUS will be covered in **Chapter 5**.

2.9 Groups Under- or Over-represented in the Literature

Based on the literature there were two main groups that were observed to be underrepresented. Male participants were observed to be underrepresented in comparison to female participants, and it was suggested that this may be because women are often more likely to access health advice and attempt weight loss than men (Fjeldsoe et al. 2016). It was observed that while older participants are equally likely to participate in weight loss maintenance interventions as younger participants, older participants were reported as being more likely to complete such interventions (Donaldson et al. 2014).

2.10 Gaps in the Literature

Based on this literature review, it is known that technology for healthy living, behaviour change and weight management is a popular topic of contemporary research, that a wide

range of technologies have been trialled for these purposes, and that technology has been shown to positively affect weight loss maintenance in the short term (between 3 and 24 months). It is not known how technology affects weight loss maintenance in the long term (over 24 months), whether chatbots may be effectively used to automate personalised messages and feedback in support of weight loss maintenance, or whether chatbots may be effectively utilised as a delivery method for weight loss maintenance interventions.

2.11 Recommendations for Further Research

Based on this review of the literature on technology for health and weight loss maintenance, the following recommendations for further research have been made:

- To investigate the effectiveness of technology for weight loss maintenance in the long term (over 24 months).
- To investigate how chatbots may be used to effectively automate personalised feedback and messages
- To investigate the extent to which chatbots may be used as a delivery method for weight loss maintenance interventions.

2.12 Conclusion

Technology for weight loss is a popular research topic, and although there have been numerous studies into this type of technology, many of them are underpowered or have not progressed beyond the pilot stage. More robust clinical trials are limited but have shown that technology may have potential for supporting weight loss in the short term or on a small scale. Research into the use of technology for weight loss maintenance is limited, and existing technologies have been observed to be effective mainly in the short term and are less effective in comparison with traditional face-to-face coaching techniques. If health authorities are to successfully tackle the global epidemic of obesity, then innovative methods must be tested and employ the use of contemporary technologies. This is never more important than in the present, where we have seen a significant shift away from the use of email and internet for communication, in favour of contemporary technologies. Chatbot based systems are gaining prevalence for their simplicity and casual interaction with their users but pose specific ethical issues which should be suitably addressed if they are to be used in any intervention. There has not yet been any solid research into the effectiveness of chatbots for weight loss maintenance.

This research project will evaluate the effectiveness of these systems through the development and testing of a custom-designed chatbot system for weight loss maintenance.

Chapter 3: Needs Analysis

3.1 Introduction

A needs analysis is conducted in order to determine the needs of a specific group, or to facilitate development or evaluation of services (Tutty & Rothery, 2010). A needs analysis may be conducted for several reasons, such as informing policy and service development or improvement, assessing satisfaction or choosing the best solution from several possibilities (Stabb, 1995). A needs analysis determines the gap between the status quo and how things "should" be including any possible reasons for such a gap and possible strategies for closing this gap (Tutty & Rothery 2010). Any chatbot designed for supporting weight loss maintenance will be most effective if it directly targets the needs of individuals who are trying to maintain weight loss and provides them with feasible, simple strategies for achieving their goals.

It is well documented that obesity is a significant problem that will not be resolved overnight (WHO, 2019, NHS 2019). There is no "one size fits all" approach to obesity management and weight loss maintenance, and it is important for individuals to develop a strategy that works for them, harnessing their strengths. In order to effectively determine *what* the chatbot should do and *how* it will help the target users, it is necessary to conduct a needs analysis using qualitative data collection methods to gather information on participants' weight loss experience, anticipated challenges to maintenance, proposed solutions to these challenges, and opinions on the use of weight loss maintenance apps.

3.2 Aim

To investigate the needs of individuals maintaining weight loss to inform the development of a weight loss maintenance chatbot.

3.3 Methods

3.3.1 Research Design

The research design employed involved a qualitative needs analysis, which had the aim of gathering participants' attitudes to weight loss maintenance and presenting these in a structured way in order to determine how the chatbot may effectively address the needs

of target users. Conducting a needs analysis is an essential part of this research for several reasons. It will provide insight into the experiences, motivations and difficulties faced by individuals maintaining weight loss, quantify their specific needs and determine how these may be met. The needs analysis will determine if apps are an acceptable solution for supporting weight loss maintenance and the extent to which users may benefit from their use, and whether individuals would consider using a chatbot to achieve their weight loss maintenance goals.

A recent review of Nutrition-related smartphone apps, “Popular Nutrition-Related Mobile Apps: A Feature Assessment" (Franco et al. 2013) identified the main features of the most popular nutrition mobile apps in Google and Apple stores and was used along with interview findings to inform chatbot development. Tutty & Rothery (2010) suggested that needs analysis may be either quantitative or qualitative. Tutty & Rothery (2010) identified four main methods for qualitative data collection during a needs analysis, but the most used are focus groups and interviews. Interviews may be used for gathering information on a one-to-one basis with participants using a predetermined process (Tutty & Rothery, 2010). Focus groups adopt a less structured approach than individual interviews but allow for discussion of the topic as a group. Focus groups seek to determine a consensus viewpoint, whereas interviews identify individual viewpoints.

Each of these methods has its own advantages and disadvantages. Focus groups are suitable for groups of between eight and twelve participants, and, while it is possible in such a setting for participants to react to each other and openly discuss their ideas the group coordinator must be effective in ensuring that discussion keeps pace and does not follow unrelated tangents. Groups do not easily allow for in-depth discussion of participant feelings or experiences and depend on a positive relationship between participants. Participants may also be reluctant to discuss sensitive personal issues in a group setting. Interviews permit deeper levels of discussion of sensitive issues that may not be comfortable for participants in a structured group setting. However, interviews are more time-consuming than focus groups and work best with a small sample size (Tutty & Rothery 2010). After consideration of the strengths and weaknesses of each of these methods, it was determined that semi-structured interviews would be used during the needs analysis, as it was anticipated that only a small sample would be required to reach data saturation (see section 3.2) and participants would potentially feel able to be honest and open about their experiences without being overhead by others.

3.3.2 Participants and Sample Size

Participants were adults (over 18 years) who had either just lost weight or had previously lost weight but regained it, interested in maintaining weight loss (or losing the weight they had regained), and had a reasonable understanding of the English language. The inclusion and exclusion criteria are summarised in **Table 3.1**. Participant demographics are discussed in section 4.1.

Table 3.1: Needs Analysis Participant Selection Criteria

Inclusion criteria	Exclusion criteria
Adults over 18 years	Not providing consent
Lost weight or had previously lost weight but regained it	
Interested in maintaining weight loss (or losing the weight they had regained)	
Reasonable understanding of the English language	

Purposive sampling was used to select interview participants. This type of sampling is based on the personal judgement of the researcher, who selects specific individuals from a population in order to achieve the research aim. This type of sampling is most effective in cases where only limited numbers of people in a population may provide enough data (Dudovskiy, 2018). For this needs analysis, it was necessary to interview only those individuals who had successfully lost weight and were maintaining it, or who had previously lost weight and had since regained it. Purposive sampling was used to select individuals from the population who fitted into these categories. It was determined that a minimum of five participants should be interviewed until saturation was reached. Saturation in qualitative research is the point where successive data collections (interviews in this needs analysis) do not yield any useful new data, or where the topics for discussion have been sufficiently represented by data that have already been collected (Saunders et al. 2017). Data saturation was reached after 15 participants.

3.3.3 Recruitment

Participants were recruited from staff and students at the Ulster University and from family and friends of the researcher. A recruitment email (see **Appendix 4**) was circulated to university staff and students via the university mailing list. A similarly worded email was also circulated to other PhD researchers and staff in the Faculty of Arts, Humanities and Social Sciences by the Faculty Administrative Officer. Staff and students

who were interested in taking part contacted the researcher at his university email address and requested a copy of the Participant Information Sheet (PIS), which is in **Appendix 5**. Family and friends of the researcher who were interested in participating requested a copy of the PIS, which was forwarded to them by email or Facebook Messenger. Individuals who met the inclusion criteria and were willing to participate signed two copies of the consent form at the start of the interview. One copy was retained by the participant and the other was filed by the researcher. The consent form is in **Appendix 6**.

3.3.4 Data collection method – Semi-structured interviews

Weight loss is highly personal in nature, and what works for one individual may not be most effective for others. Additionally, individuals who have struggled to lose and maintain weight may be reluctant to discuss their experiences in a group setting. It was determined that individual interviews would be the most appropriate data collection method for this needs analysis, and that these interviews should be semi-structured. Semi-structured interviews are designed to be highly flexible, creating scope for the interviewee to discuss their answers in detail, while also permitting the interviewer to exercise their own judgement as to the degree of freedom granted to the interviewee. A more spontaneous and relaxed procedure may result in more honest, and often unexpected answers from participants (Kvale, 2007). For this needs analysis, topics for discussion were identified, and these were developed into structured questions and associated probes for use during the interviews. A full interview schedule listing questions and their associated probes is in **Appendix 7**. Participants were asked ten questions during the interview. Some examples of questions and probes are listed in **Table 3.2.**

Table 3.2: Example Needs Analysis Interview Questions

Topic	Question	Probes
Future goals	Congratulations on your weight loss! Do you have any plans for maintaining this weight loss?	Are you going to maintain it? Do you want to maintain it long term? How long for? Do you have any weight loss maintenance goals and why? Do you want to lose further weight? Are you trying to keep the weight off for a special event or function?

Topic	Question	Probes
Challenges	What challenges do you consider you will experience in maintaining weight loss?	How easily tempted are you? When are you most likely to be at risk? Do things like stress, tiredness, or emotions affect your behaviour?
Personalised messages	Do you consider that a personalised message would motivate you to try and maintain weight loss?	How would you personalise the message? Who would the message come from? If you knew you were getting a message from a virtual person, how would that make you feel?

3.3.5 Procedures

Interviews were conducted in the Jordanstown and Belfast campuses of the Ulster University. One participant was unable to travel to the university and was interviewed using a room provided by Magheralin Parish Church, as this was the most convenient location for the participant. Interviews lasted approximately 30 minutes and consisted of three parts. Part one discussed goals for future weight loss or maintenance and strategies for achieving these, challenges and solutions and opinions on app use. During part two, participants were asked for their opinion on personalised messages and shown a handout identifying different types and tones of messages, which they were asked to comment on. Part three was a demonstration of the WeightMentor prototype chatbot. Participants were given a briefing before the interview started, which explained the purpose of the interviews, what participants would be asked to talk about, and how the data collected would be used. The text of the briefing is in **Appendix 7** as part of the interview schedule. Participants who had consented to take part completed a participant fact sheet, which is in **Appendix 8**, and collected demographic information about the participant (name, age range, gender, height, weight, BMI, weight loss and weight loss method).

3.3.5.1 Interview part 1: Future goals, strategies, challenges and solutions

This part of the interview consisted of six questions. Participants were first asked to identify goals for future weight loss or maintenance, such as preparing for a holiday or other event, fitting into clothes, improving fitness and health, or attempting to return to a previous weight following weight gain, and suggest strategies they were using or planning

to use in order to achieve these goals, such as joining a slimming club. Participants were then asked to consider the sort of challenges they would face, such as temptation or stress and suggest ways to overcome these challenges, e.g. social accountability, keeping active etc. Thirdly, participants were asked about their use of apps for maintaining weight loss, which apps were available, and which app functions were most or least helpful and why.

3.3.5.2 Interview part 2: Personalised messages

During this part of the interview, participants were initially asked to provide their views regarding personalised messages and encouraged to think about how they would personalise a message, if there was any particular person (e.g. a friend or family member) from whom a personalised message would be particularly helpful, and what they thought of the possibility that a message could come from a "virtual" person. Participants were then shown a handout (see **Appendix 9**) which outlined five different types of message (requests for feedback, reflection, generic messages, tailored messages and tips) and five different message tones (authoritative, empathetic, congratulatory, commiserating and encouraging), based on research by Woolford et al. (2011), Smith et al. (2014) and Pollard et al. (2016).

3.3.5.3 Interview part 3: Chatbot demonstration

The final part of the interview consisted of a demonstration of the WeightMentor chatbot prototype. At this stage, the chatbot was still in development, and although the main functions were present, responses were still limited, and variability and Smalltalk (see **section 4.3.4.2.** in chapter 4) had not yet been implemented. The purpose of the demonstration was to give needs analysis participants an idea of what the finished chatbot might be like and how it would function, to determine how participants would respond to such a chatbot and what they thought of it. The researcher used a laptop to demonstrate the chatbot's main functions (initial greeting, self-reporting, reviewing self-reporting trends, motivation, goodbye message) and discussed how it would benefit potential users. Needs analysis participants were asked to indicate whether they liked the *WeightMentor* chatbot, they would use it (and how often) and whether they felt it would be beneficial for weight loss maintenance.

3.3.6 Data Analysis

Interviews were transcribed and analysed using thematic analysis to identify key themes from the interviews. Thematic analysis identifies and organises patterns in qualitative data and was chosen for this needs analysis study because it allows a wide variety of topics to be analysed, permits in-depth exploration of data and maintains flexibility of interpretation of data (Castleberry & Nolen, 2018). Phrases that were of relevance to the research were extracted from the transcripts and recorded using post-it notes. A tally of each phrase was kept using an Excel spreadsheet and once all key phrases were identified, these were transposed into themes. Key themes were recorded in a table along with the number of occurrences and a collection of quotes related to each theme. This coding table is in **Appendix 10**. Nvivo 12 was used to analyse the interview transcripts as part of the structured thematic analysis process. Transcripts were read in order to identify topics which occurred more than once. These topics were used to form codes. A total of thirty-seven codes were identified from the fifteen transcripts. These codes were combined into a list of ten sub-themes, which were finally combined into a list of five verbose themes. These five themes were later refined to make them shorter and more to the point. Finally, the transcripts were read again to identify quotes that supported each theme.

3.3.7 Ethics

As for all studies within this PhD project, ethical approval was obtained from the Communication Ethics Filter Committee and University Research Ethics Committee, and the study was conducted in compliance with the Ulster University Code of Practice for Professional Integrity in the Conduct of Research and the Policy for the Governance of Research Involving Human Participants. The study was designed to comply with the five core principles of beneficence, non-maleficence, honesty and integrity, confidentiality and informed consent. Interviews were conducted in a way that provided a positive, friendly environment in which participants could feel free to discuss as much or as little as they felt they wanted to, and participants were at no point forced to discuss topics that they were uncomfortable with. If any participant preferred to skip a topic or not answer a question, they could request to do so. If, at any point, a participant had become upset or distressed by a discussion topic, the interview would have been stopped. Participant confidentiality was respected throughout the study. Where it was necessary in an interview to refer to anything said by a previous participant, this was done without naming the participant or stating anything else that could make it possible to identify the previous

participant. Informed consent was obtained from all participants prior to participation, by asking participants to read the PIS (see **Appendix 5**) which covered the purpose and aim of the study, the number of participants required, what participants would be expected to do, how data would be collected and used, and how participant anonymity and confidentiality would be respected. Participants signed a consent form before commencing their interview, on the basis that they had read this PIS. Personal data relating to participants were collected and stored in compliance with GDPR 2018.

3.4 Findings

3.4.1 Participant Demographics

Participant demographics are summarised in **Table 3.3**. All participants (n=15) were adults (aged 18+) who had either lost weight and were now maintaining weight loss or had previously lost weight but since regained it. Overall, there was almost the same number of male participants (n=7, 47%) as female participants (n=8, 53%). The median age range was 41-45 years and the mode was Over 50 years. All participants had lost weight though some had regained it, mean weight loss was 11.14±5.29kg and median was 10kg. Mean participant weight overall was 82.00±27.05kg (BMI 27.61±6.89kg/m^2), and the median was 68.5kg (25.3kg/m^2). The shortest weight loss duration was 2 months, the longest was 24 months, the mean was 6.50±5.89 months and the mode duration was 3 months (5 participants, 33.33%). Participants had used a range of methods for weight loss and maintenance, with "self-regulation" the most used method.

Table 3.3: Needs Analysis Participant Demographics (n=15)

ID	Interview No.	Age (Years)	Gender	Height (m)	Weight (kg)	BMI (kg/m^2)	Weight Loss (kg)	Duration (months)	Method
P01	I01	26-30	M	1.88	84.0	23.8	10.0	3	Self-Regulation
P02	I02	Over 50	F	1.57	63.5	25.8	16.3	12	Slimming World
P03	I03	Over 50	F	1.68	62.0	22.0	4.0	6	Self-Regulation
P04	I04	41-45	F	1.52	55.0	23.8	6.0	6	Self-Reg./Diet Plan
P05	I05	Over 50	M	1.80	150.6	46.5	16.8	3	Self-Regulation
P06	I06	36-40	F	1.78	79.0	24.9	6.0	4	Weight Watchers

ID	Interview No.	Age (Years)	Gender	Height (m)	Weight (kg)	BMI (kg/m^2)	Weight Loss (kg)	Duration (months)	Method
P07	I07	26-30	M	1.78	86.2	27.2	20.0	3	Weight Watchers
P08	I08	26-30	M	1.80	88.0	27.2	7.0	2	Self-Reg./Diet Plan/Exercise
P09	I09	18-25	F	1.63	68.5	25.8	7.7	6	Self-Regulation
P10	I10	41-45	F	1.60	63.6	24.8	6.4	2	Self-Regulation
P11	I11	41-45	F	1.80	101.6	31.4	7.5	3	Slimming World/Self Reg./Couch to 5K
P12	I12	36-40	F	1.68	65.8	23.3	19.5	12	Weight Loss Apps
P13	I13	41-45	M	1.71	65.0	22.2	14.0	24	Self-Regulation
P14	I14	Over 50	M	1.63	67.1	25.3	12.7	9	Diet Plan
P15	P15	Over 50	M	1.80	130.2	40.2	13.2	3	Diet Plan

3.4.2 Key Themes

From the thematic analysis, five key themes were identified, these were: "*Weight loss maintenance is a challenge", "Social contact can be a double-edged sword", "Apps are popular"," Personalised messages are more useful" and "Chatbots have potential for weight loss maintenance"*. Themes are summarised in **Table 3.4**. Full details of codes and sub-themes identified during thematic analysis can be found in the coding table in **Appendix 10**.

Table 3.4: Key Themes from Interviews

Key theme	Quotes
1. Weight loss maintenance is a challenge	*"If I'm hungry I sometimes make the wrong choices"* *"I don't drink, I don't smoke but I feel I have as much addiction to say crisps, or chocolate"* *"if I don't have my breakfast in the morning, by the time I come to college and I go through the mall and I smell the bacon, I want bacon!"*
2. Social contact can be a double-edged sword	*"If someone knows [you're losing weight] they'll invite you out less, but sometimes people say, 'It's only one drink', 'it's only one chocolate bar'. That's not the point!"* *"It's not really an issue for [friends who aren't overweight], but at the other end of the scale, people who are overweight look at you and think you have nothing to worry about."* *"With exercising, sometimes I'll go to the gym with a friend and we'll do the same activity together, having that social support definitely helps."*

Key theme	Quotes
	"If you are [at work events] they are always throwing everything at you to get you drunk"
3. Apps are popular	*"Sometimes it's just, can you be bothered inputting the data…"* *"It takes a lot more time to input manually and stuff. It was just laziness and I got out of the way of it"* *"It is the fact that it is connected to everyday living, whereas the phone is a little bit remote"* *"I didn't like it because you had to say how much you were eating, how much calories they ask, yeah. So many things you had to do."*
4. Personalised messages are more useful	*"Tailored messages...feel like they relate to what you are doing"* *"why would I want to go and download an app just to get patronizing or random messages that have nothing to do with me. If you are going to get me to download an app you need to make it specific for me"* *"I think if it just, if an authoritative one just keeps telling you stuff and you are just like, leave me alone, stop being annoying"* *"If it was too cheesy, I'd just throw it away!"* *"I would feel a bit patronised by an insincere message…"*
5. Chatbots have potential for weight loss maintenance	*"I would use it because it is engaging and to me it seems a lot more fluent how to enter data"* *"I liked it because it keeps me accountable and would be a good way to keep people motivated"* *"I think there are potentially too many questions…I think someone could potentially get frustrated."* *"I liked the humour of the hello and goodbye…I think as well if you could create some kind of 'character', that might be quite useful.*

3.4.3 Description of Key Themes

3.4.3.1 Key Theme 1: Weight loss maintenance is a challenge

Participants acknowledged that weight loss maintenance is a challenge, and the key reported significant challenges were stress, emotion, tiredness and temptation. Participants admitted that often comfort eating was their default response to these challenges, however this can be highly addictive and lead to compulsive behaviour. Participants stated that a regular routine helps provide a distraction to mitigate challenges and reduces the risk of unhealthy behaviours such as snacking.

Participants who had regained weight were particularly mindful of the difficulty of weight loss maintenance and acknowledged the need for proactivity. For example, "*If I could maintain exercise and my diet, I'd lose all the weight again!"* (Participant #7). Participants acknowledged that the convenience of certain foods was a strong influencer, especially if they were tired. "*Sometimes you don't want to do any cooking when you get*

home, so you order a pizza" (Participant #13). They linked emotion and stress with unhealthy eating habits, for example, *"If you're stressed it's easy to [eat out] or buy a high-calorie meal"* (Participant #9), *"If I am stressed, yes I start, I can be constantly eating..."* (Participant #10). *"I would comfort eat, yes, I am an emotional eater so that is probably the biggest issue"* (Participant #11). Hunger was a factor in making unhealthy food choices for some participants, *"If I'm hungry I sometimes make the wrong choices"* (Participant #12). Some participants suggested that skipping breakfast created a temptation for indulgence, for example, "*if I don't have my breakfast in the morning, by the time I come to college and I go through the mall and I smell the bacon, I want bacon!"* (Participant #6). Others acknowledged that snack foods can themselves be an addiction as strong as alcohol or cigarettes, such as "*I don't drink, I don't smoke but I feel I have as much addiction to say crisps, or chocolate"* (Participant #2).

Although participants did acknowledge that they have their own weaknesses and face their own challenges, there was good awareness across all participants of practical solutions to cope with these challenges. Slimming clubs or activities such as Parkrun or Couch to 5K worked well for some participants, for example, "*I want to get [a slimming club award] and that is a big, big thing that is keeping me going…"* (Participant #2) , *"….[a community running group] worked really well, I have started a habit now of going out at least three times a week running…"* (Participant #11),. However, others reported that these organised activities would not work for them, for example, "*I would find [slimming clubs] a waste of time. If I'm going to self-regulate, I can do it by myself*" (Participant #13). Participants said that a regular routine, whether physical activity or purposeful work, was helpful in preventing unhealthy habits, for example, "*unless I am out doing my activity or whatever, there is a tendency to go to the cupboard and maybe have a few snacks, where during the day is fine, keeping active*" (Participant #4); and "*When I had drink at the weekends, you had quite a few tots, mine was whisky and coke, then you get the munchies and then you are on the phone and it is just a progressive cycle, whereas now I caravan quite a bit, I bought a caravan to study, I know it sounds crazy but it got me out of the house and I went away and studied...temptation used to be a problem for me, but now it isn't"* (Participant #5). "*The biggest issue is to try and change that behaviour because...and the running does help because I feel better when I run, so it means, yes...and also just try to plan a little bit better, when I plan my eating it goes better"* (Participant #11). Regular physical activity was also reported as effective for reducing stress levels, for example, "*I think that exercise helps...and because I have to do*

some physical activity every day, or at least every other day, I have made that part of my routine, I find it probably helps with the stress because I think that when you stop and you start getting, eating badly it can then become a downward spiral quite quickly" (Participant #6).

In summary, participants reported a range of challenges of losing weight and weight maintenance. These challenges focus on balancing diet (energy intake) and physical activity (energy expenditure), which is the key for weight loss maintenance and helps reduce temptation and stress in weight management.

3.4.3.2 Key Theme 2: Social contact can be a double-edged sword

Participants recognised that social contact can be a double-edged sword, as it can provide motivation and accountability but can also be a source of powerful temptations, as not everyone appreciates or understands the difficulty of weight loss.

Participants who were using physical activity to maintain weight loss reported the benefits of exercising with other people, for example, "*With exercising, sometimes I'll go to the gym with a friend and we'll do the same activity together, having that social support definitely helps."* (Participant #8). Slimming clubs can be helpful for participants who found it useful to be accountable and look for encouragement from others, for example, "*You have the weigh-in, I think it helps because I think everybody is friendly and there is different people with different weights, so in some ways you feel you are not the worst off, but with some people you have aspirations too"* (Participant #6).

However, not all forms of social contact were beneficial, and participants did report that sometimes other people failed to understand the challenges of maintaining weight loss. For example, *"sometimes people say, "It's only one drink, it's only one chocolate bar. That's not the point!"* (Participant #1); and "*It's not really an issue for them, and at the other end of the scale, people who are overweight look at you and think you have nothing to worry about."* (Participant #12). Some participants reported that social contact leads to temptation as "*If you are [at work events] they are always throwing everything at you to get you drunk"* (Participant #5).

In summary, social support for weight maintenance is a mixture of positive and negative aspects. It is beneficial for accountability and encouragement but can also be a distraction or reinforce unhealthy habits.

3.4.3.3 Key Theme 3: Apps are popular

The findings indicate that apps are popular because they are convenient and fit in well with users' everyday lives, and the most important app function is the ability to track and review the user's progress. However, it was reported that apps are less useful if they are too complex or time consuming.

Almost all participants (n=14 out of 15) used apps for weight loss and for weight loss maintenance. Participants who used apps stated they liked the convenience, for example "*I like it because it runs in the background, you can kind of set it and forget about it, yeah it's good"* (Participant #7). If an app is too time consuming it will be less useful as *"Sometimes it's just, can you be bothered inputting the data…"* (Participant #1), and "*It takes a lot more time to input manually and stuff. It was just laziness and I got out of the way of it"* (Participant #4). Although not strictly apps, Fitbit-type devices were reported to be very popular. The physical connection to the body makes these devices part of everyday life, which adds to convenience, for example, *"…it's the fact that it's connected to everyday living. The phone is just a little bit remote."* (Participant #5). The most desirable reported feature of apps was progress tracking, which for many participants was the main reason for using apps , as "*I like being able to track what I'm doing" (Participant #15), "If you're doing exercise it is useful to see how many calories you have used or whatever…",* (Participant #3) and "*I like it because you can set up your goals and it will tell you how close you are to it"* (Participant #8).

In summary, mobile apps are popular and the users prefer them to be convenient and allow users to track their progress without being too time consuming.

3.4.3.4 Key Theme 4: Personalised messages are more useful

The findings indicate that personalising messages makes them more useful and relevant to the recipient. The tone of messages was important to participants in the context of engagement. Generic messages were viewed as too impersonal by most participants and

sentiments such as empathy that should be subtle in order to protect against appearing insincere or patronising, and overdone authority could sound bossy.

Participants were asked to give their opinions on different types and tones of message. Participants generally agreed that generic messages, i.e. messages that have not been personalised in any way, were too impersonal. For example, "*There is nothing wrong with generic messages, but I just think that it is...yes it's similar to a tip but to me if it is not a tip then it is not as useful"* (Participant #1), and "*I think tailored messages are best because it feels like you are engaging with it…generic messages would be less useful"* (Participant #1) and "*Why would I want to go and download an app just to get patronizing or random messages that have nothing to do with me. If you are going to get me to download an app you need to make it specific for me",* (Participant #10). It was widely accepted that message sentiment is important but must be subtle to avoid sounding insincere or patronising, for example "I would feel a bit patronised by an insincere message…" (Participant #2) and "*If it was too cheesy, I'd just throw it away!"* (Participant #10) and to avoid downplaying the issue, "*If you make people become too comfortable there's not much point…"* (Participant #7). Participants responded similarly to authoritative messages, and while some participants stated that authority could be beneficial, it would depend on the recipient's personality and the tone of the message , as one participant articulated: "*I guess it depends on your audience member or the person who is receiving that message, how they like being talked to"* (Participant #8), and "*So authoritative, the thing is, this is a tricky one because some people I am sure like being told things softly"* (Participant #6). Most participants however expressed a negative opinion of authoritative messages, saying that they appeared to be too bossy, as "*Authoritative is just trying to tell you what to do"* (Participant #12), *"I don't like being told what to do"* (Participant #13) and "*[Authoritative messages are] a bit bossy"* (Participant #14), *and "Authoritative messages are just like bossing me about"* (Participant #15).

In summary, messages should be personalised for the individual user, and sentiment or authority should be subtle.

3.4.3.5 Key Theme 5: Chatbots have potential for weight loss maintenance

3.4.3.5.1 Chatbots

As has been previously documented, chatbots are computer systems that simulate human conversation (Knowledge@Wharton, 2016), and they have several potential advantages for weight loss maintenance. It has been observed that chatbot-based therapeutic health interventions can improve patient compliance (Barak et al. 2009) and engagement (Fadil & Gabrielli, 2017), and are capable of reacting to emotions and creating empathy when programmed to do so (Barak et al. 2009). Their user interface is conversation driven, highly user-friendly and simple (Personetics, 2016). Chatbots may be integrated with existing social media tools and messaging platforms such as Facebook Messenger, which is currently the most popular social networking app and is installed on most people's smart phones (UKOM, 2018). Integrating chatbots with existing social medial tools provides the user with an interface that is familiar without the need to install additional apps, of which there is already an abundance (Personetics, 2016). Chatbots are fully automated, they do not normally require input from a human operator and are available to users 24 hours a day (Personetics 2016), potentially at times when the user most needs help, and may also automate messages while still retaining personalisation (Fadhil & Gabrielli, 2017). These capabilities effectively resolve the shortcomings of text-messaging based motivational systems for weight loss maintenance, such as the need for a human operator to reply to messages, and the difficulties associated with personalising automated text messages (Donaldson et al. 2014).

As discussed in **section 2.6.8.3.1**, the use of chatbots for therapeutic interventions raises ethical issues regarding trust. As has been previously documented, chatbots may be unable to adequately manage and identify risk, and it is considered unethical to convince a human user that they are talking to another human rather than a chatbot. Chatbots and other artificial-intelligence systems may not yet offer an effective alternative to a trained health professional when providing medical advice (Bickmore et al. 2018).

3.4.3.5.2 Participant feedback on the WeightMentor chatbot

Findings from this needs analysis suggested that the prototype WeightMentor chatbot was easy to use and could help with motivation, but that interactions should be kept to a

minimum to avoid frustration. The idea of a "personality" was received positively, and participants indicated that they felt it made the chatbot seem more natural.

Participants generally liked the casual, conversation-driven nature of the WeightMentor chatbot, as *"I would use it because it is engaging and to me it seems a lot more fluent how to enter data"* (Participant #1), and "*I would use it because it is natural and not too generic"* (Participant #2). Participants reported that the chatbot functions were useful, and that it could potentially be a very useful tool for motivation, as *"I like the graph function and the motivational aspect, I think it could be quite useful"* (Participant #8), and "*I liked it because it keeps me accountable and would be a good way to keep people motivated"* (Participant #9). However, it was generally accepted that interactions with the chatbot should be kept as minimal as possible as users could become frustrated by complex interactions. For example, "*I think there are potentially too many questions [in the activity logging] ...I think someone could potentially get frustrated."* (Participant #11), and that there should be no obligation for a user to interact with the chatbot daily.

Participants stated that users should instead be able to use the chatbot as and when they feel the need to as *"I would want to use it when I want to, without having to use it every day, or whatever."* (Participant #2). The concept of a chatbot "personality" was received positively by participants, particularly if it was possible to choose the personality, as "*I would like to be able to choose the personality. It would be nice to have an upbeat personality one day and more serious the next."* (Participant #3) and "*I liked the humour of the hello and goodbye…I think as well if you could create some kind of 'character', that might be quite useful"* (Participant #11).

Participants concluded that they could see the merits of using WeightMentor chatbot for weight loss maintenance as it was viewed as engaging and simple to use, however it was felt that interactions should be kept to a minimum, there should be freedom to use the chatbot as and when needed (rather than a daily obligation to do so), and the chatbot's personality should be developed further.

In summary, chatbots could be useful for weight loss maintenance in that they are easy to use and accessible 24 hours per day. The prototype WeightMentor has shown potential at this stage in both development and analysis. Interactions should be kept to a minimum and the chatbot's personality should be developed to make it natural and friendly

3.5 Discussion

The purpose of the needs analysis was to identify the needs and requirements of individuals who are maintaining weight loss and determine how these needs may be addressed by a weight loss maintenance chatbot.

3.5.1 Needs for weight maintenance

Evidence generated from the interviews suggests that the needs of those who had lost weight and sought to maintain it were (1) to support the user at times when they are most vulnerable (e.g. tired, stressed), (2) facilitate progress tracking, with appropriate encouragement, (3) provide a tool that fits into the user's daily life with minimal inconvenience, (4) provide feedback that is positive and encouraging without being patronising.

Interview participants indicated that they were most vulnerable when stressed or tired, for example after a long day at work or university, and may need help and support in the evenings, or weekends. Findings from the *LEAP Beep* trial (Donaldson et al. 2014) highlighted the lack of out-of-hours support as a limitation of text-message based interventions, as these require a human operator to read the message and respond with appropriate feedback. Participants found progress tracking to be the most useful of app functions, as it allowed them to see how they had progressed. Progress tracking has been seen to be of benefit in weight loss maintenance interventions (Donaldson et al. 2014, Fjeldsoe et al. 2016). Participants said that the most useful apps are those which integrate with everyday life without being inconvenient to use or access. The FitBit is particularly popular because it may be worn on the wrist and is unobtrusive and simple to use. Participants said that feedback should be positive and encouraging without being patronising, which is consistent with findings from Pollard et al. 2016, and is discussed further in section 5.2.

3.5.2 Personalised messages

Most participants reported they would find personal messages helpful. However, it was generally agreed that messages should always be personalised, in order to ensure that they are specific to the individual recipient. While some participants said they would view generic messages as similar to "tip of the day" type messages in that they would seem

helpful but not particularly useful, most said they would ignore generic messages and showed strong preference for messages that were personally relevant to them. This is consistent with the findings of Woolford et al. (2011), Smith et al. (2014), Stephens et al. (2015), Pollard et al. (2016) and Cohen et al. (2017). Participants in a focus group study by Smith et al. (2014) found personalised messages "less automated" and felt they had a more human touch. The importance of message personalisation was more prominent among younger participants. This is consistent with a study by Stephens et al. (2015), which found that in a world where the individual is bombarded with "generic" emails and notifications, such as those from banks etc. it is easier to ignore anything that is not personally relevant. Message tone was also recognised as important. Consistent with work by Smith et al. (2014) and Pollard et al. (2011), participants appreciated empathetic messages, and viewed supportive, encouraging messages as more acceptable than authoritative messages. Participants did state, however, that encouragement and empathy should be subtle, in order to avoid sounding patronising. Pollard et al. (2016) reported that "Generation Y Engaging" messages which were specifically targeted towards young people using language they understand were in some cases seen as patronising or trying too hard.

3.5.3 Apps

Participants who used apps for weight loss or weight maintenance stated that it was helpful to be able to report and review their progress and suggested that they would be more likely to use weight loss apps that offered this functionality. This is consistent with findings by Burke et al. (2012), who reported that electronic diaries with personalised feedback resulted in increased intervention engagement and better weight loss. Participants in this study acknowledged that it is important for apps to be convenient and to fit in with their day-to-day life. Participants viewed apps as time-consuming with limitations with respect to usefulness. This is consistent with the perceptions reported by participants in the LEAP Beep trial (Donaldson et al. 2014) and the Get Healthy Stay Healthy trial (Fjeldsoe et al. 2016) who were satisfied with the convenience of receiving text-message based feedback.

3.5.4 The WeightMentor Chatbot

Participants generally reacted positively to the WeightMentor prototype, which was described as "engaging" and "friendly". As previously discussed, chatbot-based health

interventions have been linked with increased participant engagement (Barak et al. 2019). The benefits of chatbot-based systems, increased user engagement, automated personalised feedback and increased usability are all cited by Fadhil et al. (2017) as a rationale for using chatbots for weight loss maintenance interventions. Research by Donaldson et al. (2014) suggested that text messaging is effective for weight loss maintenance because text messages are delivered directly to a participant's phone and may be read without downloading any additional apps or logging in to a website. Donaldson et al. (2014) also indicated that participants readily engage with this level of convenience. The convenience of engaging with a chatbot has been identified by participants in this review as a desirable feature.

Individuals who are losing weight or maintaining weight loss face specific challenges and need support to overcome these challenges. Smartphone apps or other technology-based tools are popular because they can be easily integrated into daily life without inconvenience and provide a means of tracking progress. Chatbots are friendly, natural and easy to use, so are engaging and may have potential for weight loss maintenance, however interactions with a chatbot should be minimal and responses should be specific to participants.

3.5.5 Implications for *WeightMentor* Chatbot Development

Feedback from participants suggests that the current features of the prototype chatbot are useful and appropriate. However, it has been suggested that the self-reporting aspect of the chatbot is a little cumbersome and may be worth scaling down. Participants expressed a view that WeightMentor would be most useful if it could be used "as and when required", rather than daily. It is possible with the current functionality to do this, however a requirement to use the chatbot on a regular basis may in fact stimulate engagement and improve effectiveness of the chatbot, thus it may be beneficial to require some degree of regularity of usage, within reason. It is also worth considering how the chatbot's personality may be developed to make it more natural, however as discussed in chapter 2, it may potentially be unethical to engineer the chatbot to appear human. Thus, while it is important for WeightMentor to be friendly and engaging, it must never seem to the users that they are talking to a human. *WeightMentor* must meet the needs that have been identified from this needs analysis. This may be achieved using the strategies summarised in **Table 3.5**. The *WeightMentor* chatbot must also support the ten *Processes*

of Behaviour Change associated with the maintenance phase of the Transtheoretical Model of Health Behaviour Change, as discussed in **section 2.3.1**. The *WeightMentor* functions support six of these ten processes.

Table 3.5: Meeting User Needs using the *WeightMentor* Chatbot

Identified need	Relevance to Transtheoretical Model	How could the chatbot address this?
Support the user at times when they are most vulnerable (e.g. tired, stressed)	Promotes *stimulus control*, by boosting the user's ability to cope when they are vulnerable and thus most likely to give in to temptation	WeightMentor will be available 24 hours a day, seven days a week
Progress tracking, with appropriate encouragement	Progress tracking supports *Consciousness Raising*, as the user becomes more aware of their eating and exercise habits, causes of negative behaviours and possible solutions. *Self-re-evaluation* and *self-liberation* are supported by appropriate feedback, which helps the user to feel positive about even small achievements. Feedback will also support *contingency management* by helping the user to associate positive behaviour with positive feedback and *counter conditioning* by providing users with a means to manage their negative behaviours and reinforce positive behaviour.	Users of WeightMentor can request a graph of their self-reported food and physical activity trends at any time. When a user self-reports, they are given feedback based on what they have reported. Positive reports elicit subtle praise and encouragement, negative reports elicit subtle encouragement.
Tool should fit into user's daily life with minimal inconvenience	Not directly relevant – although *WeightMentor* will be most effective in supporting the maintenance phase of the Transtheoretical Model if the user is not inconvenienced by its use.	WeightMentor may be accessed from a user's phone using Facebook messenger, the most popular social media app User is not obliged to report in every day, but may use the chatbot only when they need to WeightMentor interactions should be as simple and minimalistic as possible
Feedback should be positive and encouraging without being patronising	*Self-re-evaluation* and *self-liberation* are supported by appropriate feedback, which helps the user to feel positive about even small achievements. Feedback will also support *contingency management* by helping the user to associate positive behaviour with positive feedback and *counter conditioning* by providing users with a means to manage their negative behaviours and reinforce positive behaviour.	Feedback will be subtle. Encouragement will focus on reminding participants about their previous achievements rather than emphasising their failures

3.6 Conclusion

There is no quick fix for obesity, and individuals who are maintaining successful weight loss are often capable of identifying strategies that work for them. Needs analysis participants identified convenience and progress tracking as the most popular selling points of weight loss apps, determined that subtle personalised messages were most useful, and acknowledged that temptation, emotion and stress are the biggest challenges to weight loss maintenance and that social contact can be either a hindrance or a help. Although participants reported that *WeightMentor's* friendly, engaging personality could be helpful, they did stress that interactions should be kept to a minimum to maintain convenience and minimise frustration. Findings from this needs analysis suggested that a chatbot could assist in addressing the needs of individuals who are maintaining weight loss.

Chapter 4: *WeightMentor* Chatbot Design & Development

4.1 Introduction

Chatbots are artificial intelligence (AI) systems capable of mimicking human conversation (Knowledge@Wharton, 2016) (refer to **Chapter 3: Needs Analysis**). One of the potential selling points for chatbots is that they are conversation driven as opposed to being mouse-and-pointer driven in the case of conventional graphical user interfaces. As the vast majority of interactions with chatbots will be text or voice based, the user interface will therefore be very rudimentary, consisting of a text entry box, chat window, and a "send message" button in the case of text-based chatbots such as *WoeBot* or *iHelpr*. Digital personal assistants or "smart speakers", such as Amazon Echo or Google Home will usually be controlled by little more than a microphone, loudspeaker, volume controls and a mute button.

4.1.1 Development Methodology

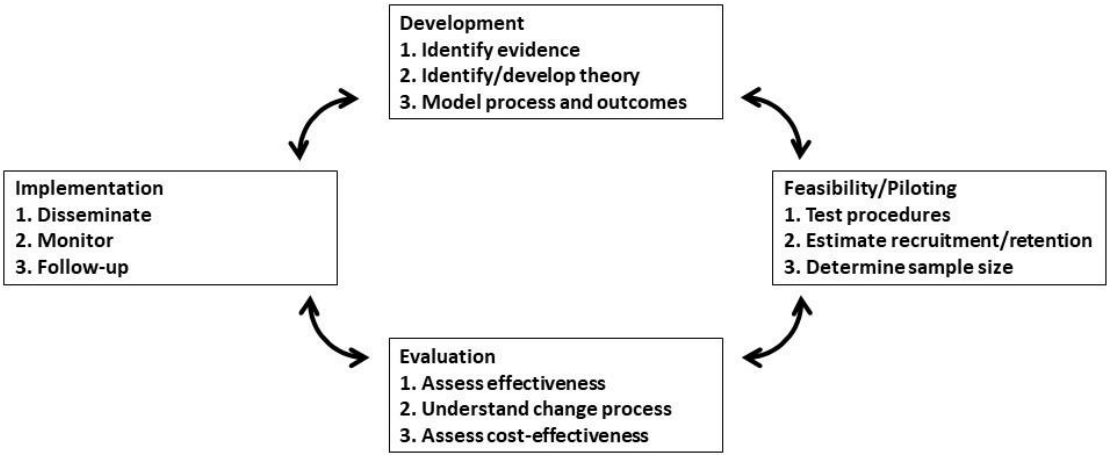


Figure 4.1: Phases of the Medical Research Council (MRC) Framework for Developing and Evaluating Complex Observations

This project will progress in accordance with the Medical Research Council (MRC) Framework for Developing and Evaluating Complex Interventions, as updated in 2019. This framework has been developed to help researchers overcome the problems associated with designing complex interventions within a healthcare setting. The framework proposes a four-phase process for development of complex interventions. A complex intervention is defined as one which has several interacting components. The four phases of the framework are illustrated in **Figure 4.1**. The five objectives identified

in chapter 1 align with the first two phases of the MRC Framework. Objectives one to three correspond to the MRC Design phase, while phases four and five correspond to the MRC Feasibility and Piloting phase.

4.1.2 System Development Lifecycle

A System Development Lifecycle (SDLC) is a vital part of the software design process. The SDLC provides a clear structure for development that allows problems to be identified early (Half, 2017) and may be resolved before they become more costly. Several different SDLCs exist, some of which are summarised in **Table 4.1.**

Table 4.1: Common SDLC Models

Model	Description
Agile Model	An iterative model, with system design progressing as a series of releases, each building on the previous. Product testing takes place at each iteration. Although quite a new model, it has quickly become the favourite of many organisations.
Lean Model	The Lean model is inspired by Lean manufacturing processes. The model seeks to focus only on one task at a time, while reducing waste (e.g. unnecessary meetings, or unnecessary documentation).
Waterfall Model	The oldest of the SDLC models, the Waterfall model is also the most straightforward. Development progresses forwards from one process to the next, with no going backwards. It is often seen as being too rigid as it does not allow for revisions, and delays at an earlier stage of the model can affect later stages.
Iterative Model	The Iterative Model is like Agile in the sense that it relies on a series of releases, which are tested and refined until the final system is completed. The Iterative Model is popular because it ensures early deliverables and reduces the cost of changes and debugging. However, the repetitive nature of the cycle makes it somewhat resource hungry.
Spiral Model	The Spiral Model is like the Iterative Model in the sense that its four phases (Planning, Risk Analysis, Engineering and Evaluation) are repeated as a "spiral" until the project is completed. This allows for multiple iterations of refinement. This model is popular for managing large projects because of its excellent risk-management component. The Spiral Model also allows stakeholder feedback to be implemented early
DevOps Model	The DevOps Model is one of the newest SDLC models. It encourages Development and Operations teams to work closely, accelerating innovation and boosting quality. This model is like iterative models because it relies on small, frequent updates. The model encourages discipline, continuous feedback and process improvement.

4.1.2.1 A Development Life Cycle for Chatbots

All the SDLCs described above have been designed for development of conventional mouse-and-pointer based systems, thus may not be easily applied to chatbot development without being adapted for this purpose, and it may be more practical to follow a lifecycle

specifically designed for chatbot development. Sheth (2016) proposed an eleven-stage chatbot development life cycle, which is presented in **Figure 4.2** and summarised in **Table 4.2.**

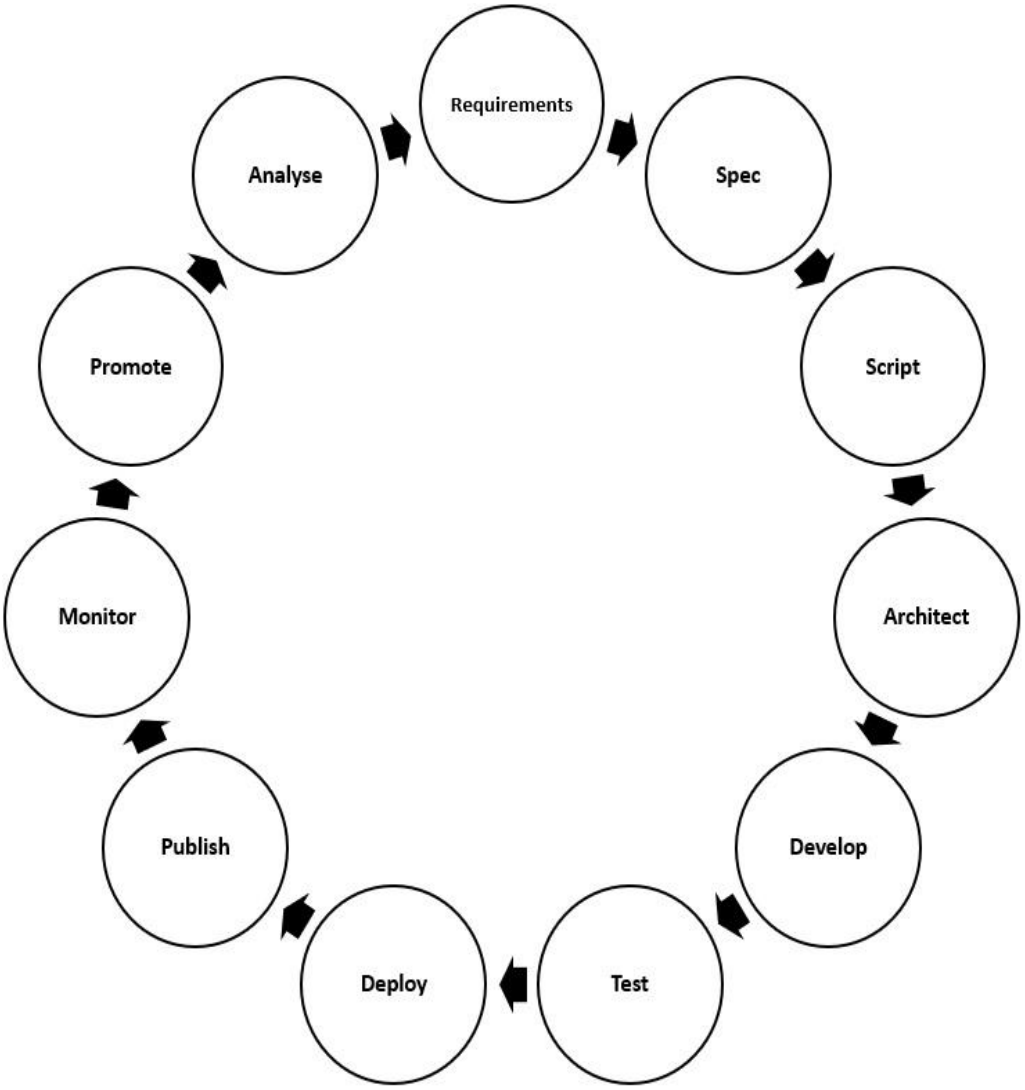


Figure 4.2: Chatbot Lifecycle

Table 4.2: Chatbot Lifecycle Summary

Step	What is involved
Requirements	Gather information from potential users and target customers to determine what requirements are. This step is like the requirements analysis step in conventional software projects.
Spec	Develop a project specification that identifies features and functions of the chatbot. The specification should address the requirements identified in the previous step.
Script	A step that is unique to chatbot design, scripting involves building conversations based on predetermined scripts. The script should guide the user towards their goal. If the chatbot is to use Natural Language Processing, this should be taken into consideration during the scripting stage, thinking about all the possible variations of a word or phrase that may occur.
Architect	The main phase of designing the front- and back ends of the chatbot. The front end will map conversation inputs to actions, while the backend will handle main chatbot functionality and links to external services.
Dev	The actual development stage where the chatbot code is written. This phase will be highly iterative, with developers alternating between writing code and testing functionality.
Test	Testing is an integral part of the development stage as discussed above, however this phase provides an opportunity to test the chatbot's integration with the messaging platform and ensure the chatbot responds in the right way at the right time.
Deploy	The completed chatbot should now be deployed to a host platform (such as Heroku)
Publish	The completed chatbot will need to be published to app stores (or approved with Facebook). Approval may take a few weeks, so this should be taken into consideration.
Monitor	The chatbot should be monitored to ensure correct functionality. Of importance is making sure the chatbot responds correctly to user conversations.
Promote	The chatbot should be marketed using social media and other advertising channels to attract traffic.
Analyse	The chatbot's performance and interactions should be analysed. This provides an opportunity to gauge efficiency and determine which conversations are not working effectively, if any.

4.1.3 UX Design Practices

Human Computer Interaction (HCI) is focused on the relationship between humans and computers, and strategies that may be implemented to facilitate easy interaction. HCI best practices often define clear rules that should be followed when designing a user interface, such as Ben Shneiderman's Eight Golden Rules (Shneiderman et al. 2016) or Nielsen's Usability Heuristics (Nielsen, 1994).

Shneiderman's Eight Golden Rules state the following:

1. The system should be consistent in terms of layout, menus and colour schemes to increase familiarity. Exceptions (such as concealing passwords) should be minimal.
2. The system should be accessible to all users. It should be intuitive and helpful for novice users but should enable expert users to improve efficiency by using shortcuts.
3. The system should give appropriate, timely feedback for every action."
4. System functions should follow a logical sequence, and the user should be informed when actions have completed (e.g. a file has been saved).
5. If it is not possible to prevent errors from occurring, the system should help users recover from them quickly.
6. Users should be able to undo and redo actions. This "encourages exploration of unfamiliar options" (Shneiderman, 2016)
7. Users should always feel in control of the system, as this will improve confidence and reduce frustration. There should be no surprises!
8. The system should reduce users' short-term memory load by reducing how much they need to remember, e.g. by pre-filling forms etc.

Nielsen's (1994) usability heuristics state the following:

1. Users should be kept informed about what the system is doing or what has happened, through timely, relevant feedback.
2. Language should be matched to the user, rather than jargon filled. Information should be presented naturally and logically, as it would appear in the real world.
3. Users should have full control over the system, and should be able to rectify mistakes, undo and redo actions without being inconvenienced.
4. Language and actions should be consistent throughout the system and should conform to platform conventions.
5. Error prevention should always take preference even over well-designed error messages.
6. System components should be designed to facilitate *recognition*, rather than *recall*. The number of things a user needs to remember should be minimised and instructions should be easy to find and access.

7. The system should cater for novice users but allow experienced users to use shortcuts to speed up tasks.
8. Irrelevant content reduces visibility and effectiveness of relevant content, thus should be minimised or eliminated entirely.
9. Error messages should be meaningful, appropriate and clearly explain the problem and help the user to resolve it.
10. A good system will be intuitive enough to make documentation unnecessary, however where documentation is needed it should be easy to follow and accessible.

When applied to interaction design, these best practices help to improve system usability (discussed in chapter 5) and create a positive user experience (UX). However, these best practices have been developed to inform the design of conventional graphical user interfaces. As already discussed, the graphical component of the chatbot user interface will exist solely for the purpose of facilitating text- or voice-based conversation, as the user interface itself will primarily be driven by conversation, either text typed into a chat window, or verbal instructions spoken into a microphone. When designing this type of user interface, it is not easy to apply principles such as those of Shneiderman or Nielsen. Conversational user interfaces need to replicate human conversation, and although the simplest way to do this is to merely model conversations on a very simple level, real human conversations can be complex, and thus conversation design will need to take this into consideration.

Conversation analysis studies have identified three basic principles of conversation (Moore et al. 2017). Firstly, *recipient design* is concerned with tailoring conversation to match the recipient's level of understanding. A good example of this is the use of jargon in medical or scientific fields. Jargon may be defined as *words and phrases used within a specific profession or field in order to convey the meaning of concepts relevant to the field.* It is often expected that individuals working in, for example, the computer science field, will use jargon in conversation with other like-minded professionals. However, it is irrelevant and confusing to use jargon in conversation with laypeople, such as patients or clients, as doing so may appear patronising, or lead the listener to believe that the speaker does not fully understand what they are talking about. Thus, the conversation should be tailored so that concepts may be described using a language the patient or client will be familiar with and understand (Moore et al. 2017). The second principle of

conversation is *minimisation*, the idea of "not using twenty words when five will do". If one of the main goals of conversation is to inform and be understood, then it is essential that conversation should be kept as simple as possible, using as few words as possible in order to maintain efficiency and promote understanding. Finally, *repair mechanisms* should be in place that can be employed when understanding fails. Two examples of these are repetition and paraphrasing (Moore et al. 2017). Repetition allows the listener to hear what has been said for a second time, while paraphrasing uses different words to express the same concept.

Conversations may be modelled on one of four types. The first of these is *ordinary conversation*, which may be more commonly known as "small talk". It is casual, informal and not usually constrained by specific rules. Ordinary conversation is used in everyday life, primarily as a means of connecting socially. The second type of conversation is *service conversation*. This is more formal, and usually reserved for situations where one person (e.g. waiting staff in restaurants, customer service agents, or salespeople) serves another person (e.g. a diner, customer, or guest). In service conversation there are two roles, service requestor and service provider. This type of conversation is more constrained, with a limited number of specific conversation paths, such as introducing oneself, explaining one's needs (as the service requester) and providing relevant information (as the service provider). *Teaching conversation* happens between a student and either a teacher (in an academic setting) or an instructor (in a training setting). The student may ask questions of the teacher/instructor in order to clarify understanding, while the teacher/instructor will use questions as a means of testing the student's knowledge. Alternatively, teacher's questions may be rhetorical or reflective, to encourage the student to think about a concept. In this type of conversation, interruption and correction (necessary aspects of all conversation types) are particularly important. Finally, *counselling conversation* is most used in "talking therapies" such as during visits to a counsellor or psychologist, although it is not unique to these situations. Counselling conversation may also occur between a consultant (such as in the field of Information Technology) and client, or indeed in any situation where one person seeks advice from another. In a counselling conversation, it is common for the "counsellee" to lead the discussion, with the "counsellor" prompting when necessary. It is not uncommon for counselling conversations to be very one-sided, with one participant doing most of the talking.

Conversation modelling is an important part of chatbot development as it will shape conversation flow and determine when inputs are required from the user and how the chatbot should respond. *WeightMentor* conversations will be a combination of *ordinary conversation*, *service conversation* and *counselling conversation*. For *ordinary conversation*, the chatbot uses a built-in *Small Talk* module that contains pre-defined conversation topics such as "Who are you?", "Are you human?" and "What is your name?". During normal use, *WeightMentor* will utilise *service conversation* to ask the user what they want to do, collect self-reporting data from them, and display trend data. *Counselling conversation* will be used when the chatbot provides motivational feedback for the user.

4.1.4 *WeightMentor* chatbot type

As chatbots have gain popularity, a variety of types have emerged, and there are a variety of ways to classify the different types. Gnewuch et al. suggest that chatbots may be very simply classified on two dimensions – Context (i.e. purpose) and Mode of Communication (e.g. text vs. speech). This taxonomy is illustrated in **Table 4.3**. However, it is also possible to classify chatbots based on their function. For example, chatbots such as *WoeBot* and *Wysa* are text-based chatbots designed with a specific role in mind, i.e. mental health management or self-guided *Cognitive Behavioural Therapy (CBT)*. The Ada Health app, and the NHS App are designed as "symptom checkers". This kind of app asks the user to describe their symptoms and suggests possible diagnoses. IKEA's chatbot is one example of a domain-specific chatbot where the domain is customer service. Chatbots used in this way can provide product information, details about store locations and opening times, or advice about returns etc.

Table 4.3: A simple taxonomy for Conversational Agent classification (Gnewuch et al. 2017)

		Context	
		General-Purpose	Domain-Specific
Primary Mode of Communication	Text-Based	ELIZA, Cleverbot	IKEA's Chatbot, WoeBot, Customer service Chatbots
	Speech-Based	Siri, Alexa, Google Assistant, Cortana	In-Car assistants, Speech-based customer service agents etc.

Not all chatbots are designed with a specific purpose in mind, however. *ELIZA*, the world's first true "chatbot", was designed as an experiment in artificial intelligence and was purely conversational in role, and *Cleverbot* is the modern-day equivalent. Other "chatbots" which have no specific purpose are speech-based agents such as Siri, Alexa and Cortana, which may be used to play music, answer questions, or even control home automation devices.

As part of the *WeightMentor* design process it will be useful to think about what type of chatbot it will be and what role it will play. *WeightMentor's* primary mode of communication will be text-based, and it will play a specific role, that of motivation and weight loss maintenance support, however it will not provide nutrition advice or guidance. The *WeightMentor* chatbot will provide some elements of casual conversation to appear relaxed and informal.

4.1.5 *WeightMentor* chatbot UX design principles

Four design principles have been identified for consideration during the development of the *WeightMentor* chatbot. These are summarised in **Table 4.4.** Design principles cover the four areas of *Conversation Leading*, *Conversational Variance*, *UX Personalisation* and *Message Brevity*. These design principles were selected based on discussion with the research team and with guidance from Emeritus Professor Michael McTear, who is a leading researcher in Conversational AI from the Ulster University. Design principles were selected in order to comply where possible with the design practices suggested by Shneiderman and Nielsen (discussed in section 4.1.3), and to ensure that the *WeightMentor* chatbot presented a positive UX. The *WeightMentor* chatbot should always lead the conversation so the user knows what is happening at every stage of their interaction and so they are not left wondering what to do next, as per Nielsen's Heuristic #1 (keeping users informed). Interactions with the *WeightMentor* chatbot will be varied in order to make the experience more interesting for the user. Although the primary functions will always remain the same, and will always be accessed in the same way using the same set of commands, as per Shneiderman's Golden Rule #1 and Nielsen's Heuristic #4 (system consistency), the specific text and interactions with the user will vary (such as during motivational feedback and greetings). This will prevent the user from becoming bored and will create the illusion of participating in a real conversation.

The *WeightMentor* chatbot will personalise interactions with the users by using their name during greetings, when asking questions and during personalised feedback. Additionally, feedback will be personalised based on the user's previously reported food and physical activity reports. This will help the user to feel that they are being addressed as a human being. This is consistent with findings from focus group studies conducted by Woolford et al. 2011, Smith et al. 2014, Stephens et al. 2015 and Pollard et al. 2016, in which participants stated that personalised messages felt more relevant to them, and that they were less likely to ignore those messages which addressed them by name or seemed to be written specifically for them. Finally, *WeightMentor* will keep messages sent to the user as short as possible. Where it is necessary to send a message that is composed of multiple sentences, these will be sent as a series of individual messages, with a suitable delay between each message to allow reading time. This will prevent users from being overwhelmed by too much text and improve readability of messages, as per Shniederman's Golden Rules #2 (universal accessibility) and #8 (reducing short-term memory load), and Nielsen's Heuristic #6 (reducing short-term memory load).

Table 4.4: *WeightMentor* Chatbot UX Design Principles

Principle	Rationale
The chatbot should always lead the conversation	The user should never be left in "limbo", wondering what to do next. The chatbot's final interaction with the user should always be a question, even if it is a simple "Anything else I can do for you?"
The chatbot's interactions should be varied	If the same greetings or responses to feedback or self-reporting are used, the conversation will become tedious and boring after a while. Variability will help maintain interest in the conversation and preserve the illusion of participating in a "human" conversation. Context awareness can be used to further enrich the user experience, with the chatbot using specific responses depending on the day of the week, time of day, or making comments specific to the user's location, e.g. weather related.
The user experience should be personalised as much as possible	The user should feel that the chatbot is talking to them as a human being rather than just another machine. Messages should not be generic but should instead be personalised based on user's name and personality.
Messages should be kept as short as possible	Long messages may be tiresome for a user to read. Messages should be limited to a maximum of 1 sentence per message. Several messages may be sent one after the other if necessary.

4.2 Aim

To design and develop *WeightMentor*, a personalised motivational chatbot for weight loss maintenance. This was based on findings from user needs analysis and current best practices.

4.3 Methods

The *WeightMentor* chatbot was designed in a series of iterative steps, which are illustrated in **Figure 4.3.**

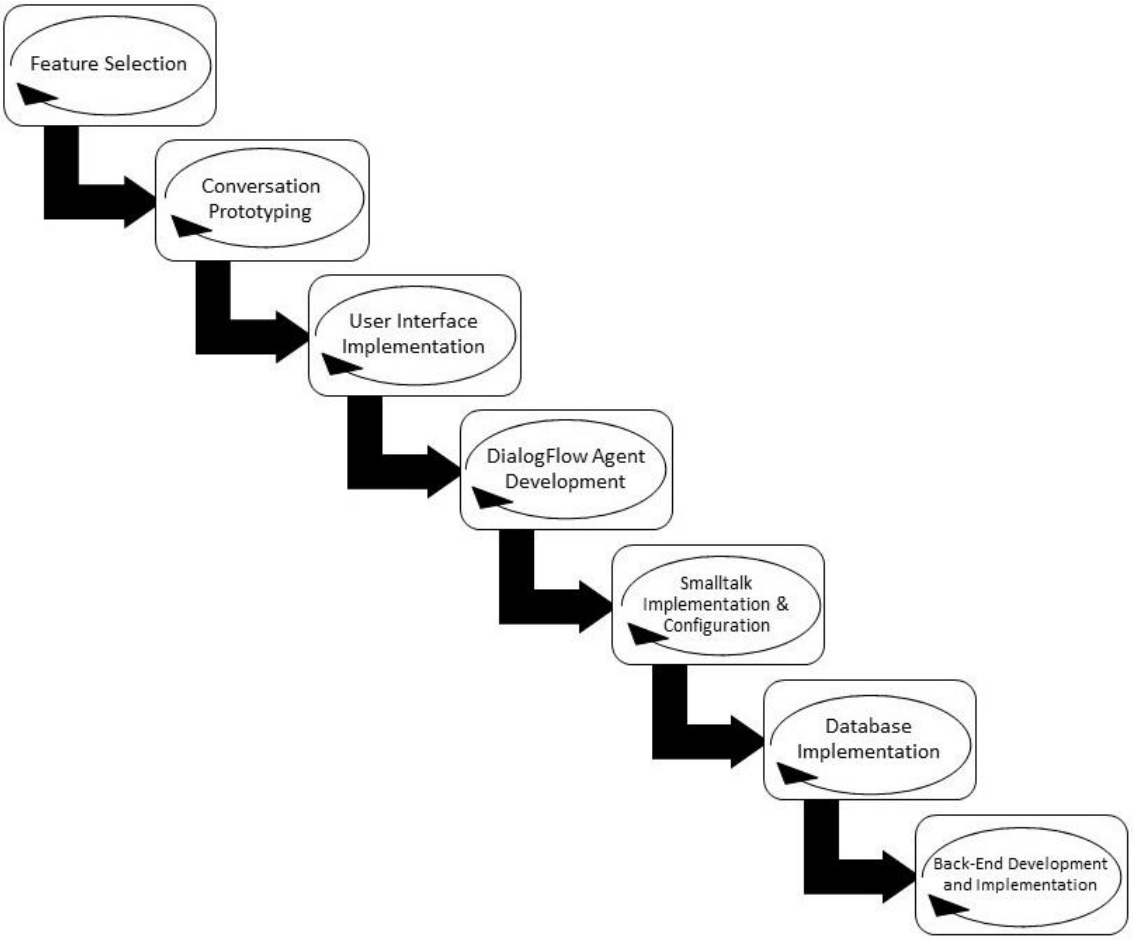


Figure 4.3: Sequence Diagram of *WeightMentor* design process

4.3.1 *WeightMentor* function selection

Before commencing the design process, the *WeightMentor* chatbot's functions were determined based on a review of top nutrition apps by Franco et al. (2016), and findings from two randomised controlled trials using text-messaging based interventions for weight loss (Donaldson et al. 2014, Fjeldsoe et al. 2016). Findings from the *Lifestyle Eating and Activity Programme (LEAP) Beep* (Donaldson et al. 2014) and *Get Healthy Stay Healthy (GHSH)* (Fjeldsoe et al. 2016) indicated that weight loss maintenance could be improved by regular self-monitoring and reporting (of physical activity and food consumption) and motivation. Thus, it was determined that self-reporting, motivation and personalised feedback should form the main functions of *WeightMentor* in order to allow the chatbot to fulfil its role as a personalised motivational tool for weight loss

maintenance. These functions will support six of the ten behaviour change processes of the Transtheoretical Model (see **sections 2.3.1.** and **3.5.5**), which are summarised in **Table 4.5** and discussed further in **chapter 7 (section 7.8)**. It was also determined that *WeightMentor* should allow for creation of a user profile which would facilitate the collection of demographic data from participants and allow participants to choose what they want the chatbot to call them.

Table 4.5: *WeightMentor* functions mapped to Transtheoretical Model processes

Function	Transtheoretical Model Process(es)
Feedback	• Counter Conditioning
Motivation	• Stimulus Control • Self-Liberation
Positive Encouragement	• Contingency Management • Self-Liberation • Self-Re-evaluation
Self-Reporting	• Consciousness Raising

4.3.2 *WeightMentor* conversation prototyping

Before building the conversation framework and writing the chatbot code, several sample conversations were designed to determine how conversations with *WeightMentor* would flow. A conversation in *WeightMentor* consists of two parts: anything the user says to the chatbot (or an action the user performs) and the chatbot's response to the user. To design conversations, an initial list of conversations was compiled in Microsoft Word (for example, "Creating a user profile", "Self-Reporting", "Checking user's progress"), which were then expanded to identify what the user would say to the chatbot during the conversation, and how the chatbot would respond. Conversation flows were drawn using Microsoft PowerPoint. A sample conversation flow is shown in **Figure 4.4.**

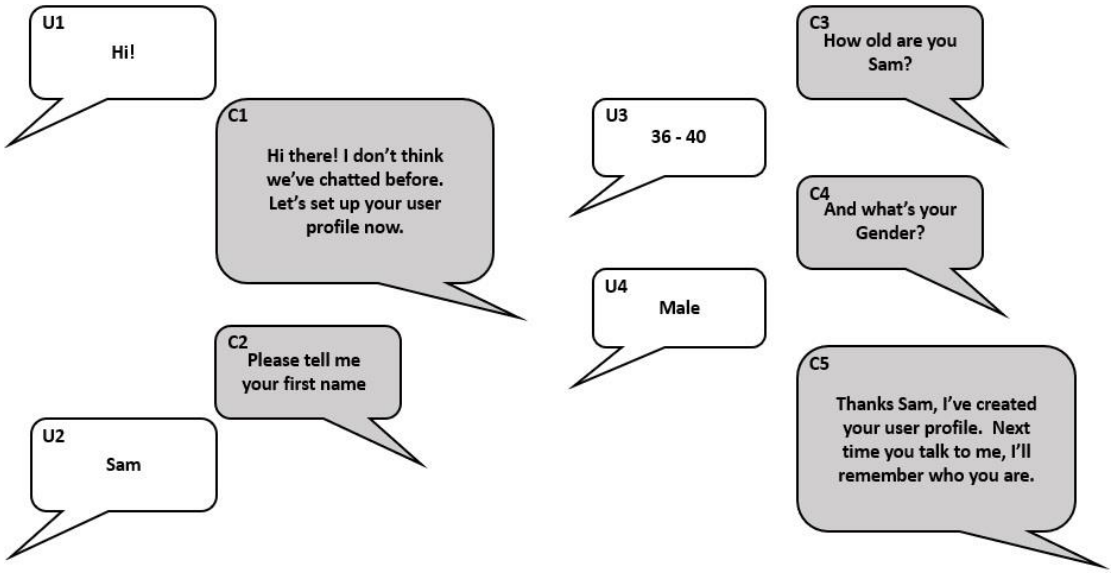


Figure 4.4: Sample conversation flow.
(User statements are marked "U", chatbot statements are marked "C")

4.3.3 Building the *WeightMentor* chatbot

The *WeightMentor* chatbot comprises several core components. The front-end and user interface is provided through Facebook Messenger. All interactions with the chatbot happen through the Facebook Messenger interface, which may be accessed from a variety of devices – smartphones, tablets, or via a web browser on a PC or Mac. Messages received through the Facebook Messenger interface are passed to a Facebook app, which parses messages and passes them to a back-end application written in NodeJS and hosted online using the Heroku platform. If the Heroku application code can handle the message programmatically it will do so, otherwise the message will be passed to an Artificial Intelligence agent built using the DialogFlow framework. Dialogflow will determine what to do with the message and either respond itself or trigger a response through the NodeJS application. Responses to messages are passed back to the Facebook app and to the Facebook Messenger interface where they are received and read by the user. A diagram showing the *WeightMentor* chatbot architecture is shown in **Figure 4.5.**

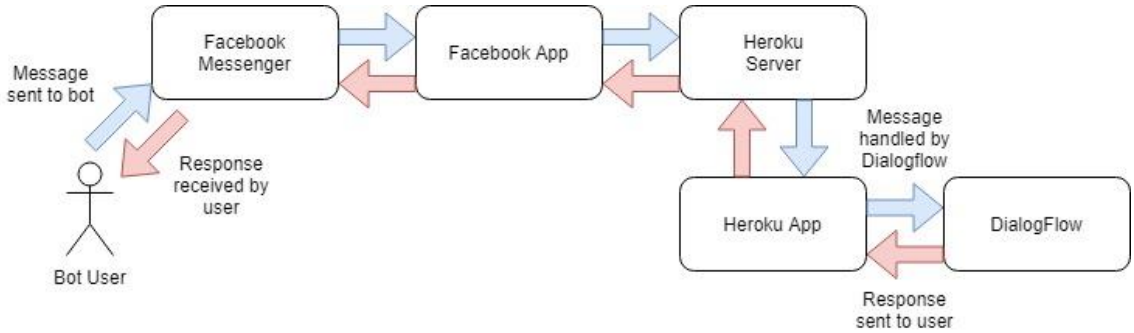


Figure 4.5: WeightMentor Architecture Diagram

4.3.3.1 User Interface: Facebook messenger

US technology firm *Personetics* (2016) discussed in their "Bot Brief" whitepaper a phenomenon called "app fatigue", which is described as frustration with the overwhelming number of smartphone apps available. It is suggested that around a quarter of smartphone apps are downloaded, used once and then abandoned (Personetics, 2016). Thus, it was desired that WeightMentor should be integrated with an existing messenger platform such as Facebook Messenger, Twitter or Skype, to eliminate the need for downloading a new app and reduce the time taken to learn a new user interface. Facebook Messenger was selected for platform integration, based on recent statistics which suggest that Facebook is the most popular social media platform (UKOM 2018, Ofcom 2018). A new *WeightMentor* Facebook page was created and connected to a *WeightMentor* Facebook app.

4.3.3.1 DialogFlow Agent

At the core of the *WeightMentor* chatbot is a *conversational AI* agent. This agent is a module that is capable of *Natural Language Understanding (NLU)*, the ability to interpret and understand human language. A range of different frameworks exist for building the agent, and while Chatfuel is the most popular, it is designed to be friendly for beginner users. It is possible to build a very basic chatbot using Chatfuel, however functionality is limited, and the machine learning (ML) capabilities are very rudimentary. Machine learning is an important aspect of conversational AI and may be implemented using an ML enabled framework such as DialogFlow.

The *Dialogflow* framework started as an independent project and is now owned by Google. Dialogflow provides all the building blocks for creating powerful agents and has extensive ML capabilities, thus was the preferred platform for building the *WeightMentor* chatbot. At the core of the *WeightMentor* chatbot is a DialogFlow agent, which handles conversations. DialogFlow agents are ML-enabled and capable of understanding human conversation. Agents have several key components. *Training phrases* are specific statements the agent has been trained to recognise (for example "hello", or "self-report"). When a user types a training phrase, the agent will recognise it and respond accordingly. Training phrases are mapped to *intents*, which provide a link between training phrases and *actions*, which are performed by the chatbot in response to an event or training phrase (e.g. send a text response or an image). Intents are generally triggered by training phrases,

but they may also be triggered by *events*, which come from external sources (e.g. clicking the "Get Started" button in Facebook messenger"). When an intent is matched to either a training phrase or event, it triggers the associated action, and the designated response is sent to the user. Where a user has supplied some personal data (e.g. name, gender) this data will be captured within the intent *context*, which sets the context or "topic" of the conversation. Data are stored as *parameters* within contexts and may be accessed so they can be used. A diagram of how the DialogFlow components interact is shown in **Figure 4.6**. The main DialogFlow components are summarised in **Table 4.6.**

Table 4.6: Main DialogFlow Components

Component	Purpose
Action	Specifies the action to perform when intent is triggered
Agent	Manages conversation flow. Capable of NLP.
Context	Specifies the context or "topic" of the conversation.
Entity	Used to extract data from user input. System defined entities (such as colour, date or time) exist, but developers may define their own as needed.
Event	Triggers an intent from an external source, such as Facebook Messenger or Google Home.
Intent	Creates a link between training phrases, actions and events.
Parameter	Used to extract data from user input, in conjunction with entities.
Response	How the agent responds to what the user says.
Training phrase	Anything the user says to the chatbot. When a training phrase is recognised by the Agent, it will trigger the associated intent.

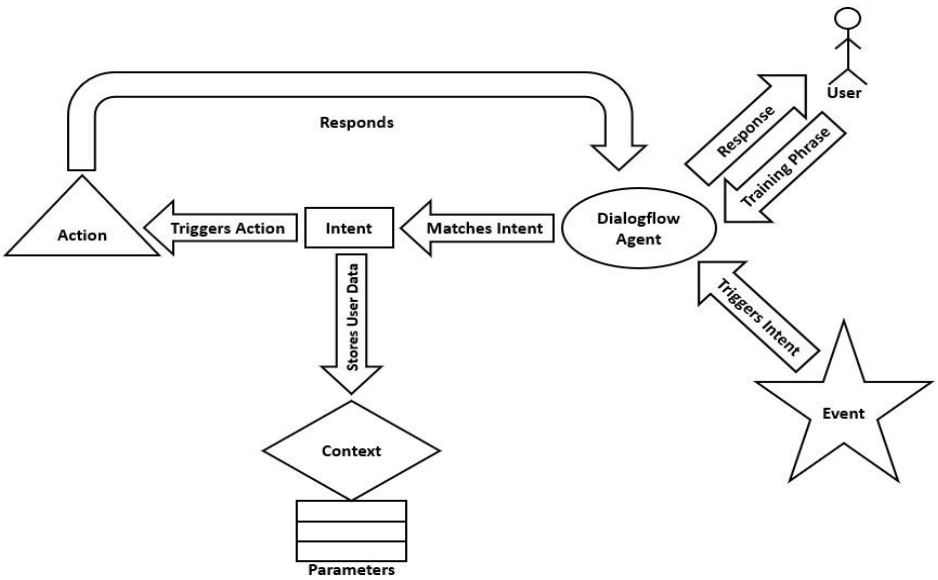
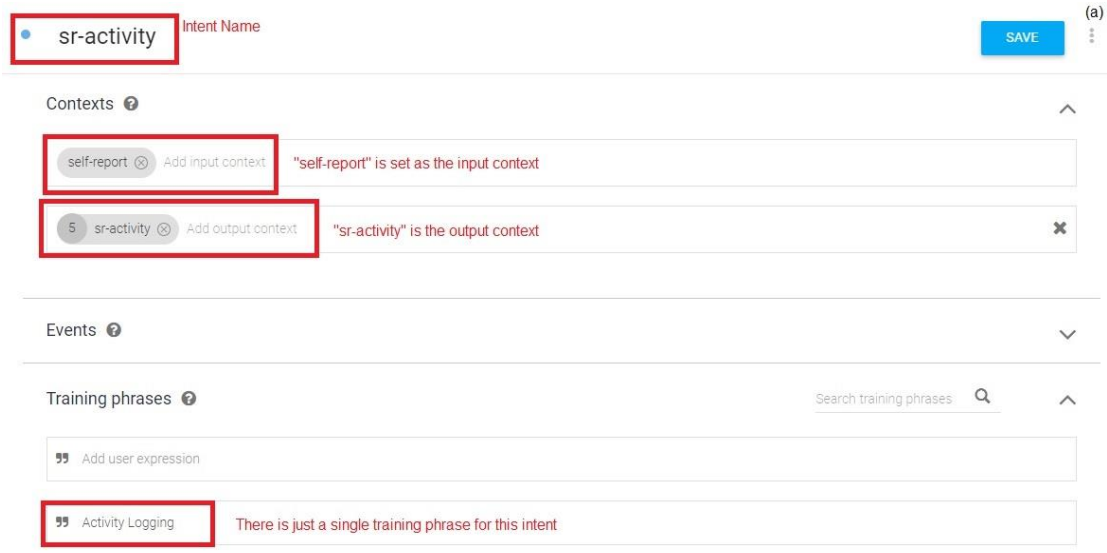


Figure 4.6: Dialogflow Agent Components

The *WeightMentor* agent was built using the Dialogflow web interface. Intents were created for each of the conversation flows identified earlier and assigned associated training phrases. Intents were assigned contexts, either an *input context*, if the intent was to be triggered as a follow-on from another intent, or an *output context*, where an intent was to trigger another intent, or both *input* and *output* contexts where the intent was part of a chain of intents. *User profile creation* represents an example of a chain of events in which an initial intent triggers one of two intents dependent upon which options are chosen by the user. In instances in which user data were to be collected, intent parameters were set. Finally, a generic response was added to each intent. Responses were not hard coded into intents, but rather were stored in a database and retrieved as needed (see section 4.3.5). **Figure 4.7** shows the Dialogflow user interface and how the component parts are implemented.



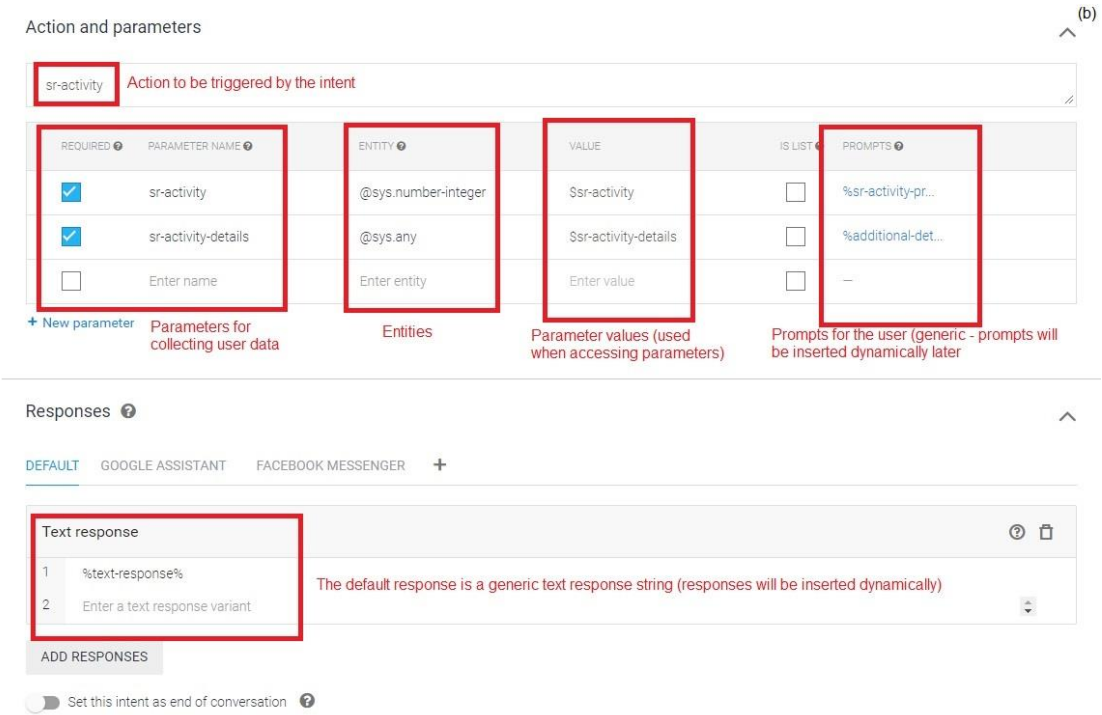


Figure 4.7: DialogFlow user interface, with key components highlighted

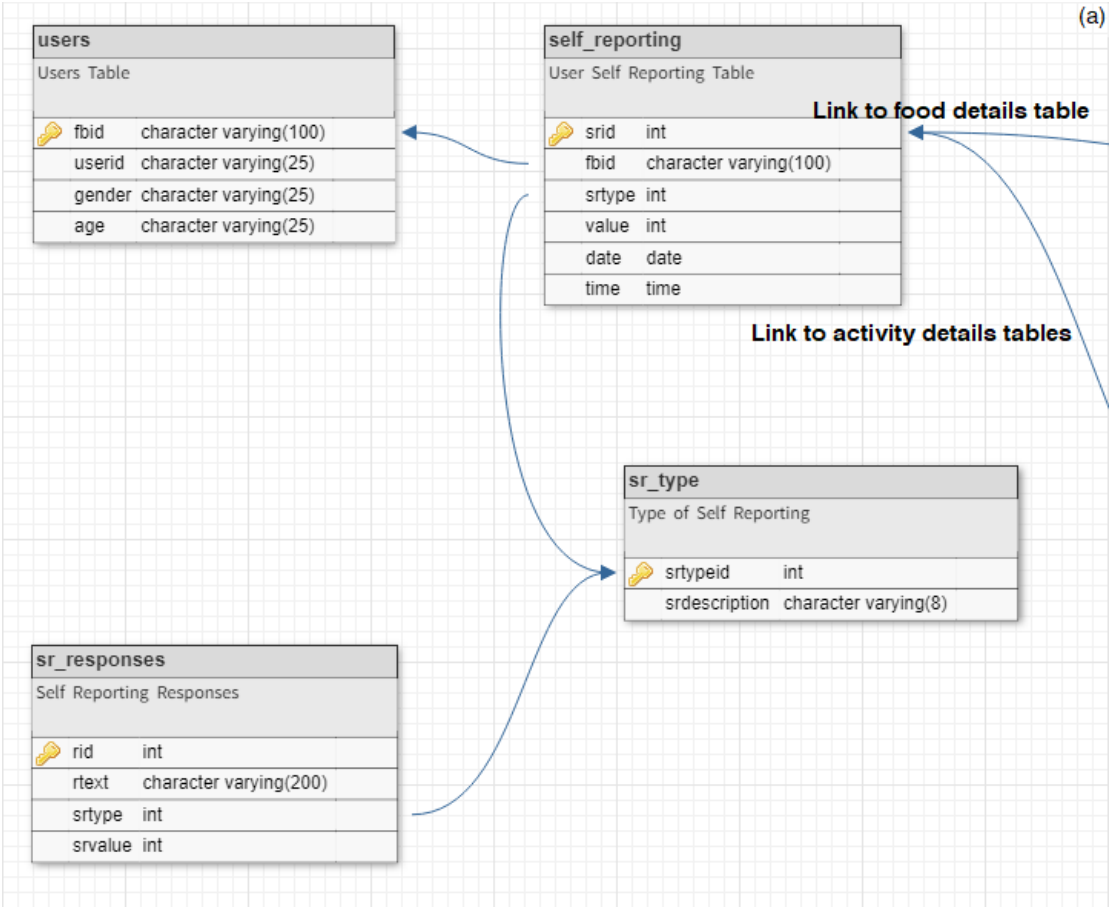
4.3.3.2 Smalltalk Agent

Included with DialogFlow are several pre-built agents, designed to extend agent functionality and interface with external Application Programming Interfaces (APIs) for navigation systems, weather forecasting, media players and home automation. One of these pre-built agents is *Smalltalk* which provides pre-defined responses to everyday "small-talk" questions outside the scope of the chatbot. Smalltalk functionality is integrated into all DialogFlow agents in later versions, but for the *WeightMentor* chatbot it was implemented as a pre-built agent. In the *WeightMentor* chatbot, Smalltalk provides responses to user questions such as "How are you?", "What is your name?" etc. Although responses to these questions are pre-defined it is possible to override these with custom responses. For example, in the *WeightMentor* chatbot, the responses to "Who are you?" and "What is your name?" were overridden to tell the user the chatbot's name and briefly explain its purpose. Responses to "Hi" or "Hello" and "Bye" or "Goodbye" were also overridden to provide custom responses using animated GIFs and customised greetings based on time of day (see section 4.3.6.4).

4.3.3.3 Database Implementation

Dialogflow is a powerful, easy to use framework that provides basic building blocks for chatbot conversation modelling; however, its functionality does have limitations. Firstly,

although it is possible to retrieve user-submitted data from parameters within the active context, it is not possible to permanently store these data. Contexts have a limited lifespan of either five interactions or ten minutes (twenty minutes in Dialogflow version 2.0), and although lifespan may be increased it is not possible to do so indefinitely. Thus, in order to permanently store and access user data it is necessary to extend DialogFlow functionality using a database, which would also permit varied self-reporting feedback and motivation. The database was developed in PostgreSQL, which is available as a free Heroku add-on. Prior to building the database, the relational schema was sketched on paper and then designed in the online tool *DbDesigner.net*. A diagram of the schema is shown in **Figure 4.8**.



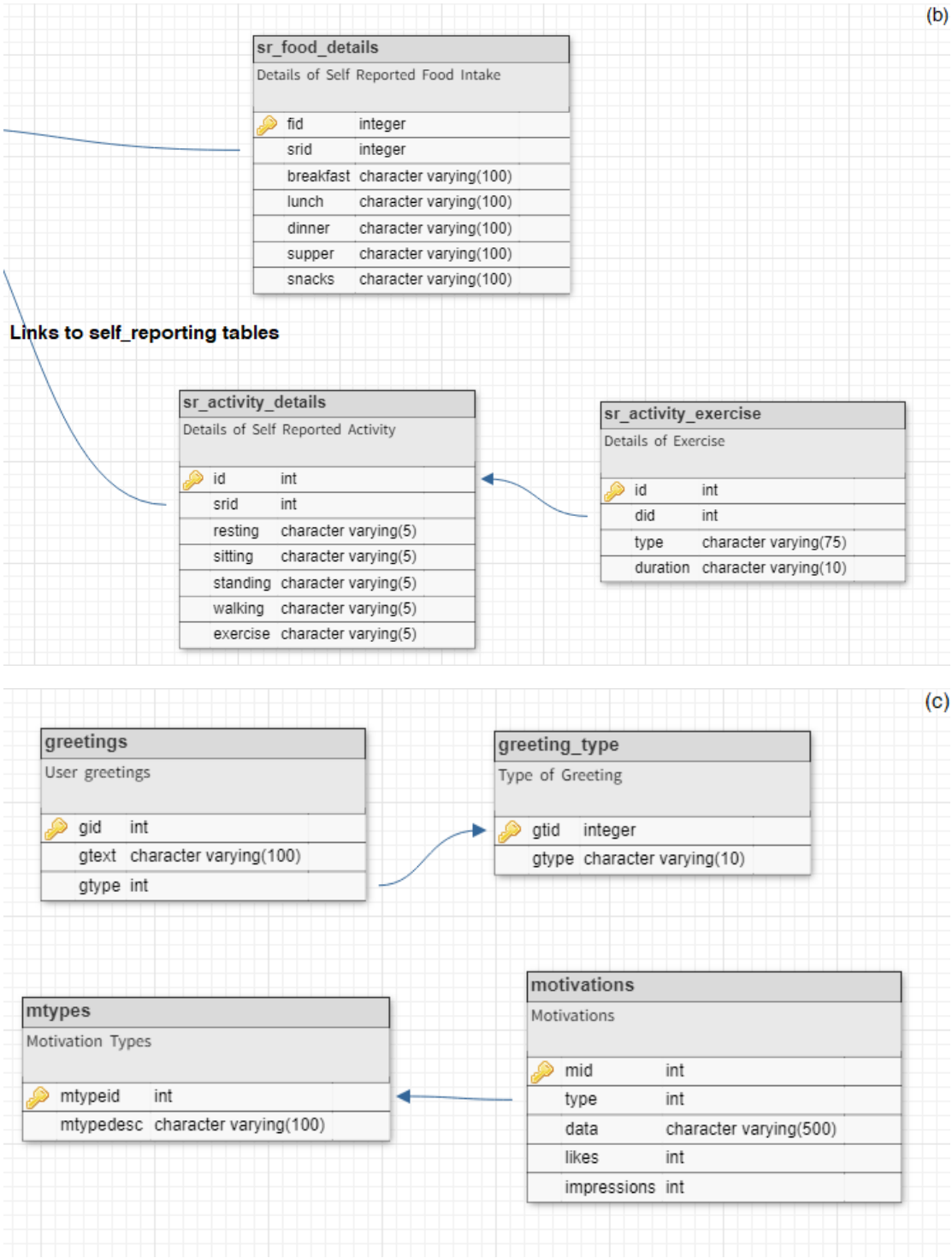


Figure 4.8: Database schema showing tables and links:
(a) Users and self-reporting tables (b) food and activity details tables (c) greetings and motivations tables

The database contains a total of eleven tables. The *users* table stores user data. The table primary key is the user's *page-scoped ID*, which is automatically generated by Facebook and unique to each user, thus is an obvious choice for the primary key in the database table. Additional data stored in the table include the *userID*, which the user will choose when they create their user profile, and may be their real name, a nickname, or something

else (such as an ID used in online gaming), *gender*, which will be *Male, Female, Other* or *Prefer not to say*, and *age*, which is collected from the user as a range (e.g. 18-25) rather than a number. Six tables store self-reporting information. The *self_reporting* table stores numeric values reported by users for physical activity and food intake, along with date and time of report (see section 3.6.2). This table is linked to the *users* table by the user's *page-scoped ID*. The type of self-report is stored in this table as a numeric value (1=physical activity, 2=food) and linked to the *sr_type* table. A third table, *sr_responses* is linked to the *sr_type* table and stores the possible responses that may be sent to the user following a self-report. Data related to self-reported physical activity are stored in two tables, *sr_activity_details* and *sr_activity_exercise*. The *sr_activity_exercise* table is linked to the *sr_activity_details* table, and the *sr_activity_details* table is linked to the *self_reporting* table. Data relating to self-reported food intake are stored in a single table, which is linked to the *self_reporting* table. Two tables store user greetings. The *greetings* table stores the greeting text and is linked to a second table, *greeting_type*, which stores the type of greeting that will be sent. The final two tables, *motivations* and *mtypes* store motivational messages that will be sent to the user.

4.3.3.4 *WeightMentor* back-end

Dialogflow includes a comprehensive API which allows the agent to be accessed and controlled externally by other programs. *Dialogflow's* basic functionality was extended in *WeightMentor* using an application written in NodeJS, which is based on JavaScript, currently one of the most popular languages for scripting web-based applications. NodeJS is a powerful, versatile language, well fit for the purpose of developing the *WeightMentor* back end application. The *WeightMentor* NodeJS application was designed using the Dialogflow and Facebook Messenger APIs, both of which are available for use at no cost. The NodeJS application is hosted online using Heroku, which is a *Platform-as-a-Service (PaaS)*. A PaaS allows developers to build, deploy and manage their own software applications without needing to implement their own server infrastructure. Heroku is suited to this purpose as it is easy to configure, includes good support for databases and has a wide range of add-ons available. The back-end application provides most of *WeightMentor's* main functions.

4.3.4 WeightMentor functions

4.3.4.1 Onboarding and User profile creation

Onboarding is a crucial part of the chatbot user experience (Martín et al. 2017). Chatbots should welcome the user and explain their scope and purpose. Doing so allows the user to manage their expectations and to gain some idea of how to interact with the chatbot from the start. As part of the onboarding process all *WeightMentor* users will create a profile for themselves the first time they use the chatbot. During profile creation the user will be asked to choose a user ID, which may be their real name or a nickname. The user ID will be used by the chatbot to facilitate recognition and to greet them personally. During the first interaction with a user, *WeightMentor* uses the *Facebook Graph API* to retrieve the user's name from their Facebook profile, and provides the option of using their real first name as their user ID or choosing something else. The first time a user interacts with *WeightMentor*, a blank user object is created in memory, which will be populated with the user's *page-scoped-ID*, user ID, gender, age range, and self-reporting data. When the user's profile is created, the data stored in this object are written to the database. In future interactions, the back-end application will check for the presence of this object to determine if the user exists in memory. If the object does not exist, then the back-end application will create a new blank user object and populate it using data loaded from the database. **Figure 4.9** shows a screenshot of the user profile creation conversation.

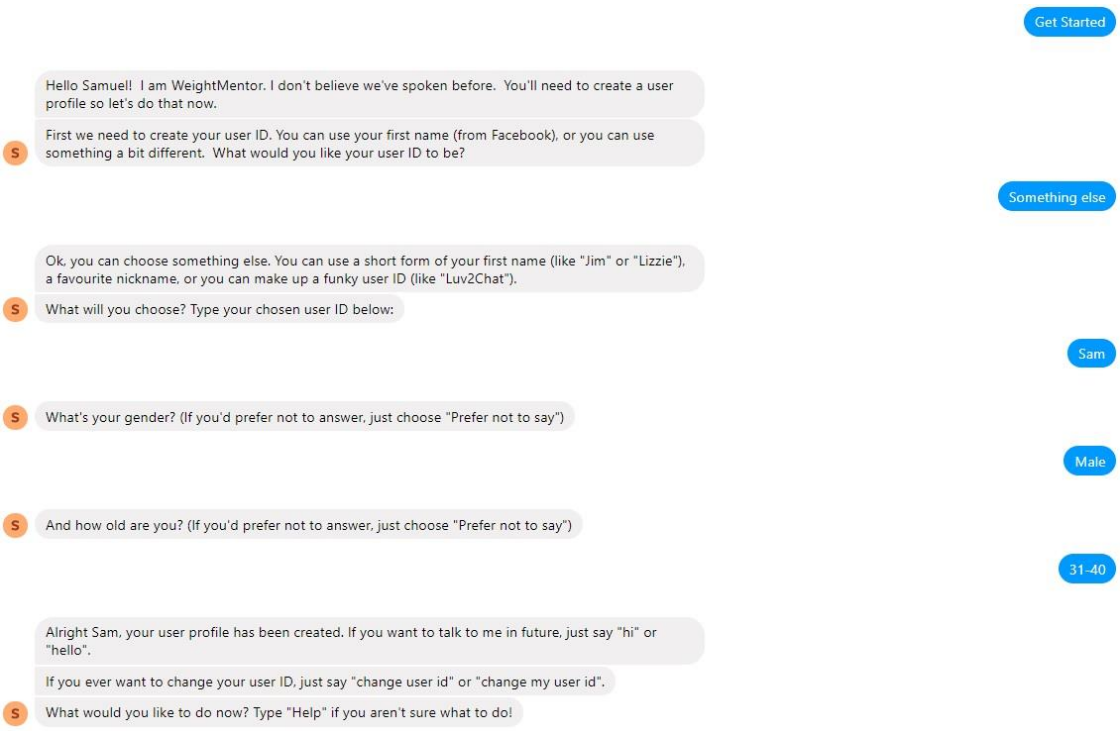


Figure 4.9: User profile creation

4.3.4.2 Self-reporting

Self-reporting in *WeightMentor* is accomplished using a five-point scale, based on the user's own perception of their physical activity and food intake. Scale points are "*Much less than usual", "Less than usual", "Usual", "More than usual"* and "*Much more than usual"*, which are recorded and stored as numerical values from 1 (much less than usual) to 5 (much more than usual). The process for self-reporting is as follows:

1. User selects "self-report" from *WeightMentor's* function menu, or types "self-report" into Facebook Messenger
2. *WeightMentor* offers the user the choice between *Physical Activity* or *Food Consumption* (assume for the purposes of this example that the user chooses *physical activity*, the process for reporting food consumption is the same)
3. The user either selects one of the options above, or types their choice into Facebook Messenger
4. *WeightMentor* presents the user with the self-reporting scale described above
5. The user selects the option that best matches their perceived physical activity level (assume for the purposes of this example that the user chooses *Usual*)
6. *WeightMentor* checks the database for previous activity reports for this user and compares any that exist with the current report

7. If enough self-reports exist (at least 5), *WeightMentor* will display a chart of the user's activity trend
8. If there are fewer than 5 self-reports, *WeightMentor* will tell the user whether their current physical activity level is higher, lower or the same as their previous report
9. The user's current report is written to the database
10. *WeightMentor* displays a personalised response based on the user's self-reported physical activity

The procedure above is handled in the *WeightMentor* back-end by a program module called *self-reporting.js*. Code for this module may be found in **Appendix 11**.

4.3.4.3 Displaying charts

Users may view charts of their self-reported physical activity and food intake trends. These may either be displayed as part of the self-reporting process (as discussed in section 3.6.2) or may be viewed by the user on demand. When designing this function, it was determined that charts should be easily viewable using a smartphone or laptop, and thus are displayed as an image within the Facebook messenger interface. Initially the *Google Charts API* was selected for displaying charts as it uses a URL (Uniform Resource Locator) - a web link - to generate the chart as an image. When this URL is sent to Facebook Messenger, the image will be displayed in the messenger interface, where it can be viewed, downloaded or printed. Unfortunately, however it was discovered during development that Google have deprecated the static chart functionality of their Charts API, thus an open-source alternative, *Image Charts* was used. Image charts is essentially a port of the Google Charts static chart functions, and functions in the same way. As with Google Charts, Image Charts generates the chart as an image from a URL. To facilitate easy building of charts a custom module that implements *Image Charts* was built within the back-end application. The module fulfils several functions: encoding chart data, plotting a standard trend chart (of all self-reported physical activity or food intake) and plotting a chart of the previous five self-reports. Code for this module can be found in **Appendix 12**.

4.3.4.3.1 Encoding chart data

Chart data was must be encoded before sending to Image Charts. In the *WeightMentor* charts module, two types of encoding are used: text encoding and simple encoding. Text

encoding sends chart data as a string of values, separated by a comma (e.g. *1,2,3,4,5*) whereas simple encoding sends a string of alphanumeric characters where the letters A to E are used to represent values from 1 to 5 (e.g. *1, 2, 3, 4, 5* would be encoded as *ABCDE*). Data sent to *Image Charts* are then identified in the chart URL as either text or simple encoding using the strings *&chd=s* or *&chd=t* respectively. Pseudocode outlining how data are encoded is shown in **Pseudocode sample 1**.

```
(a) CONVERT CHART DATA TO
    TEXT_STRING (COMMA SEPARATED)
    ADD &chd=t TO START OF
    TEXT_STRING
    OUTPUT TEXT_STRING
```

```
(b) FOR EACH VALUE IN CHART DATA
        CONVERT VALUE TO
        A CHARACTER A - E
        ADD ENCODED VALUE TO
        DATA_STRING
    ADD &chd=s TO START OF
    DATA_STRING
    OUTPUT DATA_STRING
```

Pseudocode sample 1: Image Chart Data Encoding
(a) Text Encoding (b) Simple Encoding

4.3.4.3.2 Plotting a standard trend chart

A standard trend chart is plotted on request by the user. This chart will show a line chart of all self-reported food or physical activity since the user started self-reporting. Charts are built using a string of options which is sent to Image Charts. This function encodes data using simple encoding (to reduce the length of the URL) and sets the chart colour to either blue (physical activity) or red (food intake), a URL is then generated, which, when sent to the Facebook Messenger interface, displays the chart. Pseudocode for this function is shown in **Pseudocode sample 2**, and a sample chart is shown in **Figure 4.10**.

```
SET CHART_OPTIONS (LINE CHART, NO AXES, CHART SIZE)
IF ACTIVITY_CHART
    SET COLOUR TO BLUE AND CHART_TITLE TO "ACTIVITY CHART"
ELSE
    SET COLOUR TO RED AND CHART_TITLE TO "FOOD CHART"
ADD COLOUR AND CHART_TITLE TO CHART_OPTIONS
IF CHART_DATA IS SHORTER THAN FIVE VALUES
    USE TEXT ENCODING TO ENCODE DATA
ELSE
    USE SIMPLE ENCODING TO ENCODE DATA
BUILD CHART_URL USING CHART_OPTIONS + CHART_DATA
OUTPUT CHART_URL
```

Pseudocode sample 2: Function to plot a standard chart

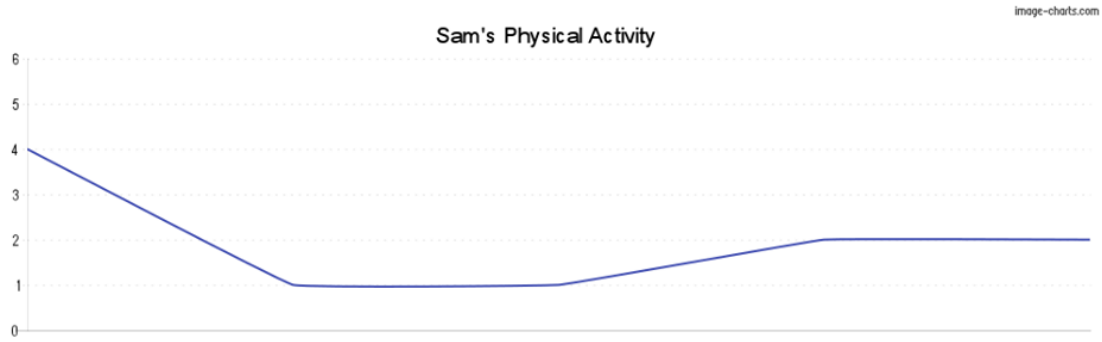


Figure 4.10: Sample standard trend chart

4.3.4.3.3 Displaying a chart of previous self-reports

A chart of previous self-reports is displayed automatically once the user has submitted at least five self-reports. *WeightMentor* will retrieve the last five self-reports from the database and display them as a two-line chart, with one line showing the previous five self-reports and the second showing the average. Charts are built in the same way as standard trend charts. Pseudocode for this function is shown in **Pseudocode sample 3**, and a sample chart is shown in **Figure 4.11**.

```
SET CHART_OPTIONS (LINE CHART, NO AXES, CHART SIZE)
IF ACTIVITY_CHART
    SET COLOUR TO BLUE, AVERAGE_COLOUR TO YELLOW AND CHART_TITLE TO
    "ACTIVITY CHART - LAST FIVE"
ELSE
    SET COLOUR TO RED, AVERAGE_COLOUR TO DARK GREEN AND CHART_TITLE
    TO "FOOD CHART - LAST FIVE"
ADD COLOUR, AVERAGE_COLOUR AND CHART_TITLE TO CHART_OPTIONS
BUILD CHART_SERIES USING LAST_FIVE_REPORTS AND AVERAGE
BUILD CHART_URL USING CHART_OPTIONS + CHART_SERIES
OUTPUT CHART_URL
```

Pseudocode sample 3: Function to plot a chart of the last five self-reports.

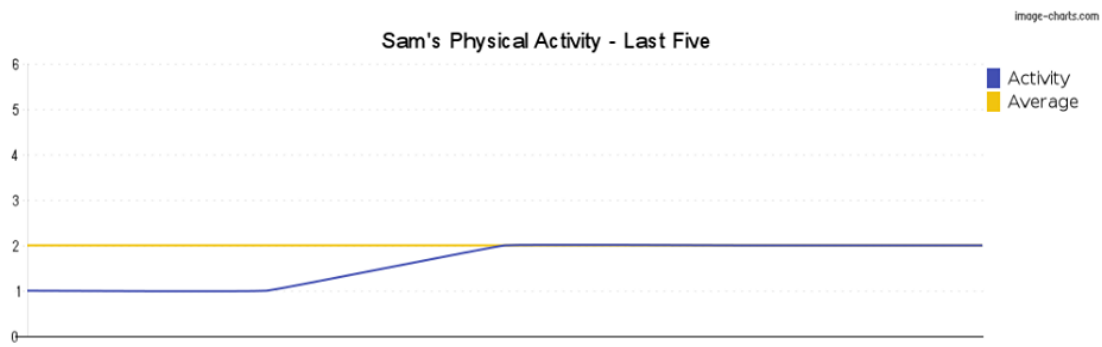


Figure 4.11: Sample "last five" chart

4.3.4.4 Variability and greetings

In order to maintain variability, a range of self-reporting responses and greetings were written for *WeightMentor* to use in conjunction with the *Smalltalk* functionality (described in section 4.3.4.2). These responses and greetings are stored in the database and retrieved at random when required. As an additional level of variability, animated GIF images are used during greetings when a user starts and finishes a conversation.

4.3.4.4.1 Variability of greeting text

Greeting text varies to reflect the time of day and uses a random selection of greetings using a random number generator. Times of day used for greeting selection are shown in **Table 4.7**. When a user interacts with *WeightMentor*, the *datetime.js* module determines which greeting should be used based on the current time of day. All appropriate greetings are then retrieved from the database and one is selected at random and displayed for the user. **Pseudocode sample 4** shows how this process is accomplished. **Figure 4.12** shows some example greetings.

Table 4.7: Times of Day for *WeightMentor* Greetings

Greeting Type	Time
Early Morning	05:00 - 07:00
Morning	07:00 - 12:00
Afternoon	12:00 - 17:00
Evening	17:00 - 22:00
Late	After 22:00

(a)
```
CHECK CURRENT_TIME
IF CURRENT_TIME IS BETWEEN
5AM AND 7AM
    USE EARLY_GREETING
ELSE IF CURRENT_TIME IS
BETWEEN 7AM AND 12PM
    USE MORNING_GREETING
ELSE IF CURRENT_TIME IS
BETWEEN 12PM AND 5PM
    USE AFTERNOON_GREETING
ELSE IF CURRENT_TIME IS
BETWEEN 5PM AND 10PM
    USE EVENING_GREETING
ELSE (CURRENT_TIME IS AFTER
10PM)
    USE LATE_GREETING
RETURN TYPE OF GREETING
```

(b)
```
GET GREETING_TIME
SELECT ALL GREETINGS FROM
DATABASE BASED ON
GREETING_TIME
PUT GREETING_TEXT INTO ARRAY
GENERATE RANDOM_NUMBER BASED
ON NUMBER OF GREETINGS IN
ARRAY
SELECT GREETING THAT
CORRESPONDS
TO RANDOM_NUMBER
OUTPUT GREETING
```

Pseudocode sample 4: Times and Greetings:
(a) getting greeting time (b) selecting a random greeting

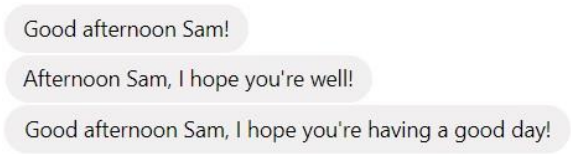


Figure 4.12: Sample Greetings

4.3.4.4.2 Generation of GIF Images

GIF images are stored in a folder on the Heroku server and are labelled “greetingXX.gif”, or “goodbyeXX.gif” where XX is a number. Images are displayed at random when a user says “hello” or “goodbye” to *WeightMentor*. The displaying of GIFs is handled by *gifs.js*, which is the GIFs module in the NodeJS code. Code for this module may be found in **Appendix 13**. A pseudocode sample of how GIFs are selected is shown in **Pseudocode sample 5**, and a sample GIF is shown in **Figure 4.13**.

```
IF TYPE_OF_GIF IS “GREETING”
    SET GIF_NAME TO “GREETING”
ELSE
    SET GIF_NAME TO “GOODBYE”
GENERATE RANDOM_NUMBER
IF RANDOM_NUMBER IS LESS THAN 10
    APPEND “0” FOLLOWED BY RANDOM_NUMBER TO GIF_NAME
ELSE
    APPEND RANDOM_NUMBER TO GIF_NAME
APPEND “.GIF” EXTENSION TO GIF_NAME
OUTPUT GIF_NAME
```

Pseudocode sample 5: Code for selection of random GIFs

Figure 4.13: Sample animated GIF

4.3.4.5 Motivation

As discussed in chapter 2 (**section 2.5**), motivation is an important aspect of weight loss maintenance, and thus the chatbot, *WeightMentor*, will provide motivational support to users. This will be accomplished using selected motivational quotes in both image and text form. Images are stored in a folder on the Heroku server, and text-based motivations are stored in the database. When a user requests motivation, a random motivational quote is retrieved (either as an image or in text form) and displayed to the user. Once a quote is displayed, *WeightMentor* asks the user if they like the quote. If the user chooses "yes", *WeightMentor* acknowledges this and displays the main menu, however if the user chooses "no", *WeightMentor* will display another random motivation. Total number of likes per motivation are recorded in the database, along with total number of "impressions", i.e. the number of times the motivation has been used. Motivation is handled by the *motivation.js* module within the NodeJS code. Code for this module can be found in **Appendix 14**. **Pseudocode sample 6** shows the procedure for motivation and updating likes and impressions, and **Figure 4.14** shows an example of the sort of motivations generated by the chatbot.

(a)
```
GET MOTIVATIONS
FROM DATABASE
GENERATE
RANDOM_NUMBER
GET MOTIVATION
BASED ON
RANDOM_NUMBER
WHILE MOTIVATION
HAS PREVIOUSLY BEEN
SENT TO USER DURING
THIS SESSION
    GET ANOTHER
    MOTIVATION
SEND MOTIVATION
UPDATE LIKES
UPDATE IMPRESSIONS
```

(b)
```
GET LIKES FROM
DATABASE
IF USER LIKED
MOTIVATION
    LIKES =
    LIKES + 1
ELSE
    LIKES =
    LIKES - 1
WRITE LIKES TO
DATABASE
```

(c)
```
GET IMPRESSIONS FROM
DATABASE
IMPRESSIONS =
IMPRESSIONS + 1
WRITE IMPRESSIONS TO
DATAB
```

Pseudocode Sample 6:
(a) Code for getting and displaying motivations (b) Code for tracking likes (c) Code for updating impressions (i.e. total number of times motivation has appeared)

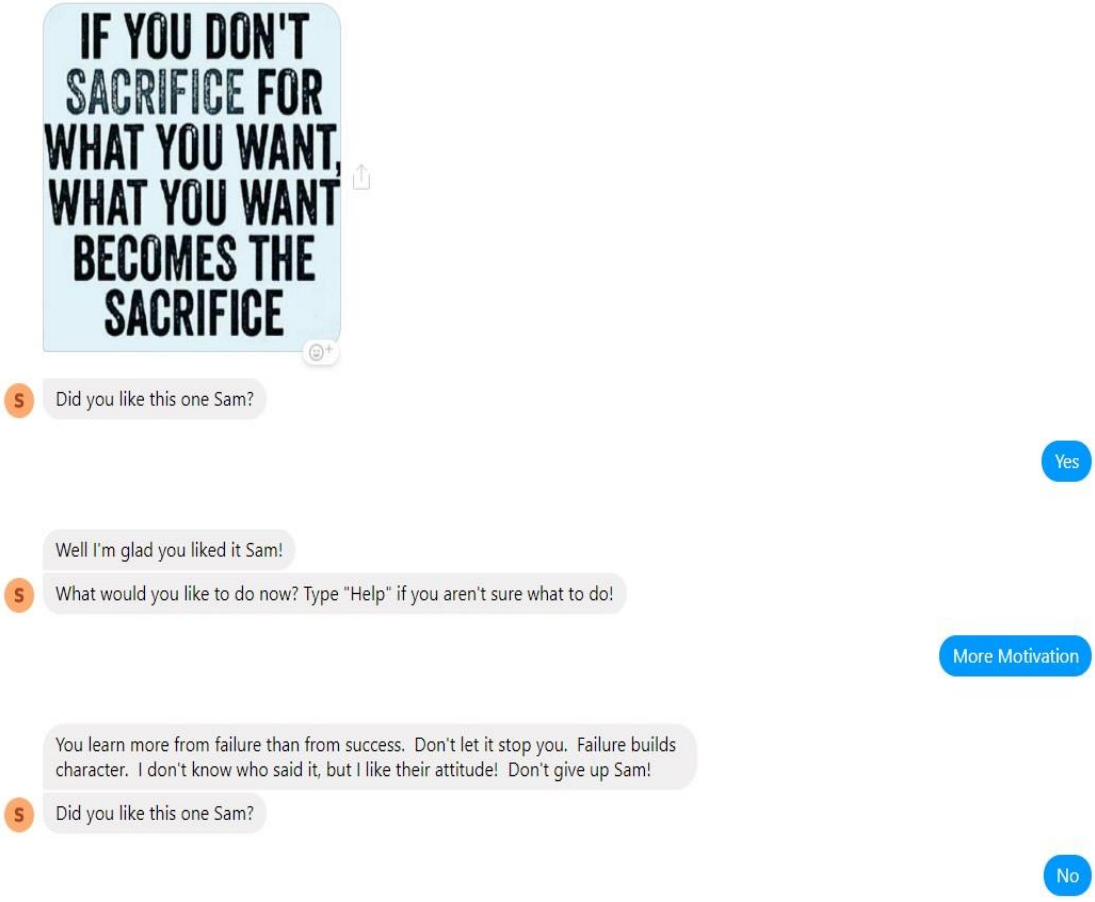


Figure 4.14: Sample Motivations

4.4 Chatbot Content Analysis

A preliminary analysis of chatbot content was conducted in July 2019. The *WeightMentor* chatbot was compared with six other chatbots related to the general area of health. Chatbots were analysed for purpose, user interface platform, response generation, dialogue initiative, input modality, target, personality/role, gender, use of

multimedia, free text entry, introduction, range of functions, motivation, use of external resources, use of social support and ratio of text to multimedia per 20 messages (based on the first 20 messages). A taxonomy of the chatbot content analysis is shown in **Table 4.8**. Findings from the content analysis are in **Tables 4.9** to **4.12**.

Table 4.8: Chatbot Content Analysis: Taxonomy
(Purpose, Platform, Response Generation, Dialogue Initiative)

Dimension	**Characteristics**				
Purpose	Advice	Support			
Platform	App	Messenger			
Response Generation	Machine-Learning	Rule-Based			
Dialogue Initiative	Chatbot-Led	User-Led	Mixture		
Input Modality	Text	Quick Replies	Both		
Personality/Role	Formal	Friendly			
Gender	Male	Female	None		
Multimedia Use	Yes	No			
Free Text Entry	None	Limited	Some	Free	
Introduction – Explains Purpose	Yes	No			
Range of Functions	Poor	Fair	Good		
Motivation	Yes	No			
External Resources	Yes	No	Unknown		
Social Support	Yes	No	Unknown		
Text to Media Ratio	Not Used	2:1	2.5:1	5:1	20:1

Table 4.9: Chatbot Content Analysis
(Purpose, Platform, Response Generation, Dialogue Initiative)

Chatbot	**Purpose**	**Platform**	**Response Generation**	**Dialogue Initiative**
Ada Health	General Health Advice/Symptom Analysis	App	Machine Learning	Mixture
Forksy	Weight Loss Advice	Messenger	Rule-Based	Chatbot-led
Jolt.ai	Weight Loss/Maintenance Support	Messenger	Rule-Based	Chatbot-led
Weight Loss Bot	Weight Loss Advice	Messenger	Rule-Based	Chatbot-led

Chatbot	Purpose	Platform	Response Generation	Dialogue Initiative
WeightMentor	*Weight Loss Maintenance Support*	*Messenger*	*Rule-Based*	*Chatbot-led*
WoeBot	Mental Health Support	App	Machine Learning	Chatbot-led
Wysa	Mental Health Support	App	Machine Learning	Mixture

Table 4.10: Chatbot Content Analysis
(Input Modality, Personality/Role, Gender)

Chatbot	Input Modality	Personality/Role	Gender
Ada Health	Both	Formal. Uses direct questions to gather symptom information from users to recommend a course of action.	Female?
Forksy	Both	Acts as an assistant. Uses self-reporting and gentle encouragement to support user.	Female?
Jolt.ai	Both	Friendly. Acts like a coach. Uses humour, graphics and social support to support user.	None
Weight Loss Bot	Both	Friendly and casual. Acts as an instructor. Uses graphics and humour to educate user.	Male
WeightMentor	Both	*Friendly. Acts as a mentor. Uses gentle encouragement, positive reinforcement and motivational quotes to support user.*	*None*
WoeBot	Both	Friendly and casual. Acts as a mental health coach. Uses daily check-ins and light-hearted humour to support user.	None
Wysa	Both	Friendly. Acts as a mental health coach. Uses regular check-ins to support the user.	None

Table 4.11: Chatbot Content Analysis
(Uses Multimedia, Free Text Entry, Introduction, Range of Functions)

Chatbot	Multimedia Use	Free Text Entry	Introduction – Explains purpose	Range of Functions
Ada Health	No	Limited	No	Fair
Forksy	No	Free	Yes	Fair
Jolt.ai	Yes	Some	Yes	Good
Weight Loss Bot	Yes	Limited	Yes	Poor
WeightMentor	*Yes*	*Free*	*No*	*Fair*
WoeBot	Some	Some	Yes	Fair
Wysa	Some	Some	No	Fair

Table 4.12: Chatbot Content Analysis
(Motivation, External Resources, Social Support, Text to Media Ratio)

Chatbot	Motivation	External Resources	Social Support	Text to Media Ratio (First 20 Messages)
Ada Health	No	Unknown	No	No Media
Forksy	No	Yes	Unknown	2:1
Jolt.ai	Yes	Unknown	Yes	5:1
Weight Loss Bot	No	Yes	No	2.5:1
WeightMentor	*Yes*	*No*	*No*	*No Media*
WoeBot	No	Unknown	No	20:1
Wysa	No	Unknown	Unknown	No Media

Four chatbots were designed for weight loss or weight loss maintenance (targeting obesity) (including *WeightMentor*), two were designed for mental health support, and one was designed for general health advice or symptom analysis. All four chatbots for weight loss or maintenance were integrated with Facebook Messenger and used rule-based artificial intelligence, as opposed to machine learning, to generate chatbot responses. Three chatbots (including both of those designed for mental health) ran on smartphones as standalone apps using machine learning. Dialogue was chatbot-led for five of the seven chatbots, and all seven used a combination of free-text and quick replies for user input.

Chatbot personalities varied, however most (n=4) adopted a friendly, casual approach to user interactions. Gentle, positive encouragement was common (n=5). Where humour was used this was appropriate and seemed to be carefully designed to support the chatbot's purpose (e.g. supporting weight loss maintenance or managing mental health. Most of the analysed chatbots (n=4) gave no indication of their assumed gender, one chatbot was determined to be male, and two were determined to be female based on the chatbot name or avatar. Only two of the chatbots offered some form of motivation functionality, and only one made use of external social support. Four chatbots used multimedia messages to create a more attractive user experience, the most favourable ratio of multimedia messages to text messages was 1:2.

4.4.1 Chatbot Text Readability

Several scales exist for assessing readability of text, but the most used are the Flesch Reading Ease Scale, and Flesch-Kincaid Grade Level (Linney, 2019). The Flesch Reading Ease scale was developed in 1948 and scores text out of 100 based on ease of

reading, with higher scores representing greater reading ease. The Flesch-Kincaid scale was developed in the 1970s as a means of interpreting Flesch scores, by converting these to equivalent education grade levels (Linney, 2019). All chatbots in the content analysis were assessed for readability using the Flesch Reading Ease scale and Flesch-Kincaid scale. Readability scores are in **Table 4.13**.

Table 4.13: Chatbot Readability Scores

Chatbot	Flesch Reading Ease Score	Flesch-Kincaid Grade Level	Equivalent Reading Age
Ada Health	77.1	4.7	10 - 11 years
Forksy	95.8	1.8	7 - 8 years
Jolt.ai	96.4	1.7	7 - 8 years
Weight Loss Bot	88.0	3.4	8 - 9 years
WeightMentor	*91.7*	*2.9*	*8 - 9 years*
WoeBot	88.6	3.2	8 - 9 years
Wysa	78.4	4.7	10 - 11 years
Mean	**88.0**	**3.2**	**8 – 9 years**

Based on these scales, the *WeightMentor* chatbot scored 91.7 on the Flesch scale, equivalent to grade 3, and easily readable by 8-9-year olds. The mean Flesch score for all chatbots was 88.0, equivalent to grade 3 on the Flesch-Kincaid scale and easily readable by 8-9-year olds. Lowest scoring chatbots were *Ada Health* and *Wysa*, scoring 77.1 and 78.4 on the Flesch scale and equivalent to grade 4.7 on the Flesch-Kincaid scale. Messages from these chatbots are easily readable by 10-11-year olds. Highest scoring chatbots were *Forksy* and *Jolt.ai*, which scored 95.8 and 96.4 on the Flesch scale, equivalent to grade 2, and easily readable by 7-8-year olds.

This content analysis study could be developed further by comparing *WeightMentor* to other types of chatbot than those related to healthcare, and by conducting more in-depth analysis into types of motivation used, conversation flows and conversation type analysis (based on Moore et al. 2017).

4.5 Discussion

WeightMentor has been developed to provide target users with a convenient, accessible tool for weight loss maintenance. The main functions of self-reporting, progress tracking and motivation were selected based on previous research suggesting that benefits may be

gained from supporting these three aspects of weight loss maintenance (Svetkey et al. 2008, Donaldson et al. 2014, Fjeldsoe et al. 2016). According to a recent review of the most popular nutrition apps on the Google and Apple stores (Franco et al. 2016), nine of the thirteen reviewed apps offered self-reporting of physical activity and a food intake logging facility but did not offer motivational coaching. The review concluded that while apps that support self-reporting are certainly popular, their functionality is limited at present, and there is potential for apps of this type to incorporate personalised feedback and nutrition recommendations (Franco et al. 2016).

Although personalised nutrition and exercise recommendations are beyond the scope of the WeightMentor chatbot, the design does allow for personalised responses based on self-reported food intake and physical activity. The chatbot's main components and functions have been carefully designed to provide the user with an engaging experience that makes use of contemporary techniques and tools for conversational AI development. Facebook Messenger is currently one of the most popular Social Media platforms (UKOM, 2018) and presents a familiar, easy to use interface for users. The use of an app that many users may already have on their smartphone eliminates the need to download a new app or to become familiar with a new user interface. Animated GIFs, randomised feedback and a variety of greetings provide variability to make the chatbot interesting and engaging for users. Personalisation is achieved by greeting the user with a userID they have chosen for themselves and recording the user's past self-reporting creates accountability and allows the user to track progress.

4.6 Conclusion

Chatbots are an emergent technology, which may have a substantial role to play in future weight loss maintenance interventions. The *WeightMentor* chatbot potentially solves problems associated with automating conventional messages, which include personalisation, lack of empathy and limited variability. *WeightMentor* is potentially of value in weight loss maintenance interventions because as a non-human agent it may freely engage with participants without being influenced by human emotions. The chatbot will not need to worry about preserving reputation or be embarrassed to discuss exercise and diet with users. Usability testing will determine the usability of the chatbot and identify any areas for improvement.

Chapter 5: Testing the usability of the "WeightMentor" chatbot for weight loss maintenance

5.1 Introduction

Chatbots are unlike traditional systems in the sense that their user interfaces do not follow conventional rules like those discussed in **section 4.1.3**. The design principles of these systems focus on attempting to replicate the complex nature of human conversation. It is therefore understandable that chatbots cannot necessarily be tested using traditional means, such as those discussed in **section 2.7**. A recent study by Baki Kocaballi et al. (2018) acknowledged that there is limited solid research into measuring the User Experience (UX) of conversational systems and evaluated six questionnaires for UX evaluation of conversational interfaces. This study concluded that there is no one questionnaire that provides enough coverage for conversational UX evaluation but acknowledged that to design such a questionnaire could potentially be very cumbersome and impractical. The authors suggested that multiple questionnaires may be necessary for assessing the usability of conversational interfaces (Baki Kocaballi et al. 2018). The most common questionnaires for measuring system usability are summarised in **Table 5.1**.

Table 5.1: Existing Usability Questionnaires
(Based on Baki Kocaballi et al. 2018)

Questionnaire	Description
AttrakDiff	AttrakDiff (Hassenzahl et al. 2003) is commonly used to measure hedonic quality of a User Interface (UI), i.e. those aspects which make a UI enjoyable to use. Although this questionnaire will also measure pragmatic quality (aspects of the UI that make it *useful*), its primary focus is hedonic quality. This questionnaire does not measure "emotional" aspects of the system, such as fun or pleasure (Baki Kocaballi et al. 2018).
Mean Opinion Scale (MOS-X)	The MOS-X (Schmidt-Nielsen, 1995) is used to evaluate perceived quality of artificial speech. This Likert scale questionnaire assesses seven different areas of speech including Listening Effort, Pronunciation, Speaking Rate and Pleasantness of voice (Baki Kocaballi et al. 2018).

Questionnaire	Description
Paradigm for Dialogue Evaluation System (PARADISE)	PARADISE is a generic measure of user satisfaction. It is widely used but has been criticised as presenting an inaccurate measure of user satisfaction, and for omitting psychometric validation and other important aspects of user satisfaction (Baki Kocaballi et al. 2018).
Speech User Interface Service Quality (SUISQ)	SUISQ is used for assessing quality of speech-based UIs. It covers *four categories: user goal orientation, speech characteristics, verbosity,* and *customer service behaviour*. It is like SASSI, but also measures *customer service behaviour* which assesses how closely system behaviour matches expected behaviour of human customer service agents (Baki Kocaballi et al. 2018).
Subjective Assessment of Speech System Interfaces (SASSI)	SASSI (Hone & Graham, 2000) is a standardised tool for measuring satisfaction with speech-based UI. It assesses six main categories: *system response accuracy, likeability, cognitive demand, annoyance, habitability,* and *speed*, but also assesses *habitability*, which is the extent to which the user knows what they are doing. It is limited in that it only focuses on speech input quality (Baki Kocaballi et al. 2018).
System Usability Scale (SUS)	SUS is the most common instrument for measuring system usability. It was designed to be a quick and easy tool and is based on a ten-item Likert scale. It provides a general assessment of usability and so is suitable for a range of applications. (Brooke, 1996)
User Experience Questionnaire (UEQ)	The User Experience Questionnaire (UEQ) is very similar to AttrakDiff, and assesses six areas of usability: *Attractiveness, Stimulation, Novelty, Perspicuity, Efficiency and Dependability*. Where AttrakDiff focuses primarily on Hedonic Quality, the UEQ focuses on both Pragmatic and Hedonic Quality (Laugwitz et al. 2008).

Two existing questionnaires were selected for evaluating the usability of WeightMentor, the SUS (Brooke, 1996) and the User Experience Questionnaire (Laugwitz et al. 2008). SUS was selected for use during usability tests because it is the most common questionnaire for usability testing and is known to be reliable and effective. The UEQ was selected to complement the SUS as it provides a wider overview of the User Experience and covers both hedonic and pragmatic quality. A third questionnaire, the Chatbot Usability Questionnaire, developed as part of this PhD was also used during *WeightMentor* usability testing and was validated in a later study (**discussed in Chapter 6**).

5.2 Aim

To test the usability of WeightMentor, a chatbot for weight loss maintenance.

5.3 Methods

5.3.1 Participants and Sample Size

Participants were adults (over 18 years). It was not necessary for participants to have lost weight or be maintaining weight loss, as this study assessed the usability of the chatbot, rather than its effectiveness for weight loss maintenance. Participants were required to be able to use a smartphone or laptop and needed to be willing to use a chatbot. Participants needed access to Facebook Messenger, however if they did not use Facebook Messenger or were unable to access it, they were able to use the researcher's mobile device and Facebook account. The inclusion and exclusion criteria are summarised in **Table 5.2**. Participant demographics are discussed in section 4.1.

Table 5.2: Usability Test Participant Selection Criteria

Inclusion criteria	Exclusion criteria
Adults aged 18+ years	Not providing consent
Able to use a smartphone or laptop	
Willing to use the chatbot, *WeightMentor*	

For the usability tests, participants were selected using convenience sampling, which selects participants from a population of individuals who are *conveniently available* to the researcher (Dudovskiy, 2018a). This type of sampling is popular because it is simple and easy to conduct and allows data collection to commence very quickly (Dudovskiy, 2018a). As the only requirements for participation were to be aged 18+, able to use a smartphone or laptop and willing to use the chatbot, *WeightMentor*, it was determined to be most convenient to select participants from staff, students and other PhD researchers within the Ulster University, and also from those individuals who had previously participated in Needs Analysis interviews. It was determined that approximately ten participants would be recruited for individual usability tests, which would adopt a more formative approach than the group tests and provide qualitative *concurrent think-aloud* data. This sample size is based on work by Nielsen & Landauer (1993) which suggests that no more than 5 to 8 participants are required to identify 80% of chatbot usability issues and would account for an attrition rate of 20% (2 participants). A sample size of 50 participants was selected for the group usability tests, as this type of test would follow a summative approach and it was desirable to gather as much data as possible from these tests in a short time.

5.3.2 Recruitment

Participants were recruited from staff, students and PhD researchers at the Ulster University and from Needs Analysis participants who had indicated that they were willing to participate in additional research studies. A recruitment email (see **Appendix 15**) was circulated to staff and students via the university mailing list. A similarly worded email was also circulated to other PhD researchers and staff in the Faculty of Arts, Humanities and Social Sciences by the Faculty Administrative Officer, and to Needs Analysis participants by the researcher. The researcher also recruited participants from within computer science classes, social work/counselling classes and tutorial groups by visiting classes and groups in person to explain the purpose of the research and distribute the Participant Information Sheet (PIS) (see **Appendix 16**) to interested students. Individuals who wished to respond to the recruitment email contacted the PhD researcher at his university e-mail address and requested a copy of the PIS. Friends of the researcher who were interested in participating were sent a copy of the PIS by email or Facebook Messenger. Individuals who met the inclusion criteria and were willing to participate signed two copies of the consent form at the start of the usability tests, one of which was retained by the participant and the other filed by the researcher. The consent form is shown in **Appendix 17**.

5.3.3 Procedures

Usability tests lasted for approximately 30 - 60 minutes. Initial group tests revealed a significant usability issue in that the chatbot was only capable of storing one user profile at a time. Consequently, the *WeightMentor* chatbot could only be used by one person at a time. While the researcher attempted to fix this issue, group usability tests were abandoned in favour of individual tests. Once a workable solution to this issue was identified and implemented, the final tests were conducted as group sessions.

Participants were given access to the WeightMentor chatbot by adding them to the *WeightMentor* Facebook application as a test user. Before commencing test tasks, participants completed a pre-test demographic questionnaire (see **Appendix 18**). This questionnaire was hosted online via Qualtrics, and participants completed it using either the researcher's laptop computer (during individual sessions) or a university computer (during group sessions). Participants attempted four tasks (see **Appendix 19** for task details). Before each task, the researcher read the task aloud and asked participants what

they thought they would need to do to complete it (to assess understanding). Participants were asked the pre-task Single Ease Question: "On a scale of 1 to 7, how difficult do you think it will be to complete this task?" and their score was recorded using the Single Ease Question recording form (see **Appendix 20)** hosted online via Qualtrics. Participants attempted each of the four tasks, and their attempt was recorded using either a MOD1000 camera if they were using their mobile phone, or screen recording software if they were using a laptop or desktop PC. After a task, whether it was completed or not, participants were asked the post-task Single Ease Question: "On a scale of 1 to 7, how difficult was this task to complete?" and, again, their score was recorded using the Single Ease Question recording form. For all usability tests, a score of 1 for a Single Ease Question represented "very difficult", while a score of 7 represented "very easy". During tasks, individual participants described what they could see on the screen, what the chatbot was asking them to do, and what they were going to click or tap on to accomplish their task. This is known as the *concurrent think-aloud protocol*. Participants' voice audio was also recorded for later analysis.

5.3.4 Questionnaires

Once participants had attempted all four tasks, they completed three questionnaires online using Qualtrics: the Post-Test Questionnaire, based on the ten item System Usability Scale (Brooke, 1996) (see **Appendix 21**), the User Experience Questionnaire (Laugwitz et al. 2008) (see **Appendix 22**), and a custom Chatbot Usability Questionnaire designed specifically for testing WeightMentor (see **Appendix 23**), which is inspired by the ALMA automated testing tool (Martín et al. 2017), discussed in section 5.3.5.1.

5.3.4.1 System Usability Scale (SUS)

The System Usability Scale (SUS) was developed by John Brooke in 1986 in order to address the need for a quick and easy scale for measuring general usability. Today, SUS is the most popular questionnaire for usability testing, having become the industry standard (Sauro, 2011). It consists of ten standardised questions which are scored on a five-point Likert scale, where 1 represents "strongly disagree" and 5 represents "strongly agree". Odd-numbered questions relate to positive aspects of the system (e.g. "I think I would like to use this system frequently") while even-numbered questions relate to negative aspects of the system (e.g. "The system was very complex") (Brooke, 1996). Over the years, extensive use of SUS in computer science research and evaluation has led

to the development of the SUS benchmark, which is based on data from hundreds of usability evaluations involving thousands of users (Sauro, 2011). The average (benchmark) score is 68.0, and thus scores greater than 68 may be considered "above average", while those less than 68.0 may be considered "below average".

Although out of 100, the SUS score is not generally treated as a percentage but may be interpreted in several ways. SUS scores may firstly be converted to grades from A+ (superior performance) to F (failing performance), with the benchmark score of 68.0 being equal to a grade C (Sauro, 2018). The benchmark score is at the 50th percentile, so, in addition to using the grading system, higher or lower scores may be assigned a percentile rank (Sauro 2018). A third type of comparison involves the use of adjectives to describe the user experience, based on Bangor et al. (2009). A total of 1,000 SUS scores were associated with a scale of seven adjectives (such as "Good", "Poor", "Excellent" etc.) used to describe usability. Using this scale, scores above 85.0 are "Excellent", anything above 71.0 is "Good", and above 51.6 is "OK". Bangor et al. (2008) also proposed evaluation of a SUS scale in terms of "acceptability", with SUS scores higher than 71 being classed as "Acceptable", and below 50 "Unacceptable". The Net Promoter Score (Reichheld, 2003) may also be used to interpret SUS Scores (Sauro, 2018). This scale is based on the question "How likely is it that you would recommend our company to a friend or colleague?" which respondents may grade from zero to ten. Respondents who score a 9 or 10 are "Promoters", those who score six or lower are "Detractors", and scores of seven and eight are "Passives". Sauro (2012) successfully correlated the NPS with the SUS, such that a SUS score of 81 or higher may be classified as a "Promoter", while "Detractor" classification may be assigned to scores of 53 or lower, and "Passive" will be assigned to any score in between (Sauro, 2012). These scales are summarised in **Table 5.3**. The SUS Questionnaire may be found in **Appendix 21**, as part of the Post-Test Questionnaire.

Table 5.3: Scales for Interpreting SUS Scores

SUS score	Grade	Percentile range	Adjective	Acceptability	Net Promoter Score
84.1 - 100	A+	96 - 100	Best Imaginable	Acceptable	Promoter
80.8 - 84.0	A	90 - 95	Excellent		

<table>
<tr><th>SUS score</th><th>Grade</th><th>Percentile range</th><th>Adjective</th><th>Acceptability</th><th>Net Promoter Score</th></tr>
<tr><td>78.9 - 80.7</td><td>A-</td><td>85 – 89</td><td rowspan="4"></td><td rowspan="6">Acceptable</td><td rowspan="2">Promoter</td></tr>
<tr><td>77.2 - 78.8</td><td>B+</td><td>80 – 84</td></tr>
<tr><td>74.1 - 77.1</td><td>B</td><td>70 – 79</td><td rowspan="6">Passive</td></tr>
<tr><td>72.6 - 74.0</td><td>B-</td><td>65 – 69</td></tr>
<tr><td>71.1 - 72.5</td><td>C+</td><td>60 - 64</td><td rowspan="3">Good</td></tr>
<tr><td>65.0 - 71.0</td><td>C</td><td>41 - 59</td><td rowspan="3">Marginal</td></tr>
<tr><td>62.7 - 64.9</td><td>C-</td><td>35 - 40</td></tr>
<tr><td>51.7 - 62.6</td><td>D</td><td>15 - 34</td><td>OK</td><td rowspan="3">Detractor</td></tr>
<tr><td>25.1 - 51.6</td><td>F</td><td>2 - 14</td><td>Poor</td><td rowspan="2">Not Acceptable</td></tr>
<tr><td>0 - 25.0</td><td>F</td><td>0 - 1.9</td><td>Worst Imaginable</td></tr>
</table>

5.3.4.2 User Experience Questionnaire (UEQ)

The User Experience Questionnaire was developed in 2005 as a means of facilitating quick assessment, giving a comprehensive impression of user experience that allows users to simply express their feelings about the system being tested (Laugwitz et al. 2008). It is based on twenty-six items, each made up of pairs of opposite terms that may be used to describe a system (e.g. “Annoying/Attractive”). A seven-point scale is used to score each item, from -3 to +3, with a neutral score of zero. Each item is linked with one of six scales used to measure the User Experience (Schrepp 2019b). Attractiveness is a measure of the overall user impression of the system (i.e. do users like it?) Perspicuity measures how easy it is to learn to use the system and to become familiar with it. Efficiency measures system reaction time and user effort required for task completion. Dependability is concerned with system security and predictability, and the degree to which the user feels in control of the system. Stimulation measures how enjoyable (i.e. exciting) it is to use the system (is it fun?). Finally, Novelty is concerned with the creativity of the system and whether it is interesting for users.

It should be noted that UEQ scores for each scale are subjective, based on participants' own opinions of the system under test. On their own, they may be used to assess the general user expectations of the system. Scale values below -0.8 are negative, values greater than 0.8 are positive, and those in the middle are neutral. These six scales may further be divided into Attractiveness, Hedonic Quality (non-task related quality - Stimulation and Novelty), and Pragmatic Quality (task-related quality - Perspicuity, Efficiency and Dependability) (Schrepp. 2019b). To compare UEQ scores with those of other systems, Schrepp (2019b) developed a benchmark based on data from analysis of a range of products (business software, social networks and websites) (Schrepp. 2019b). As with SUS, a benchmark has been developed for the UEQ. This is updated annually and is based on analysis of data from hundreds of evaluations involving many thousands of participants (Schrepp 2019b). Benchmark scores for each category are presented in **Table 5.4.** A sample UEQ can be found in **Appendix 22**.

Table 5.4: UEQ Benchmark Scores

	Attractiveness	**Perspicuity**	**Efficiency**	**Dependability**	**Stimulation**	**Novelty**
Excellent	≥1.75	≥1.9	≥1.78	≥1.65	≥1.55	≥1.4
Good	≥1.52 <1.75	≥1.56 <1.9	≥1.47 <1.78	≥1.48 <1.65	≥1.31 <1.55	≥1.05 <1.4
Above Average	≥1.17 <1.52	≥1.08 <1.56	≥0.98 <1.47	≥1.14 <1.48	≥0.99 <1.31	≥0.71 ≥1.05
Below Average	≥0.7 <1.17	≥0.64 <1.08	≥0.54 <0.98	≥0.78 <1.14	≥0.5 <0.99	≥0.3 <0.71
Bad	<0.7	<0.64	<0.54	<0.78	<0.5	<0.3

5.3.5 Chatbot Usability Questionnaire

The Chatbot Usability Questionnaire (CUQ) has been developed as an alternative to the System Usability Scale. The CUQ is a sixteen-item questionnaire, similar in design to the SUS, but with questions that are more specific to chatbots. Eight CUQ questions relate to positive aspects of the chatbot, and eight relate to negative aspects. Each question is scored out of five. CUQ Questions are listed in **Table 5.6** and a sample CUQ can be found in **Appendix 23**.

5.3.5.1 The ALMA chatbot test tool

The CUQ is inspired by the ALMA automated chatbot testing tool (Martín et al. 2017). This tool, which is deployed as an extension for the Google Chrome web browser is designed to work with Facebook Messenger-based chatbots and provides a short heuristic evaluation that helps to identify problems and issues the chatbot's users may experience during normal use. Results are presented graphically in a "radar" diagram (see **Figure 5.1**) and textually as a list of suggestions for improving the chatbot's design. ALMA assesses seven areas of the chatbot usability experience (UX), these are summarised in **Table 5.5**.

Table 5.5: Chatbot UX aspects measured by the ALMA tool

Area	Description	Example tests
Answering	What elements does the chatbot send and how well does it do it?	Does the chatbot use media/emojis? Are messages well formatted/do they contain typos?
Error Management	How well does the chatbot deal with errors that are going to happen?	Does the chatbot have different responses for the same error? Do answers offer a solution?
Intelligence	Does the chatbot have any intelligence?	Can the chatbot understand sentences where key words are missing? Can the chatbot understand multiple contexts within the same sentence (e.g. "What's the weather like in Rome? And London?"
Navigation	How easy is it to work through the conversation and its different flows?	How many steps are needed to get meaningful information from the chatbot? Does the chatbot allow you to go back and change answers to previous questions?
Onboarding	Can users understand what the chatbot is for, and how to interact with it, from the very beginning?	Does the chatbot introduce itself? Does it explain its scope?
Personality	Does the chatbot have a personality that fits with users and the ongoing conversation?	Does the chatbot have a profile picture? Do the chatbot's voice and tone fit with the purpose?
Understanding	If users are permitted to type freely, how well can the chatbot understand them?	Can the chatbot understand variations of answers (e.g. "It would be nice" instead of a simple "Yes")? Can the chatbot understand/recognise different languages?

5.3.5.1.1 ALMA *WeightMentor* Chatbot testing

The ALMA test tool was used prior to commencing recruitment for the *WeightMentor* usability tests. The tool was installed on the Chrome browser on the researcher's laptop PC, and the researcher followed the instructions given by the tool while interacting with the chatbot and recording the results in the ALMA Chrome extension. The results of the ALMA test of the *WeightMentor* chatbot are shown in **Figure 5.1**. From **Figure 5.1** it may be observed that *WeightMentor's* average score was 57%, and the chatbot scored highest in *Error Management* and *Chatbot Answering*. *WeightMentor* scored lowest in *Chatbot Understanding* and *Navigation*.

Figure 5.1: ALMA Radar diagram presenting *WeightMentor* chatbot results

5.3.5.1.2 Development of the CUQ based on the ALMA tool

The CUQ was inspired by the ALMA chatbot test tool. A series of questions related to each of the seven ALMA areas were devised by the researcher, and the questionnaire was further developed to include questions related to positive and negative aspects, to be comparable to the System Usability Questionnaire (SUS) (see section 5.3.4.1). The final set of questions were agreed with the research team.

5.3.5.1.2 CUQ Question Selection

Questions were selected for inclusion in the CUQ by the researcher and by Dr Raymond Bond and Emeritus Professor Michael McTear from the School of Computing. Questions were selected to evaluate the seven aspects of Chatbot Usability that were considered most important by the research team, most of which were measured by the ALMA tool. Questions were written to assess each aspect using balanced pairs of positive and negative questions for each aspect. Question areas identified by the research team are presented in **Table 5.6**, which also shows how these aspects relate to those measured by the ALMA tool, and the possible questions identified for each aspect. Refined questions which were selected for inclusion in the final questionnaire are shown in italics. The final questions are listed in **Table 5.7.**

Table 5.6: Chatbot UX aspects used during CUQ question design

Area	Measured by ALMA?	Possible Questions	
		Positive	**Negative**
Personality	Yes	The chatbot had a nice personality	The chatbot's personality was not very nice
		The chatbot's personality was realistic and engaging	*The chatbot seemed too robotic*
Welcome	Yes[a]	The chatbot was welcoming	The chatbot was unfriendly
		The chatbot was welcoming during initial setup	*The chatbot seemed very unfriendly*
		The chatbot introduction was helpful	The chatbot did not introduce itself well
		The chatbot explained its scope and purpose well	*The chatbot gave no indication as to its purpose*

Area	Measured by ALMA?	Possible Questions	
		Positive	**Negative**
Navigation	Yes	I could navigate the chatbot easily	I found myself getting lost in the chatbot
		The chatbot was easy to navigate	*It would be easy to get confused when using the chatbot*
Intelligence & Understanding	Yes[b]	*The chatbot understood me well*	*The chatbot failed to recognise a lot of my inputs*
Response	Yes[c]	*Chatbot responses were useful, appropriate and informative*	*Chatbot responses were irrelevant*
Error Handling	Yes	*The chatbot coped well with any errors or mistakes*	*The chatbot seemed unable to handle any errors*
Ease of Use	No	The chatbot was easy to use	The chatbot was difficult to use
			The chatbot was very complex

a - The ALMA chatbot test tool evaluates this aspect as "Onboarding"
b – The ALMA chatbot test tool evaluates this aspect as two separate aspects, "Intelligence" and "Chatbot Understanding"
c – The ALMA chatbot test tool evaluates this aspect as "Chatbot Answering"

Table 5.7: CUQ Questions

Question	Text
1	The chatbot's personality was realistic and engaging
2	The chatbot seemed too robotic
3	The chatbot was welcoming during initial setup
4	The chatbot seemed very unfriendly
5	The chatbot explained its scope and purpose well
6	The chatbot gave no indication as to its purpose
7	The chatbot was easy to navigate
8	It would be easy to get confused when using the chatbot
9	The chatbot understood me well
10	The chatbot failed to recognise a lot of my inputs
11	Chatbot responses were useful, appropriate and informative
12	Chatbot responses were irrelevant
13	The chatbot coped well with any errors or mistakes
14	The chatbot seemed unable to cope with any errors
15	The chatbot was easy to use
16	The chatbot was very complex

5.3.6 Data Analysis

The usability questionnaires, pre-test questionnaire and single ease questions were digitised on Qualtrics. Data from each questionnaire were exported from Qualtrics as a Comma Separated Value (CSV) file and imported into R Studio or Microsoft Excel. SUS and UEQ scores were calculated using the SUS Score Calculation Tool (MeasuringUX.com) and the UEQ Data Analysis Tool respectively, for analysis.

5.3.6.1 System Usability Scale

Equation 1 shows the mathematical formula for calculating SUS scores.

$$\overline{SUS} = \frac{1}{n}\sum_{i=1}^{n} norm \cdot \sum_{j=1}^{m} \begin{cases} q_{i,j} - 1, q_{i,j}\, mod\ 2 > 0 \\ 5 - q_{i,j}, otherwise \end{cases} \cdot \qquad (1)$$

where n is the number of subjects (questionnaires); m is set to 10 (number of questions); $q_{i,j}$ is the individual score per question per participant and norm is set as 2.5.

The simplest means of calculating SUS scores is to use the SUS Score Calculation Tool from MeasuringUX.com, however scores may also be calculated manually using the following procedure:

1. For each question, assign a score from 1 to 5 based on the level of agreement with the statement in the question (i.e. “Strongly agree” is worth 5 points, “Neutral” is worth 3 points, “Strongly disagree” is worth 1 point)
2. Calculate the sum of all the **odd-numbered** (positive) questions.
3. Calculate the sum of all the **even-numbered** (negative) questions.
4. Subtract 5 from the score *step 2.*
5. Subtract the score from *step 3* from 25.
6. Add the scores from *steps 4 and 5*. This gives a score out of 40.
7. Multiply the score from *step 6* by 2.5. This gives a score out of 100.

The final SUS score is out of 100. Mean SUS Score was calculated using R Studio, along with mean and standard deviation of responses to individual SUS questions.

5.3.6.2 User Experience Questionnaire

The UEQ Data Analysis Tool (Schrepp, 2019a) analyses the UEQ Questionnaire results and calculates the mean and standard deviation for each scale, which are presented graphically. It is also possible to conduct comparison with benchmark using the tool. By default, the UEQ does not calculate a single score, as is the case with SUS; instead, a single score is calculated for each of the six UEQ scales. In order to measure correlation across questionnaires, a mean score was calculated for each participant, based on the mean score across all six scales.

5.3.6.3 Chatbot Usability Questionnaire

Data from the CUQ were exported from Qualtrics as a CSV and imported into Microsoft Excel. Scores were calculated based on sixteen questions from the questionnaire. The first fourteen questions were selected, along with question sixteen and seventeen, which were swapped. This procedure created a sixteen-item instrument in which the odd numbered questions were positive and even numbered questions were negative. Scores were calculated out of 100 using the formula in **Equation 2**.

$$\overline{CUQ} = \frac{(\sum_{n=1}^{m} 2n - 1 - 8) + (40 - \sum_{n=1}^{m} 2n)}{64} \times 100 \qquad (2)$$

where m = 16 (number of questions), n = individual question score per participant

Score calculation was facilitated using a spreadsheet, however scores may also be calculated manually using a similar procedure to SUS, as detailed below:

1. For each question, assign a score from 1 to 5 based on the level of agreement with the statement in the question (i.e. “Strongly agree” is worth 5 points, “Neutral” is worth 3 points, “Strongly disagree” is worth 1 point).
2. Calculate the sum of all the **odd-numbered** (positive) questions.
3. Calculate the sum of all the **even-numbered** (negative) questions.
4. Subtract 8 from the score from *step 2*.
5. Subtract the score from *step 3* from 40.
6. Add the scores from *steps 4 and 5*. This gives a score out of 64.
7. Divide the score from *step 6* by 64 and multiply the answer by 100. This gives a score out of 100.

5.3.6.4 Task Completion Times

As discussed in the protocol section above, participants' attempts at each task were video and audio recorded. Videos of each task were reviewed by the PhD researcher and start and stop times for each task were recorded, per participant, in a spreadsheet. Where participants were asked to repeat a task (i.e. tasks 2 and 4), start and stop times were recorded for each iteration of the task. Task completion time (in seconds) per participant was calculated along with the mean time per task. For repeated tasks, the mean completion time overall was also calculated. A benchmark task completion time was established by recording the time taken for the researcher to complete each task.

5.3.6.5 Usability Issues

Usability issues were identified from participant feedback, submitted through the post-test questionnaire and from concurrent think-aloud audio recording. Usability issues were listed in a spreadsheet in the order in which tests were conducted. Issues reported by the first participant were treated as unique, and subsequent issues were counted as unique only if they had not been identified by previous participants. Unique usability issues identified by participants were compared with the best- and worst-case testing scenarios, which are used to identify the smallest and largest numbers of participants required to identify most of the usability issues. In the best-case testing scenario, participants are sorted in descending order, based on the number of unique issues identified per participant. Where participants identified the same number of issues, these were listed in chronological order. In this scenario the first participant identified the most usability issues, and the last participant identified the least. In the worst-case testing scenario, participants are sorted in ascending order, based on the number of unique issues identified per participant. In this scenario the first participant identified the fewest usability issues, and the last participant identified the most. Line charts of each scenario were compared to identify the optimum number of users (the point where a plateau is reached) and determine the mean number of users required to identify most of the usability issues.

5.3.7 Ethics

Ethical approval was obtained from the Communication Ethics Filter Committee and University Research Ethics Committee, and the study was conducted in compliance with the Ulster University Code of Practice for Professional Integrity in the Conduct of

Research and the Policy for the Governance of Research Involving Human Participants. The study was designed to comply with the five core principles of beneficence, non-maleficence, honesty and integrity, confidentiality and informed consent, which are discussed in chapter 3. This research study was designated low risk as participants are not required to participate in sensitive discussion or disclose highly personal information. Identified potential risks to participants included inconvenience of access and potential distress if participants were unable to complete tasks during the usability tests. Inconvenience of access was mitigated by conducting usability tests in a location that was convenient for participants. Most participants were recruited from staff and students at the Jordanstown campus of the Ulster University; thus, usability tests were conducted on this campus. Participants who were recruited through computer science and social work/counselling classes were given the option of participating in usability tests during their normal scheduled practical/tutorial sessions. Participants recruited from outside the university were given the option to either attend a usability test on the most convenient of the university campuses or in a location of their choosing. To mitigate the distress that participants may have experienced if they were unable to complete a task, it was clearly explained to participants in the PIS, during recruitment and during usability tests that they would not be penalised if they were unable to complete tasks. Participants were reassured that the usability tests were about the chatbot, *WeightMentor*, rather than the user.

Personal data collected during usability tests were collected in compliance with GDPR 2018. Personal data collected by the chatbot, *WeightMentor*, were stored on the database for the purposes of the usability test and were deleted once the test was completed.

5.4 Results

5.4.1 Participant Demographics

A total of 50 participants took part in usability tests. There were more female participants (29, 58%) than male (21, 42%). Participant ages ranged from 18-25 years to over 50 years, and mode age range was 18-25 years (28 participants, 56%). Most participants ranked their technical ability with mobile devices as “high” (27 participants, 54%), and their computer literacy as 4 (22 participants, 44%). Mean technical ability was 4.3±0.80, the median and mode were 5. Mean computer literacy was 4.1±0.8, and the median and mode were both 4. More participants had never used apps or devices for weight loss and

weight loss maintenance (30 participants, 60%) than had previously used them (20 participants, 40%). Participants who had previously used apps for weight and weight loss maintenance reported using a variety of apps and devices, the most reported being My Fitness Pal (10 participants, 50%). Half of those who had used apps reported using more than one app or device. Participant demographics are summarised in **Tables 5.8 to 5.10.**

Table 5.8 Usability Testing Participant Demographics (Gender, Age Range, Occupation, First Language(n=50))

Demographics		n	%
Gender	Female	29	58%
	Male	21	42%
Age Range	18-25	28	56%
	26-30	7	14%
	31-40	3	6%
	41-45	3	6%
	46-50	3	6%
	Over 50	3	6%
	Mode	18-25	
Occupation	Student	25	50%
	PhD Researcher	12	24%
	Non-Student	13	26%
First Language	English	43	86%
	Non-English	7	14%

Table 5.9 Usability Testing Participant Demographics (Technical Ability and Computer Literacy (n=50))

Demographics		n	%
Technical Ability	1 Low	0	0
	2	0	0
	3	10	20%
	4	13	26%
	5 High	27	54%
	Mean	4.3±0.8	
	Median	5	
	Mode	5	
	1 Novice	0	0%

Demographics		n	%
Computer Literacy	2	0	0%
	3	12	24%
	4	22	44%
	5 Expert	16	32%
Computer Literacy (Continued)	Mean	4.1±0.8	
	Median	4	
	Mode	4	

Table 5.10 Usability Testing Participant Demographics (Previous Use of Weight Loss/Maintenance Software & Apps/Devices Used (n=50))

Demographics		n	%
Previously used weight loss/maintenance software	Yes	20	40%
	No	30	60%
Apps/Devices Used	FitBit	3	15%
	Health App	1	5%
	Lose It!	1	5%
	Map My Run	2	10%
	Mealtime	1	5%
	Moves	1	5%
	My Fitness Pal	10	50%
	NHS Apps	1	5%
	Noom	2	10%
	Slimming World	2	10%
	Unspecified	7	35%
	Weight Watchers	1	5%
	Most Common	**My Fitness Pal**	
	Multiple Apps	**10**	50%

5.4.2 System Usability Scale (SUS) Results

5.4.2.1 SUS Scores

Results from the System Usability Scale were collated, and SUS scores for each participant were calculated as detailed in the Methods section above. SUS Scores for individual participants are listed in **Appendix 24**. A box plot of SUS scores is shown in **Figure 5.2**.

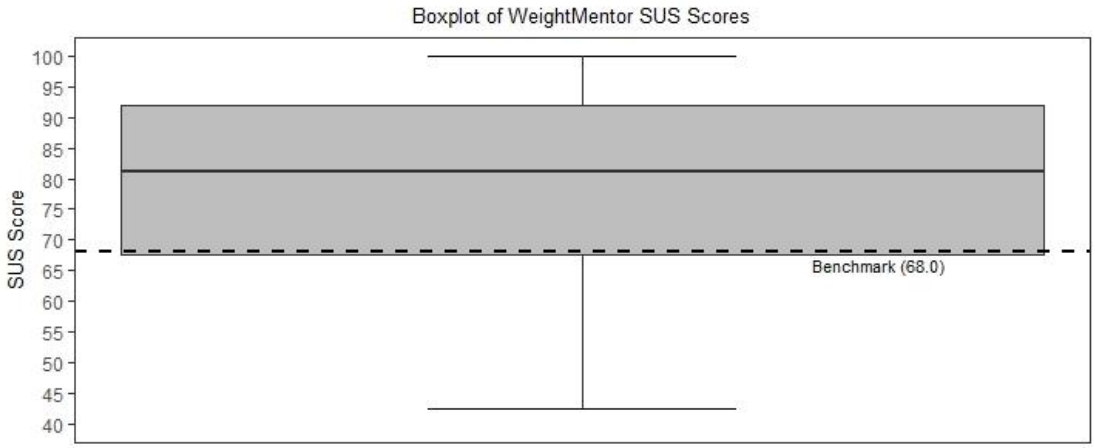


Figure 5.2 Boxplot of *WeightMentor* SUS Scores (Benchmark shown at 68.0)

The highest SUS Score (out of 100.0) was 100.0. The lowest was 42.5. The mean SUS score was 78.0±16.2. The median SUS Score was 81.3. Based on the different scales for interpreting SUS scores (discussed above), a score of 78.0 falls in the 80 - 84th percentile. It is equivalent to grade B+, "Good" on the adjective scale, "Acceptable" on the acceptability scale, and NPS class "Promoter".

5.4.2.2 Participant SUS Question Responses

Mean scores per question are shown in **Table 5.11**. The lowest scoring positive questions were Q1 (mean 3.3±1.0): "I think I would like to use this system frequently" and Q5 (3.8±0.9): "I found the various functions in the system were well integrated". All other positive questions scored highly (mean > 4.0). The highest scoring positive questions were Q3 (mean 4.5±0.7): "I thought the system was easy to use", and Q7 (mean 4.4±0.9): "I would imagine that most people would learn to use this system quickly". Four of the five negative question scores were below 2.0 (mean between 1.5±1.0 and 1.9±1.1). The lowest scoring negative questions were Q4 (mean 1.5±1.0): "I think I would need the support of a technical person to be able to use this system" and Q10 (mean 1.6±1.0): "I

needed to learn a lot of things before I could get going with this system". Highest scoring negative questions were Q8 (mean 2.1±1.1): "I found the system very cumbersome to use", and Q6 (mean 1.9±1.1): "I thought there was too much inconsistency in this system". Bar and box plots of SUS question responses are shown in **Figure 5.3**.

Table 5.11: Mean *WeightMentor* SUS Question Scores

Question	Aspect	Mean
1	Positive	3.3±1.0
2	Negative	1.7±0.9
3	Positive	4.5±0.7
4	Negative	1.5±1.0
5	Positive	3.8±0.9
6	Negative	1.9±1.1
7	Positive	4.4±0.9
8	Negative	2.1±1.1
9	Positive	4.1±1.2
10	Negative	1.6±1.0

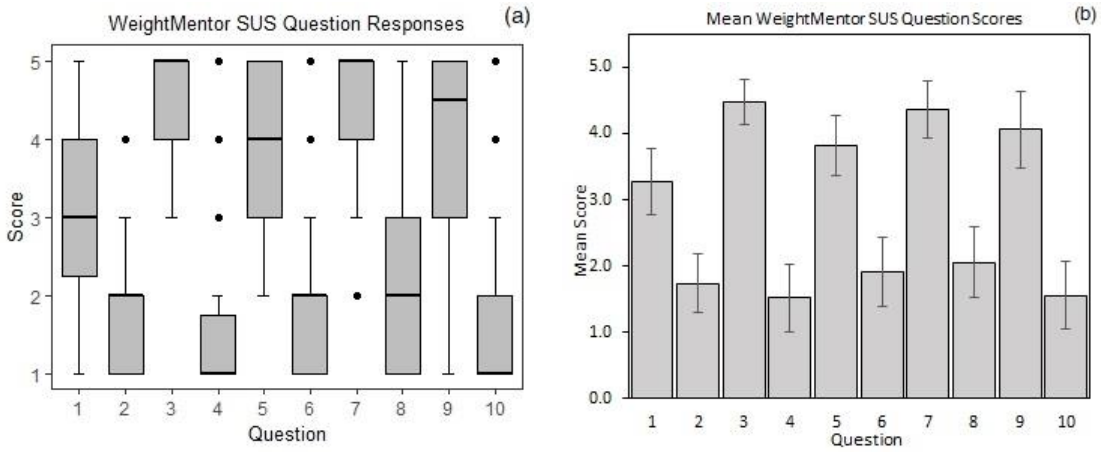


Figure 5.3 *WeightMentor* SUS Question Responses
(a) SUS question scores (b) Mean SUS question scores. Black dots represent outliers. Error bars represent SD

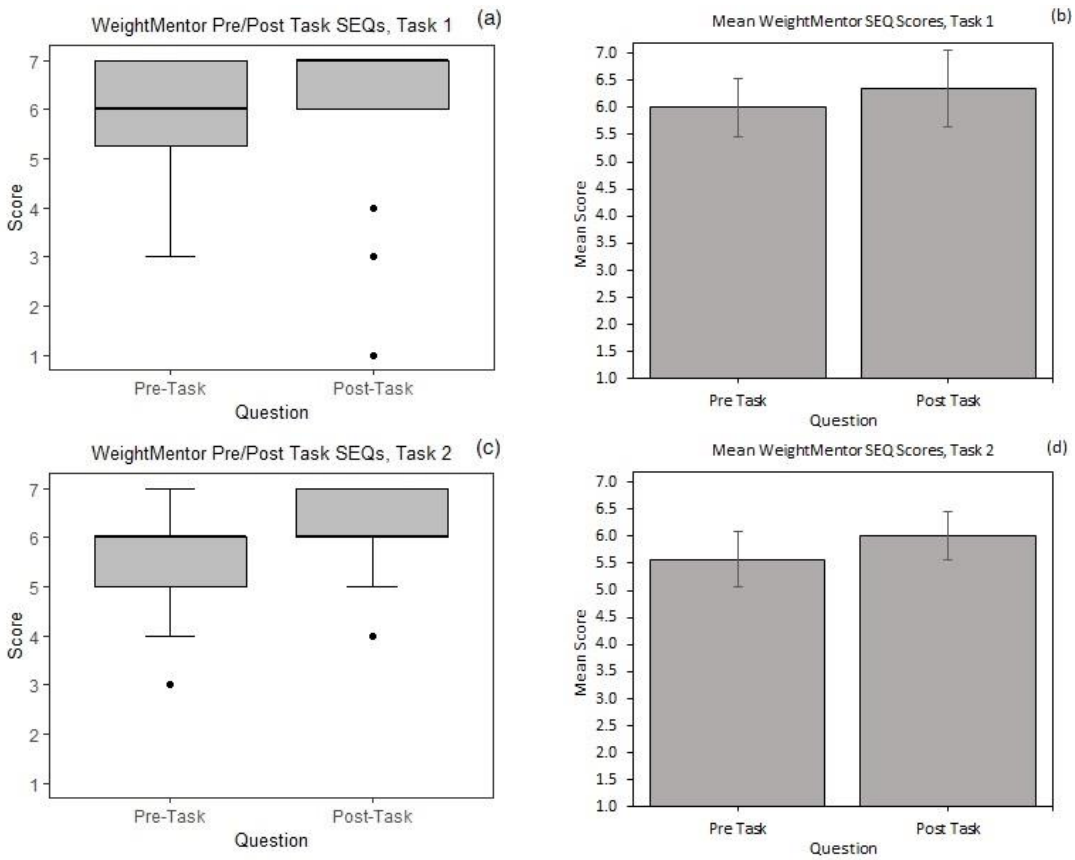
5.4.3 Single Ease Question Responses

The mean pre- and post-task scores, standard deviations and p-values are shown in **Table 5.12**, along with the Spearman correlation coefficients and associated p-values. Spearman correlation was used for SUS questions because these use a 5-point Likert scale, and Spearman correlation is better suited to this type of scale. The highest-scoring task pre-completion was task 1 (mean 6.0±1.1), and the lowest was task 2 (mean 5.6±1.0).

The highest-scoring task, post-completion was task 1 (mean 6.5±1.1), and the lowest was task 2 (mean 6.0±0.9). Differences between scores were significant ($p < 0.05$) for all tasks. Correlation between pre-and post-tasks was almost zero for tasks 1 and 2, weak for task 3 (< 0.5) and moderate (> 0.5 and < 0.7) for task 4. Correlation was not significant ($p > 0.05$) for tasks 1 &2 but was significant for tasks 3 & 4 ($p < 0.05$). Bar and box plots of pre- and post-task SEQ scores are shown in **Figure 5.4**.

Table 5.12: *WeightMentor* Pre- and Post-Task Means (with Standard Deviations) and p-values

Task	Mean score		Wilcox	Spearman correlation	
	Pre-task	Post-task	p-value	rho	p-value
1	6.0±1.1	6.5±1.1	0.002	-0.03	0.85
2	5.6±1.0	6.0±0.9	0.01	0.08	0.59
3	5.7±1.0	6.2±0.9	0.02	0.32	0.02
4	5.7±1.1	6.4±1.2	0.0002	0.61	0.000003



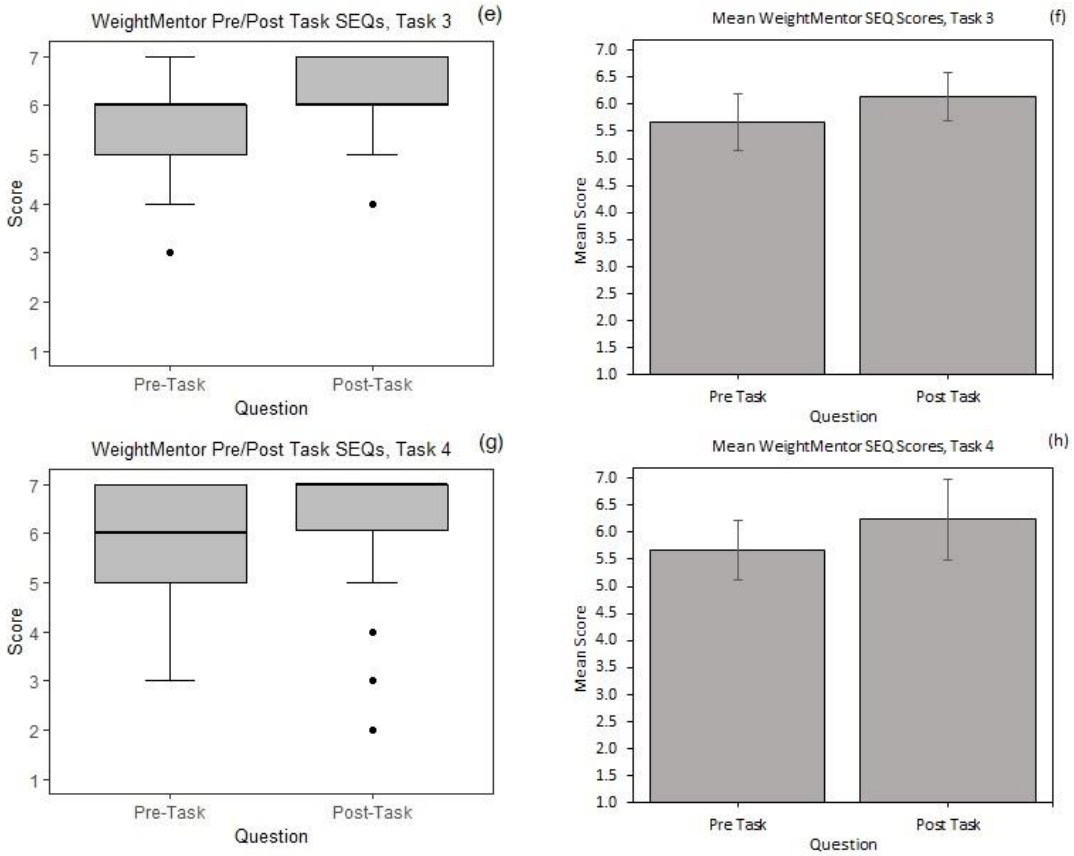


Figure 5.4: *WeightMentor* Pre- and post-task scores
(a) Task 1 box plot (b) Task 1 bar plot (c) Task 2 box plot (d) Task 2 bar plot (e) Task 3 box plot (f) Task 3 bar plot (g) Task 4 box plot (h) Task 4 bar plot. Black dots represent outliers. Error bars represent SD.

5.4.4 User Experience Questionnaire (UEQ) Responses

Results from the User Experience Questionnaire (UEQ) were collated and analysed as discussed in the Methods section above. Mean scores for each scale of the UEQ are presented in **Table 5.13** and graphically in **Figure 5.5**. Scores for Attractiveness, Hedonic and Pragmatic Quality are presented in **Table 5.14** and graphically in **Figure 5.5**.

Table 5.13: *WeightMentor* Mean UEQ Scores

Scale	Mean score
Attractiveness	1.6±0.9
Perspicuity	2.0±1.2
Efficiency	1.5±1.1
Dependability	1.3±0.9
Stimulation	1.2±1.1
Novelty	1.0±1.1

Table 5.14: *WeightMentor* Attractiveness, Pragmatic Quality and Hedonic Quality Scores

Category	Value
Attractiveness	1.8
Pragmatic Quality	1.9
Hedonic Quality	1.5

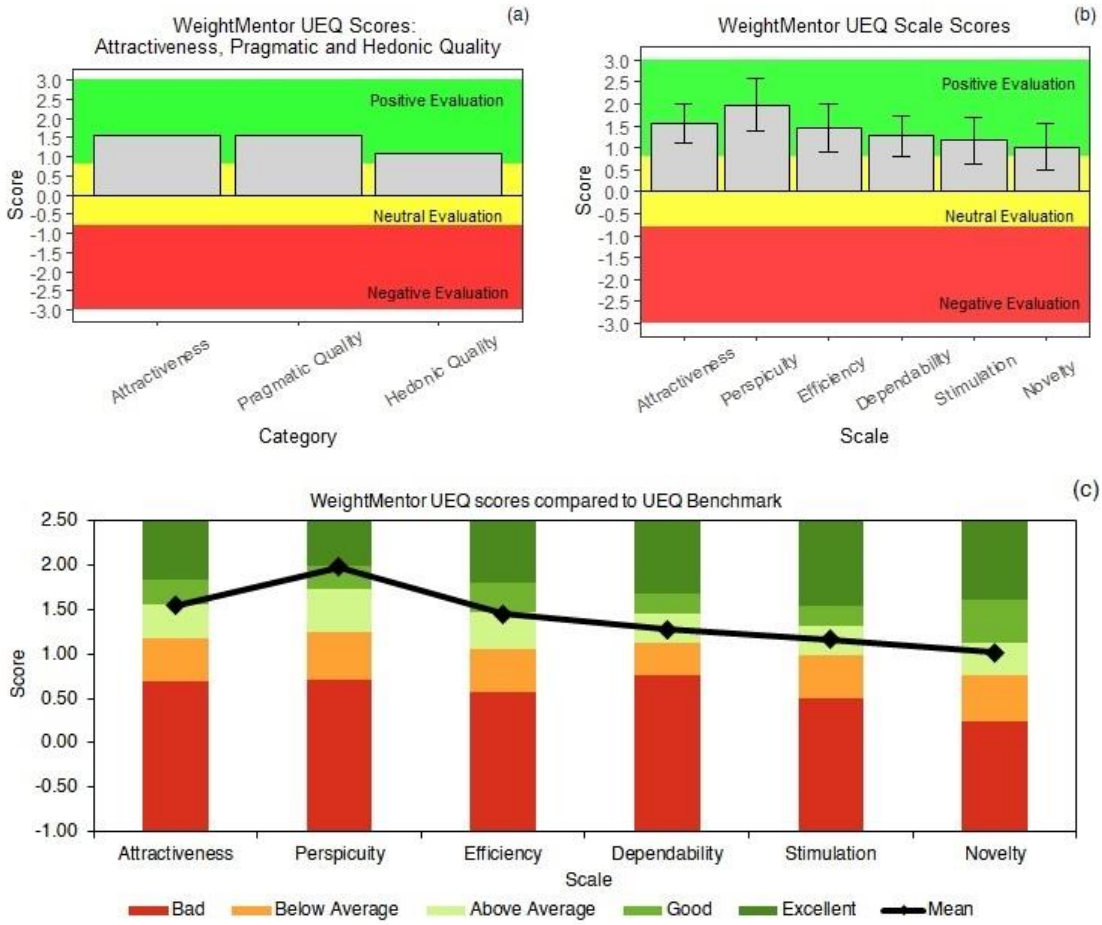


Figure 5.5: *WeightMentor* UEQ Category and Scale Scores
(a) UEQ category scores. (b) UEQ scale scores. Error bars represent SD. (c) WeightMentor UEQ scores compared to benchmark.

From **Figure 5.5** it can be determined that user opinions of WeightMentor were overwhelmingly positive (greater than mean +0.8). **Figure 5.5(c)** shows how *WeightMentor's* UEQ scores compare to the benchmark, and it can be seen that WeightMentor may be classed as "Good" for Attractiveness (mean 1.6±0.9), "Excellent" for Perspicuity (mean 2.0±1.2), and above average for Efficiency (mean 1.5±1.1) Dependability (mean 1.3±0.9), Stimulation (mean 1.2±1.1) and Novelty (mean 1.0±1.1).

5.4.5 Chatbot Usability Questionnaire (CUQ)

5.4.5.1 CUQ Scores

Results from the CUQ were collated, and scores for each participant were calculated as detailed in the Methods section above. Scores for all participants are recorded in **Appendix 25**. A box plot of scores is shown in **Figure 5.6**. The highest score (out of 100.0) was 100.0. The lowest was 39.1. The mean score was 72.2±13.9. The median score was 73.4. Correlation between CUQ question responses is shown in **Figure 5.7**.

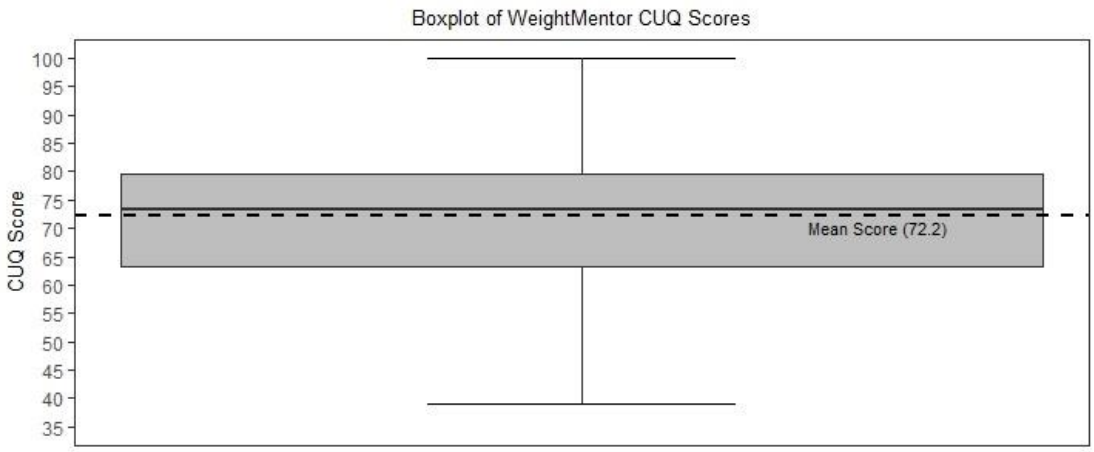


Figure 5.6: Boxplot of *WeightMentor* CUQ Scores
(mean score shown at 72.2)

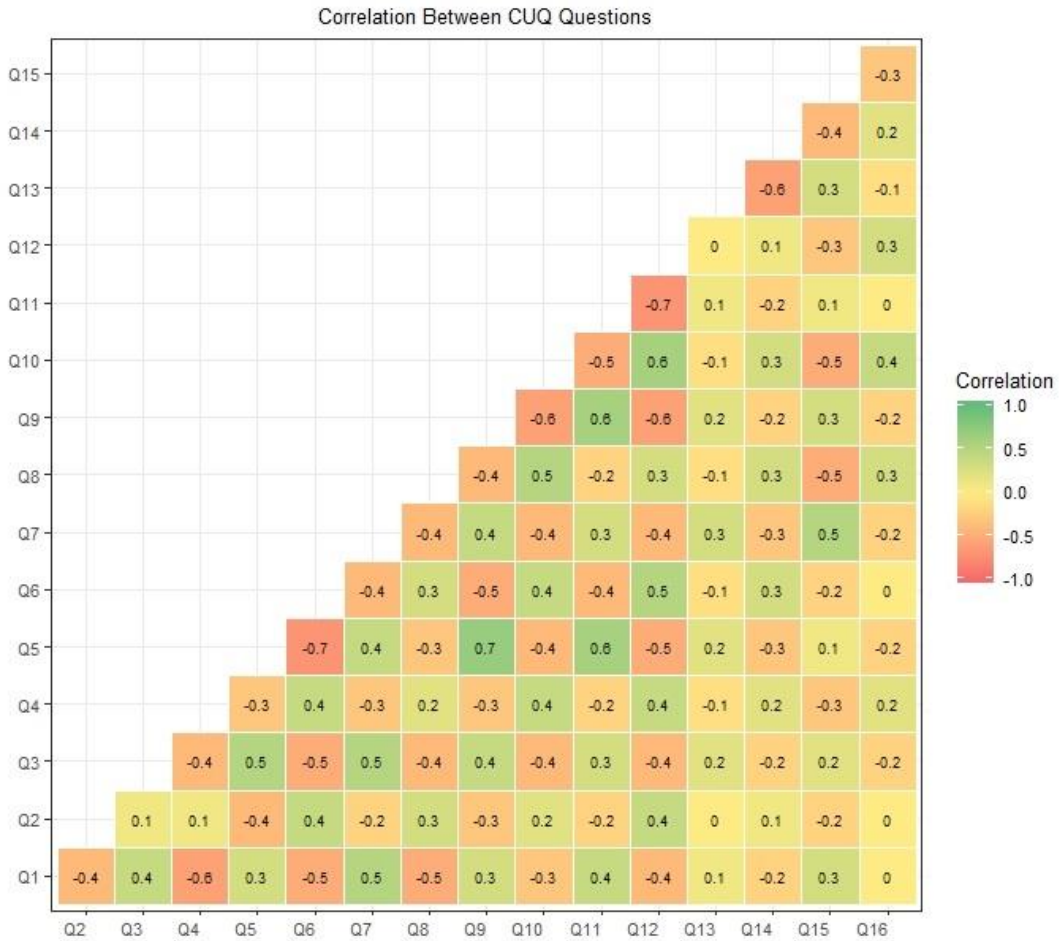


Figure 5.7: Correlation Matrix: *WeightMentor* CUQ Question scores

In **Figure 5.7**, correlation varies widely between CUQ questions. The strongest correlations were between Q5 and Q9, Q5 and Q11, Q9 and Q11, and Q10 and Q12. The weakest correlations were between Q1 and Q4, Q5 and Q6, Q9 and Q10, Q9 and Q12, and Q11 and Q12.

5.4.5.2 Participant CUQ Question Responses

Participant responses to individual CUQ questions are shown in **Figure 5.8**. Mean scores per question are shown in **Table 5.15**. The lowest scoring positive question was Q13 (mean 3.2±1.0): "The chatbot coped well with any errors or mistakes". Q1 ("The chatbot's personality was realistic and engaging", mean 3.7±1.0), Q5 ("The chatbot explained its scope and purpose well", mean 3.7±1.1), Q9 ("The chatbot understood me well", mean 3.8±0.8) and Q11 ("Chatbot responses were useful, appropriate and informative", mean 3.9±0.9) all scored low. The highest scoring positive question was Q15 ("The chatbot was very easy to use", mean 4.5±0.6). Q3 ("The chatbot was welcoming during initial setup", mean 4.2±0.8) and Q7 ("The chatbot was easy to navigate", mean 4.2±1.0) also scored highly.

Negative question scores were between 1.4±0.8 and 2.9±1.1. The highest scoring negative question was Q2 ("The chatbot seemed too robotic", mean 2.6±0.7). Q14 ("The chatbot seemed unable to handle any errors", mean 2.5±0.9), Q16 ("The chatbot was very complex", mean 2.2±1.1) and Q6 ("The chatbot gave no indication as to its purpose", mean 2.2±1.2) all scored highly. The lowest scoring negative question was Q4 ("The chatbot seemed very unfriendly", mean 1.4±0.8). Q8 ("It would be easy to get confused when using the chatbot", mean 2.0±1.0), Q10 ("The chatbot failed to recognise a lot of my inputs", mean 1.9±1.0) and Q12 ("Chatbot responses were not relevant", mean 2.0±1.0) also scored low.

Table 5.15: Mean *WeightMentor* CUQ Question Scores, per question

Question	Aspect	Mean
1	Positive	3.7±1.0
2	Negative	2.9±1.1
3	Positive	4.2±0.8
4	Negative	1.4±0.8
5	Positive	3.7±1.1
6	Negative	2.2±1.2
7	Positive	4.2±1.0
8	Negative	2.0±1.0
9	Positive	3.8±0.8
10	Negative	1.9±1.0
11	Positive	3.9±0.9
12	Negative	2.0±1.0
13	Positive	3.2±1.0
14	Negative	2.5±0.9
15	Positive	4.5±0.6
16	Negative	2.2±1.1

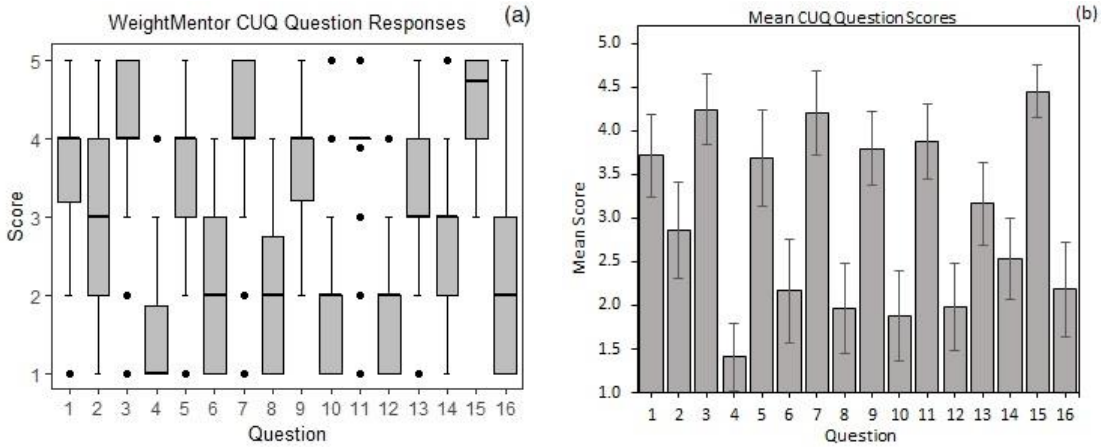


Figure 5.8: *WeightMentor* CUQ Question Scores
(a) Boxplot of scores per question (b) Mean scores per question. Black dots represent outliers. Error bars represent SD

5.4.6 Correlation Analysis of Usability Questionnaires

It is not ordinarily possible to measure correlation of the UEQ with other single-score usability metrics, as the UEQ provides six scores for each of the scales. Thus, in order to determine correlation between the UEQ and SUS and CUQ, it was necessary to manually calculate a single score for this questionnaire. This was accomplished by calculating the mean UEQ score across all scales, per participant, which could then be correlated with CUQ and SUS scores. Scatter plot comparisons, with p-values and correlation coefficients, are shown in **Figure 5.9**. The correlation is strong across all questionnaires but is strongest between SUS and UEQ. All correlations are statistically significant as p-values are below 0.05 (5%). Multiple regression was used to determine if CUQ score correlated with SUS and UEQ Mean. Results of this are shown in **Table 5.16**.

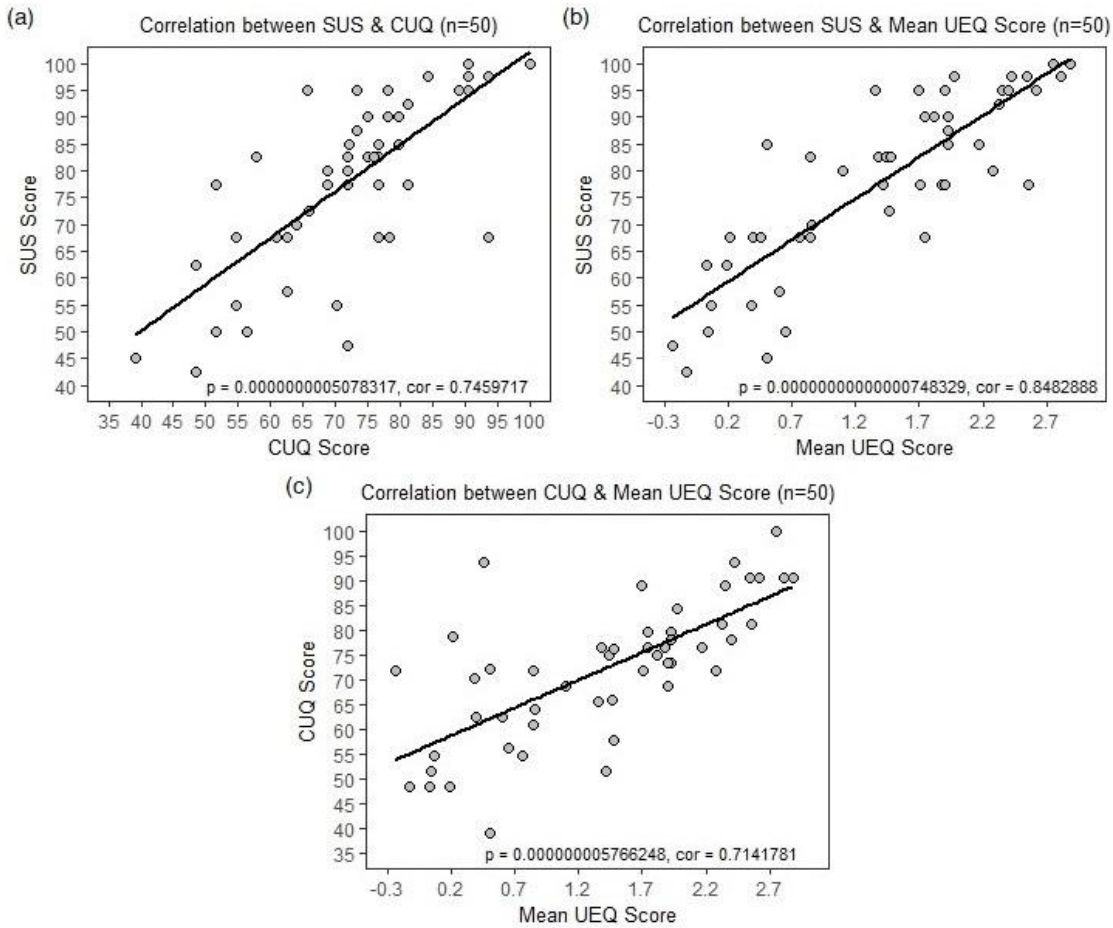


Figure 5.9: Correlation between *WeightMentor* usability questionnaires
(a) SUS and CUQ (b) SUS and Mean UEQ Score (c) CUQ and Mean UEQ Score

Table 5.16: *WeightMentor* Multiple Regression Results
Independent variables are SUS and UEQ Mean, and dependent variable is CUQ

Coefficients	Estimate	Std. Error	t-value	p-value
Intercept	32.15	8.99	3.58	0.0008
SUS	0.43	0.15	2.80	0.007
UEQ	4.57	2.81	1.63	0.11
Residual Standard Error	9.20 on 47 degrees of freedom			
Multiple R-squared	0.58	*Adjusted R-Squared*		0.56
F-statistic	32.46 on 2 and 47 DF		*p-value*	0.000000001

5.4.7 Task Completion Times

Mean task completion times for each task (with task repetitions and p-values) are shown in **Table 5.17** along with benchmark times. Mean task completion times overall are shown in **Table 5.18**. Based on the tables, in general participants completed tasks faster each time they repeated the same one, however time differences were not significant ($p < 0.05$) after the first repetition of task 2. Tasks 1, 2.2, 2.4 and all repetitions of task 3 were

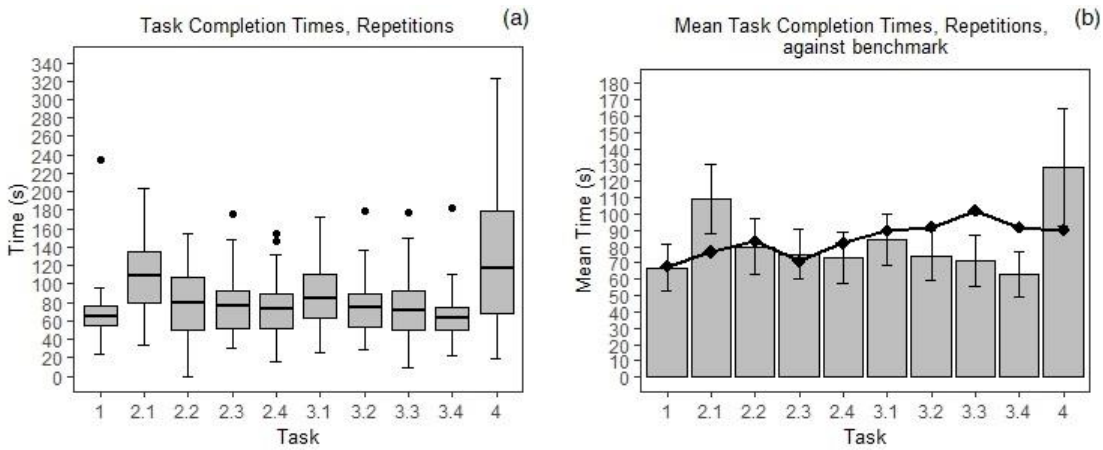
completed faster than benchmark times. Task completion times are presented graphically in **Figure 5.10**. The two slowest tasks compared to benchmark were task 2.1 and task 4. Overall, tasks 2 and 4 were slower than benchmark.

Table 5.17: Mean *WeightMentor* Task Completion Times (With Repetitions), Per Task

Task	Mean	p-value	Benchmark	Difference	% Difference
1	66.9±29.0	---	68.0	-1.1	-1.7%
2.1	109.4±42.1	---	77.0	32.4	29.6%
2.2	80.0±34.1	0.0008	83.0	-3.0	-3.7%
2.3	75.4±30.1	0.35	71.0	4.4	6.0%
2.4	73.5±31.3	0.59	82.0	-8.5	-11.6%
3.1	84.2±32.1	---	90.0	-5.8	-6.8%
3.2	74.4±30.0	0.095	92.0	-17.6	-23.7%
3.3	71.0±31.4	0.40	102.0	-31.1	-43.8%
3.4	63.0±28.0	0.09	92.0	-29.0	-46.1%
4	128.5±72.6	---	90.0	38.5	29.9%

Table 5.18: Mean *WeightMentor* Task Completion Times (Overall), Per Task

Task	Mean	Benchmark	Difference	% Difference
1	67.0±29.0	68.0	-1.1	-1.7%
2	84.6±37.4	78.3	23.8	23.3%
3	73.1±31.1	94.0	-20.9	-28.5%
4	128.5±72.6	90.0	38.5	29.9%



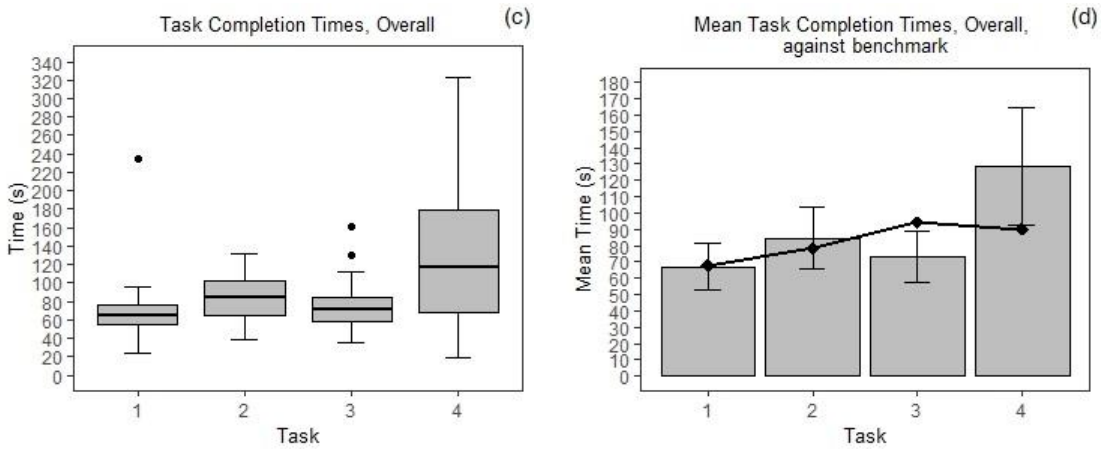


Figure 5.10: *WeightMentor* Task Completion Times
(a) Box plot of Task completion times (with repetitions), per task (b) Mean Task completion times (with repetitions), per task, compared to benchmark. (c) Task completion times (overall), per task (d) Mean task completion times (overall), per task, compared to benchmark. Black dots represent outliers. Error bars represent SD

5.4.8 Usability Issues

In total, fifty-nine issues were identified by participants during the usability tests. These were identified based on feedback submitted during the post-test survey, and from concurrent think-aloud audio data. Line charts of best-case, worst-case and chronological order scenarios are shown in **Figure 5.11**. The best-case graph plateaus at 26 participants (the optimum number), the chronological graph plateaus at 30 to 50 participants (mean 40) and the worst-case graph never plateaus (50 participants). The mean optimum number of participants is 39.

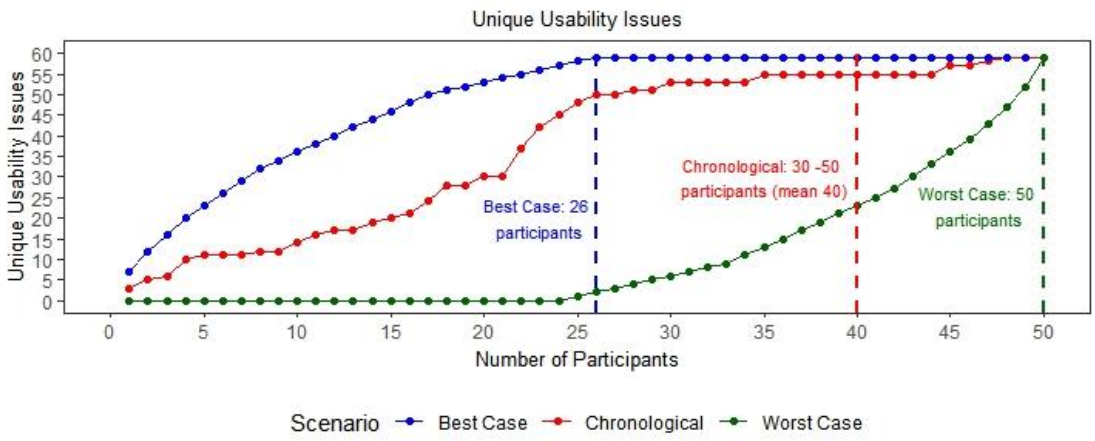


Figure 5.11: Unique *WeightMentor* usability issues identified

5.5 Discussion

WeightMentor's SUS score was 78.0±16.2, which is well above the average SUS score of 68.0. Given that chatbots are theoretically more intuitive and user friendly than

conventional mouse-and-pointer driven systems, it is perhaps expected that *WeightMentor's* SUS score should be well above benchmark. However, it is worth remembering that SUS is designed to assess general usability of computer systems, and it is unknown how many of the systems used to set the SUS benchmark are chatbots. Thus, a SUS score may not accurately reflect *WeightMentor's* usability in comparison with other chatbots. As discussed in the introduction, Baki Kocaballi et al. (2018) suggested that multiple questionnaires may be more appropriate for assessing conversational user interfaces to give a more comprehensive evaluation of usability. WeightMentor UEQ scores were all above +0.8, suggesting that participants reacted positively to the user experience. The chatbot also scored highly against the benchmark. However, as with SUS, it is unknown how many of the systems used in the benchmark were chatbots, and thus it is impossible to determine how WeightMentor compares to other chatbots. The mean CUQ score of 72.16 is slightly lower than the mean SUS score, however CUQ results cannot be treated as reliable because the questionnaire had not yet been validated at the time of testing. These positive scores improve *WeightMentor's* potential as a weight loss maintenance chatbot as they suggest that the chatbot is highly usable. Nielsen (2012) suggests that "usability [of a system] is a necessary condition of survival" and that users may abandon systems which are unusable. If *WeightMentor* is intended to support six of the ten *Processes of Behaviour Change* associated with the maintenance phase of the Transtheoretical Model of Health Behaviour Change (as discussed in **section 2.3.1**) then high usability is of utmost importance to reduce the risk of abandonment by users.

Positive SUS questions (i.e. odd-numbered questions) all scored above 3.0, suggesting that, in general, participants were quite positive about the chatbot usability. The two lowest scoring questions were Q1 and Q5. Low scores here suggest that some users did not feel that WeightMentor was something they would use frequently (Q1), and some users did not find that the functions of the chatbot were well integrated (Q5). These were the only two questions to score lower than 4.0. The highest positive questions were Q3 and Q7, indicating that, in general, participants felt WeightMentor was easy to use, and that most people could learn to use the chatbot quite quickly. This is consistent with current wisdom, which suggests that chatbots are highly user friendly and easy to use. In general, negative question scores were low (below 2.0). The low scores for questions Q4 and Q10 are consistent with high scores for Q3 and Q7, which relate to ease of learning and use. However, Q6 and Q8 both scored highly (mean 1.9±1.1 and 2.1±1.1), and the

implication of these higher scores is that for some participants, WeightMentor was too cumbersome and not consistent.

Positive CUQ responses were all above 3.0. Low positive scores suggest that participants were not satisfied with error handling (Q13), personality (Q1), explanation of scope and purpose (Q5), understanding (Q9) and usefulness of responses (Q11). However, ease of use (Q15) and navigation (Q7) scored highly, again consistent with the simple and user-friendly nature of conversation-driven systems. A high score for Q3 suggests that the chatbot was perceived as welcoming. Questionnaire scores are consistent with work by Fitzpatrick et al. (2017), which suggested that the *WoeBot* mental health chatbot (discussed in chapter 2) was highly acceptable and popular with users, and also with findings from a study by Saenz et al. (2017), which used SUS to evaluate the usability of three chatbots, the most popular of which scored 81.9 out of 100, higher than both the SUS benchmark and *WeightMentor*. A study by Budiu, 2018 suggested that contemporary technology users are not generally hostile towards chatbots, and that user attitudes "range from neutral to slightly positive".

Post-task mean SEQ scores were higher than pre-task SEQ means for all tasks. This suggests that participants' expectations of the chatbot were exceeded during the usability tests. Although pre-task scores were higher for task 1 than for all other tasks, there was a greater difference between pre- and post-task scores for tasks 2 to 4 than for task 1. Differences were significant for all tasks. Sauro, 2012(b) suggests that the average SEQ score is between 5.3 and 5.5. Overall, all *WeightMentor* SEQ scores were well above this average, and the implication of this is that participants perceive chatbot-based tasks to be easier than conventional GUI-based system tasks. Sauro's work is based on 400 tasks and 10,000 users (Sauro, 2012b), however it is unknown how many of these tasks were chatbot-based. Correlation between pre-and post-task questions was almost zero for the first two tasks, weak (0.32) for task 3 and moderate (> 0.61) for task 4. Correlation was significant for tasks 3 and 4. One possible explanation for this is that for the first two tasks, participants may not have been sure how difficult each task would be (particularly if they had never used a chatbot) thus would have been unsure how to rate the ease of each task, however by tasks 3 and 4 participants were more familiar with the operation of the chatbot and thus were able to more accurately gauge task difficulty.

For repetitive tasks (i.e. tasks 2 and 3), it was anticipated that task completion times would decrease with each repetition. However, it was observed that while task completion times did decrease in this way, there was no significant difference between the second and third and fourth repetitions of task 2, or between any of the repetitions of task 3. The implication of this is that a new chatbot user will reach optimum performance after just one attempt at a task. Additionally, since task 3 was very similar in procedure to task 2 (self-reporting), it makes sense that there should be no significant difference in completion times for this task, as participants would already have been familiar with self-reporting procedures, having completed task 2. The implication of these findings is that users become proficient in chatbots very quickly.

Mean task completion times were compared to benchmark times, which were established by recording completion times per task for the PhD researcher. Although in general the mean completion times per task were above the benchmark, interestingly, mean task completion times for tasks 1 and 3 were below the benchmark. One possible explanation for this is that during benchmarking the PhD researcher completed task 3 in full, supplying extra information relating to food consumption, whereas during tests, only 50% of participants (n=25) did so. Additionally, at the time of benchmarking the PhD researcher was already very familiar with each of the four tasks, thus knew what each one involved. Task 1 required users to create a user profile, and as many users of technology are already familiar with user profile creation for websites such as Amazon or Google, it may be argued that this task was nothing new for participants. Additionally, task 3 was very similar in procedure to task 2, and participants would already have been familiar with the procedure after four repetitions of task 2. Conversely, tasks 2 and 4 were potentially new activities for users and so may have taken longer to complete as users worked carefully through each task and considered what they needed to do.

Correlation analysis between surveys showed that correlation was strongest between the UEQ and CUQ. This suggests that UEQ and CUQ may be measuring similar aspects of usability. Multiple regression results suggest that 58% of the variance in CUQ score is explained by SUS and UEQ, although given that the p-values for each of these is greater than 0.05, neither questionnaire is significant in predicting CUQ scores. The implication is that SUS and UEQ are, as predicted, perhaps less effective at measuring the usability of chatbots.

Analysis of usability issues shows that in the best-case scenario the line graph plateaus at 26 users, representing the optimum number of users for identifying all the unique usability issues. In the worst-case scenario, the graph does not plateau, and, instead, continues upwards until all usability issues have been identified at 50 users. In chronological order, the graph begins to plateau at 30 users, but continues to rise until the 50th user. The mean optimum number of users for chronological order is 40. Based on the three-line graphs, the optimum number of users required to find most of the usability issues is 39. Research by Nielsen & Landauer (1993) determined that no more than 5 - 8 users were required to identify 80% of unique usability issues. This research involved conventional software systems however, and the implication is that it is much more difficult to identify usability issues in chatbots with a small number of users.

5.6 Future Work

Based on the *WeightMentor* usability testing results, it is proposed to determine how to improve *WeightMentor* to increase the mean scores of positive SUS questions 1 and 5. A low score for question 1 suggests participants did not feel they would like to use *WeightMentor* frequently. This may be linked with the low score for question 5, which suggests that participants did not find that *WeightMentor's* main functions were well integrated. Further research and discussion with participants would help to identify ways to improve functionality and potentially increase the desirability of the chatbot.

It is also proposed to determine how to address low mean scores for CUQ question 1 (personality), question 5 (explanation of scope and purpose), question 9 (understanding), question 11 (usefulness of responses) and question 13 (error handling). Based on Martín et al. 2017 these are important aspects of chatbot usability and *WeightMentor* clearly needs to be refined in order to improve these aspects of the user experience.

5.7 Conclusion

Usability testing is an important part of software development, however chatbots may require different methods than those used for conventional systems. Questionnaire scores for the WeightMentor chatbot were favourable, and correlation was found to be stronger between CUQ and UEQ, however multiple regression results showed that CUQ scores are not completely explained by UEQ and SUS, thus these tools may not be completely effective for assessing chatbot usability. A validated CUQ would potentially be more

effective as a usability tool. It was observed that 39 users were required to identify most of the usability issues in the chatbot. This challenges previous research, which suggested that 5 to 8 users were the optimum number for identifying usability issues.

Chapter 6: Validating the Chatbot Usability Questionnaire

6.1 Introduction

As discussed in chapter 5, there are currently numerous tools for measuring system usability. The vast majority of these are designed for assessing traditional "mouse-and-pointer" systems and thus may not be optimal for chatbot usability measurement. The Chatbot Usability Questionnaire (CUQ) was developed and trialled during the *WeightMentor* usability tests, based on the ALMA chatbot test tool (Martín et al. 2017), which assesses seven aspects of the chatbot usability experience: Personality, Onboarding, Navigation, Understanding, Responses, Error Management and Intelligence. The sixteen items in the questionnaire will in theory provide a more comprehensive analysis of chatbot usability, however as with any new measurement instrument, the CUQ must be validated before it may be considered useful for assessing chatbot usability and user experience.

6.1.1 Reliability

Spector (1992) suggests that a good scale will demonstrate both reliability and validity. Scale reliability is a measure of consistency, and there are three main types. *Test-retest reliability* relates to *consistency of a scale over time*. If usability test participants ("raters") complete the same questionnaire for the same system, each participant's score should be about the same for each repetition. This is also known as *intra-rater reliability*. *Internal consistency* is concerned with correlation of the items within the questionnaire, and those which measure the same construct should show strong positive correlation (Spector, 1992). Ideally a scale will demonstrate both types of reliability, but even one of these will be considered acceptable. The third type, Inter-rater reliability, relates to consistency of scores from *different participants* (Spector, 1992). If 100 usability test participants complete a questionnaire and it is observed that individual scores are very close, the results may be said to possess *inter-rater reliability*. In this study, only intra-rater reliability was measured.

6.1.2 Validity

Validity is the extent to which a scale measures what it is designed to measure. There are several types of validity, the most primitive being *face validity*. This is the idea that a scale is valid simply because to experts and respondents (i.e. untrained observers) it looks like it *might* measure what it is intended to measure (Holden, 2010). Although face validity results in greater acceptance by respondents, it does not offer conclusive guarantees of technical validity, thus its usefulness is limited (Holden, 2010). Technical validity is related to the extent to which a scale measures what it is supposed to measure. Three of the most measured types of technical validity are *content validity*, *criterion-related validity* and *construct validity* (DeVellis, 1991).

6.1.2.3 Content validity

Content validity is the degree to which the questionnaire items represent the concept that is being measured, ensuring that questions sufficiently cover all aspects of the concept, without asking questions that are irrelevant (Parahoo, 2014). In the context of a usability questionnaire, for example, the questionnaire will show content validity if the questions within it adequately assess all aspects of usability. Content validity is similar to face validity in that it is subjective, however it is based on the opinions of suitably qualified "experts" in the concept that is being measured, rather than opinions of untrained respondents as is the case with face validity. Experts can be asked to rate the relevance of each questionnaire item, and the level of agreement between each individual expert may then be measured (Parahoo, 2014).

6.1.2.4 Criterion-related validity

Criterion-related validity, or just *criterion validity*, is a measure of the correlation between the new instrument (i.e. the one which is being validated) and an existing instrument known to be valid (DePoy & Gitlin, 2016). Ideally the existing instrument should be a "gold standard" (Bellamy, 2015). There are two main types of criterion-related validity. *Concurrent validity* may be demonstrated by simultaneous comparison of the two instruments (Lin & Yao, 2014a), for example during a usability test when participants complete both an accepted standard usability questionnaire (e.g. System Usability Scale (SUS)) and a newly developed questionnaire. An instrument is said to have *predictive validity* if the scores can predict the outcome of future performance using a different

measure later (Lin & Yao, 2014b). For example, during a usability test the scores may indicate those individuals most likely to use the chatbot to maintain weight loss and then their actual use of the chatbot would be measured at a later time and the scores compared to determine whether the former indication was correct (i.e. valid).

6.1.2.5 Construct validity

In this type of validity, the term "construct" does not refer to construction of an experiment or assessment, but rather to theoretical concepts or factors, for example beliefs, perceptions and attitudes, which need to be measured but are intangible and thus it may be difficult to do so accurately without a validated tool. Construct validity may be assessed using techniques such as *Factor Analysis* (Preedy & Watson, 2009). Construct validity may be further divided into two types: convergent and discriminant validity. *Convergent validity* is demonstrated by a correlation between two constructs which are related to each other. In the context of a chatbot usability questionnaire, for example, two questions assessing the same aspect of chatbot usability (e.g. personality) correlation between responses to the two questions will be very high. *Discriminant validity* on the other hand, may be demonstrated by the absence of correlation between two constructs. In other words, if two questions from the usability questionnaire were selected which relate to different aspects of usability (e.g. personality vs intelligence), correlation between results for each question would be zero or close to zero.

Whilst this introduction provides a brief overview of survey or scale validation methods, this study attempts to only validate a small number of aspects of the CUQ. This includes measuring for construct validity and intra-rater reliability (or test-retest validity).

6.2 Aim

To validate the CUQ for measuring the usability of chatbots, for construct validity and intra-rater reliability.

6.3 Methods

6.3.1 Participants and Sample Size

Participants were adults (over 18 years). Although the *WeightMentor* chatbot is designed for weight loss maintenance, the CUQ is designed for evaluating the usability of any type of chatbot. Thus, it was not necessary for participants to have an interest in weight loss, or to be trying to lose or maintain weight. Selection criteria are summarised in **Table 6.1**. Participant demographics are discussed in section 4.1.

Table 6.1: CUQ Validation Study: Participant selection criteria

Inclusion criteria	Exclusion criteria
Adults over 18 years	Not providing consent
Reasonable understanding of the English language	
Basic IT/smartphone skills	
Willing to use a chatbot	

For the CUQ validation study, participants were selected using convenience sampling, which has previously been discussed in chapter 5. Since the only requirements for participation were to be aged 18+, have reasonable understanding of the English language, basic IT/smartphone skills and willingness to use a chatbot, it was determined to be most convenient to select participants from staff, students and other PhD Researchers within Ulster University, and also from individuals on the researcher's Facebook friends list. It was also determined that approximately 30 participants would be required for this research study, although it was anticipated that not all participants would be willing to participate in round 2 of the study (see section 3.4). Thus, the aim was to recruit 50 participants in order to ensure enough participants would complete both rounds of the study.

6.3.3 Recruitment

A recruitment email (see **Appendix 26**) was circulated to university staff and students via the university mailing list and to other PhD researchers and staff in the Faculty of Arts, Humanities and Social Sciences and the School of Computing by the Faculty and School Administrative Officers. A similarly worded invitation was posted to Facebook by the PhD researcher. Individuals who were interested in taking part visited the link to the study website and downloaded a copy of the Participant Information Sheet (PIS), which

is in **Appendix 27**. Individuals who were eligible to participate and were willing to do so signed the consent form (via Qualtrics) using the link on the study website. The consent form is in **Appendix 28**. Consenting participants completed a demographic questionnaire via Qualtrics, using the link on the study website. The demographic questionnaire is in **Appendix 29**.

6.3.4 Online Resources

For participant convenience, the study was designed to be conducted online. To facilitate this, the researcher designed a simple website that provided participants with easy access to the Participant Information Sheet (PIS) consent form, demographic questionnaire, CUQ and the three chatbots. The website and PIS were hosted on university webspace provided by the School of Computing, and the consent form and questionnaires were hosted on Qualtrics. Once participants had signed the consent form, they were assigned a randomly generated participant ID, which they would use when completing their demographic questionnaire and CUQs. This ensured that participant responses could be tracked without compromising anonymity. After signing the consent form, participants were directed to complete the demographic questionnaire, and once they had done so, were taken to a page where they could access the chatbots and the CUQ.

6.3.5 Chatbot Selection

A panel of experts evaluated three chatbots using their expertise and pre-determined criteria. The chatbots selected were *WoeBot*, a smartphone-based mental health chatbot, *Weight Loss Bot*, a Facebook Messenger-based chatbot for educating users about weight loss, and a very rudimentary simulation of a chatbot-based flight booking system, designed by the PhD researcher for the purposes of the research study, using a website called *Landbot.io*. Evaluation criteria were developed by the PhD researcher based on areas of assessment used by the ALMA chatbot test tool (Martín et al. 2017) and are shown in **Table 6.2**. Based on these criteria, the PhD researcher rated *WoeBot* as “good quality”, *Weight Loss Bot* as “average quality”, and *FlyBot* as “poor quality”. Each chatbot was then evaluated by the expert panel, made up of Dr Raymond Bond, Dr Patrick McAllister, KTP Associate and Emeritus Professor Michael McTear, all from the School of Computing. Each panel member used their expertise to independently rank the chatbots in order of usability. Then the panel members met as a committee to discuss and provide a consensus rating. None of the panel members knew how the three chatbots had

been categorised by the PhD researcher, but all agreed that *WoeBot* was “good quality”, *Weight Loss Bot* was “average quality” and *FlyBot* was “poor quality”, consistent with the PhD researcher’s original evaluation.

Table 6.2: Chatbot Classification Criteria

User Experience Aspect	Description	Chatbot Classification		
		Good	**Average**	**Poor**
Personality	• Does the chatbot seem friendly? • Is it robotic? • Does the personality fit with its role/purpose?	• Friendly and casual • Personality appropriate to purpose	• Personality is friendly but somewhat limited and robotic	• Personality is very limited, almost robotic • Personality not appropriate to purpose (e.g. inappropriate use of humour)
Onboarding	• Does the chatbot welcome the user and explain how they should use it? • Is the user able to "dive straight in?"	• Chatbot is very welcoming, greeting user by their first name (gets their name and uses it!) • Chatbot explains purpose and functions very well • User can start using chatbot straight away with very little prompting	• Greeting is limited to just a simple "Hello" (no first name use) • Explanation of purpose/function is limited • User needs some prompting before they can use the chatbot	• Chatbot greeting is non-existent • Chatbot does not explain purpose or function • User is unable to "dive straight in" and requires considerable help or instructions for first use
Understanding	• How well does the chatbot understand the user?	• Chatbot understands the user well, and copes well with different variations of responses	• Chatbot sometimes struggles with variations in user response	• Chatbot can only understand user if their responses exactly match expected inputs
Answering	• How does the chatbot answer the user? • Are its responses varied? • Are they relevant?	• Chatbot's answers are varied, interesting and relevant, with a good mix of humour (if appropriate), and multimedia (e.g. GIFs)	• Chatbot answers are relevant but a little dull • Limited use of multimedia or relevant humour	• No variability in answers • Answers are not relevant or are confusing • No humour (or inappropriate humour) or multimedia in responses
Navigation	• How quickly can the user get what they need from the chatbot? • Does the user know where they	• Chatbot is very easy to navigate • User always knows where they are, how they got there, and what they need to do	• Chatbot is navigable, but user sometimes gets lost or needs reminded about what they need to do	• Chatbot is too complex • User finds that they get lost easily, or is left wondering what to do

User Experience Aspect	Description	Chatbot Classification		
		Good	Average	Poor
	are and what they are doing?			
Error Management	• How does the chatbot respond to errors? • Does it become confused? • If it does, can it recover? • Does it ask the user to clarify what they meant?	• Chatbot handles and responds to all errors, notifying the user that something went wrong (e.g. "I didn't understand that") and trying to repair the error (e.g. by asking the user to clarify what they meant)	• Chatbot error handling is only very basic and offers limited options for recovery	• Poor error handling - chatbot either crashes when an error occurs (leaving the user wondering what has just happened) or generates a cryptic error message
Intelligence	• Is the chatbot intelligent? • Can it remember things? • Can it manage conversation context?	• Chatbot seems to be intelligent • Chatbot remembers the user's name and manages the context of the conversation (e.g. remembering information it has been given)	• Chatbot seems intelligent but only has limited memory capabilities	• Chatbot intelligence very limited • Chatbot is unable to remember user's name (or does not use it) • Chatbot forgets information the user has already given it

6.3.4 Questionnaire Validation

6.3.4.1 Round 1

Participants visited the study website, viewed the PIS and then signed the consent form online via Qualtrics. Participants then completed the demographic questionnaire and were taken to the web page with instructions for validating the chatbots. Participants were free to evaluate the three chatbots in any order they liked. Participants selected one of the three chatbots, visited its website and followed any setup instructions, then spent no more than 5 to 8 minutes evaluating the chatbot. When participants had finished using the chatbot they returned to the study website and clicked the link to the CUQ, which they completed for the chatbot they had just evaluated. They then repeated this process for the other two chatbots.

6.3.4.2 Round 2

Approximately two weeks after completing round one, the researcher sent each participant a link to round 2 of the study. Participants visited this page and, as for round one, were free to evaluate the chatbots in any order. Participants were able to revisit and use the chatbots again if they felt they needed to do so, again spending no more than 5 to 8 minutes using each one. Participants then completed a CUQ for each of the three chatbots, following the same procedure as during round 1.

6.3.5 Data Analysis

6.3.5.1 CUQ Score Analysis

The CUQ was digitised on Qualtrics. Data from the questionnaire were exported from Qualtrics as a CSV and imported into Microsoft Excel. Scores were calculated out of 100 using the procedure discussed in **chapter 5, section 3.6.3**. The mean score and standard deviation were calculated for each of the three chatbots.

6.3.5.2 Test-Retest Reliability Analysis

Test-retest reliability (or intra-rater reliability) was measured at questionnaire level by using paired t-tests to compare mean chatbot scores at each round. Pearson correlation tests were used to determine correlation between chatbot scores at each round. Test-Retest reliability was measured at question level by calculating the mean score per question for each chatbot and using Spearman correlation tests to determine correlation between question responses (ratings) at each round. Statistical significance for t-tests was assumed where $p < 0.05$. Correlation was assumed to be poor when $r < 0.5$, and strong when r was > 0.7.

6.3.5.3 Construct Validity Analysis

Construct validity was measured at questionnaire level by using paired t-tests to compare CUQ scores of each of the three chatbots. Construct validity was measured at question level using paired t-tests to compare responses to each question for each chatbot, and Spearman correlation tests to measure correlation. Statistical significance for t-tests was assumed where $p < 0.05$. Correlation was assumed weak when $r < 0.5$, and strong when $r > 0.7$.

6.3.5.4 Exploratory Factor Analysis

Exploratory Factor Analysis was used to identify the factors within the CUQ, and determine which questions were strongly correlated with each factor. Findings from the factor analysis were used to determine which questions could be removed from the CUQ in order to reduce its size and increase its efficiency. McCaffrey (2017), suggests that identified factors are only significant if the Sum of Squares loadings for each factor are greater than 1.0, thus a cut-off point of 1.0 was used to determine which factors were significant.

6.3.6 Ethics

Ethical approval was obtained from the Communication Ethics Filter Committee and the study was conducted in compliance with the Ulster University Code of Practice for Professional Integrity in the Conduct of Research and the Policy for the Governance of Research Involving Human Participants. The study was designed to comply with the five core principles of beneficence, non-maleficence, honesty and integrity, confidentiality and informed consent, which are discussed in chapter 3. Although this study was designated low risk, identified potential risks included anonymity, confidentiality and inconvenience to participants. Anonymity and confidentiality were assured by collecting only a minimum amount of personal data for demographic purposes, and by assigning a unique six-digit ID number to participants, which they used instead of their name when completing their demographics questionnaire and CUQs. All participants were able to download their own copy of the PIS and thus were able to provide informed consent. Participants could have potentially been inconvenienced by the need to use each of the three chatbots and complete two CUQs for each one, however this inconvenience was reduced by the study website, which gave participants access to all chatbots and the CUQ from any computer or mobile device, thus making it possible for participants to evaluate chatbots and complete questionnaires at their own convenience. Personal data collected during usability tests were collected in compliance with GDPR 2018.

6.4 Results

6.4.1 Participant Demographics

There was a total of 156 CUQ survey completions (26 participants completed the CUQ for 3 different chatbots and for 2 rounds → 26*3*2 = 156). There were more female participants (15, 58%) than male (11 participants, 42%). Participant ages ranged from 18-25 years to over 50 years, and mode age range was 31-40 years (6 participants, 23%). Most participants (22, 85%) reported that English was their first language. Most participants ranked their technical ability with mobile devices as either 3 or 4 (9 participants, 35%). Mean technical ability was 3.8±0.90, the median and mode were 5. Participant demographics are summarised in **Tables 6.3 to 6.5**.

Table 6.3: CUQ Validation Participant Demographics (Gender, Age Range (n=26))

Demographics		n	%
Gender	Female	15	58%
	Male	11	42%
Age Range	18-25	3	12%
	26-30	5	19%
	31-40	6	23%
	41-45	3	12%
	46-50	4	15%
	Over 50	5	19%
	Mode	31-40	

Table 6.4: CUQ Validation Participant Demographics (Occupation, First Language(n=26))

Demographics		n	%
Occupation	Student	1	3%
	PhD Researcher	9	35%
	Non-Student	16	62%
First Language	English	22	85%
	Non-English	4	15%

Table 6.5: CUQ Validation Participant Demographics (Technical Ability (n=26))

Demographics		n	%
Technical Ability	1 Low	0	0%
	2	1	3%
	3	9	35%
	4	9	35%
	5 High	7	27%
	Mean	**3.8±0.9**	
	Median	**4**	
	Mode	**3/4**	

6.4.2 Reliability of Expert Panel – Chatbot Selection

As discussed in **section 6.3.5**, a panel of experts evaluated the three chatbots based on pre-determined criteria (see **table 6.2**). Although the panel members were aware of which three chatbots had been selected prior to their evaluation, they were unaware how each chatbot had been rated by the PhD researcher. During the final committee stage of the chatbot selection, the experts unanimously agreed on the "good", "average" and "poor" quality ratings for *WoeBot*, *Weight Loss Bot* and *FlyBot* respectively, which were consistent with the ratings given to each chatbot by the PhD researcher. Thus, it may be determined that the inter-rater reliability of the three experts plus the researcher was 100%, since all panel members unanimously agreed ratings without prior knowledge of how the other panel members had rated each chatbot. During future validation studies (see **section 6.7**) it would be prudent to evaluate several chatbots per quality category using a larger panel of experts.

6.4.3 Test-Retest Reliability

6.4.3.1 Questionnaire Level

Mean CUQ scores for each chatbot per round were calculated using the procedure discussed in **chapter 5, section 3.6.3**. Chatbot scores per round were compared using t-tests and Pearson correlation tests (as discussed in section 3.5.2). Mean scores and

standard deviations are shown in **Table 6.6**, along with t-test p-values and correlation coefficients. Boxplots and scatter plots of chatbot CUQ scores are presented in **Figures 6.1 to 6.3**. T-Test values are greater than 0.05 for all three chatbots, thus differences in scores across rounds are not significant. Correlations are weaker for *WoeBot* rounds than for *Weight Loss Bot* and *FlyBot* rounds, but all correlations are strong ($r > 0.7$) suggesting intra-rater reliability.

Table 6.6: Comparison of chatbot CUQ scores by round

Chatbot	Round	Score	T-Test	Correlation
WoeBot	1	72.4±17.9	0.29	0.74
	2	74.9±13.6		
Weight Loss Bot	1	69.2±16.9	0.14	0.84
	2	66.4±16.3		
FlyBot	1	57.3±16.9	0.33	0.81
	2	59.4±17.7		

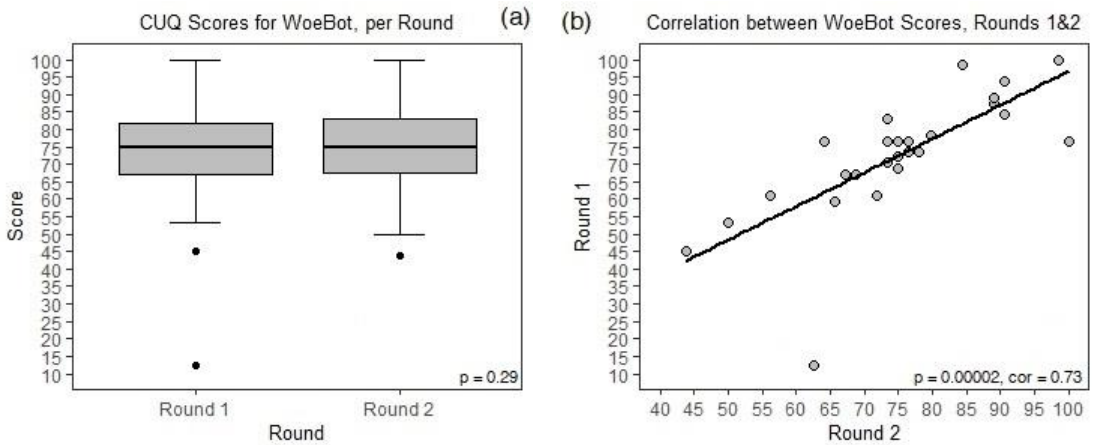


Figure 6.1: WoeBot CUQ Scores, per round
(a) Boxplot (b) Scatterplot. Error Bars represent SD.

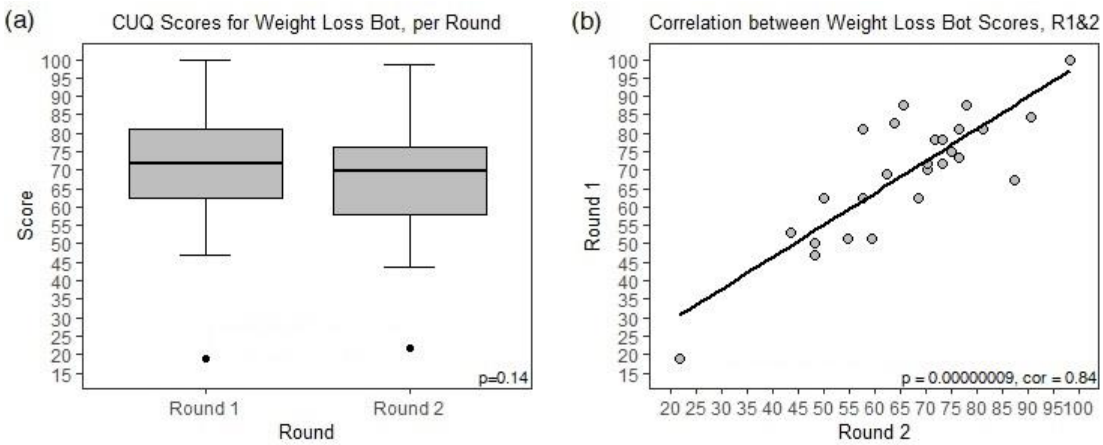


Figure 6.2: *Weight Loss Bot* CUQ Scores, per round
(a) Boxplot (b) Scatterplot. Error Bars represent SD.

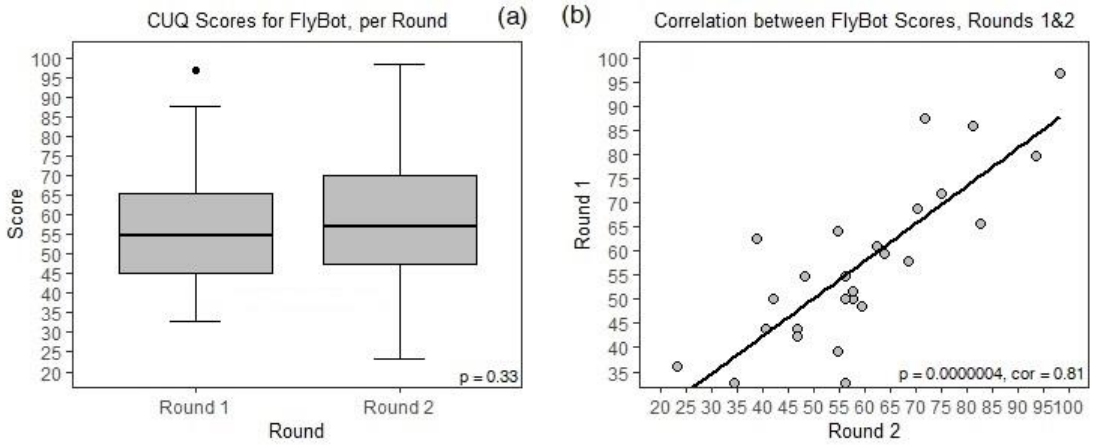
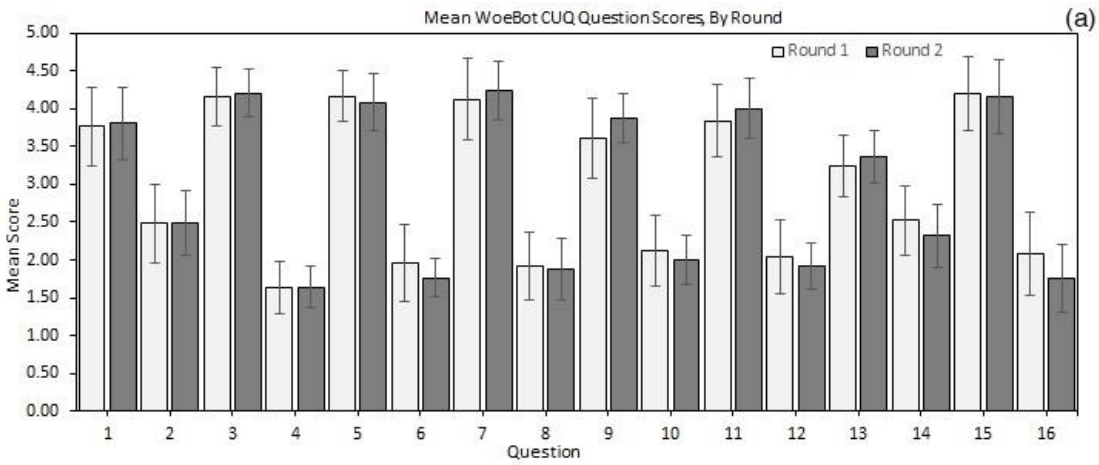


Figure 6.3: *FlyBot* CUQ Scores, per round
(a) Boxplot (b) Scatterplot. Error Bars represent SD.

6.4.3.2 Question Level

Chatbot question scores per round were compared as discussed in section 3.5.2. Bar plots are presented in **Figure 6.4**. P-values were generally not significant (greater than 0.05) for CUQ questions per round, and correlation was generally moderate (r was greater than 0.3 and r was less than 0.7). Visual inspection of the bar plots at question level per round, shows strong similarity in ratings per round. Mean question scores per chatbot at each round are in **Appendix 30**, along with p-values and correlation coefficients.



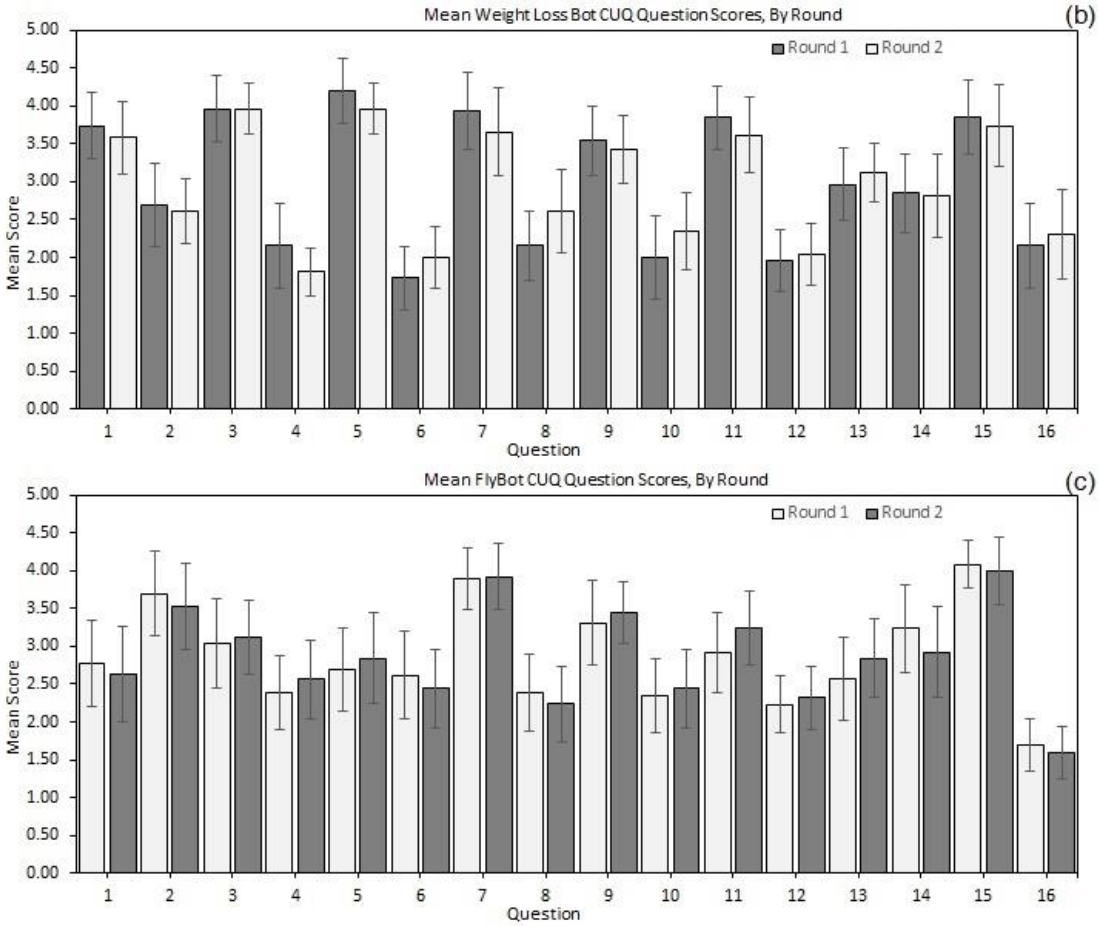


Figure 6.4: Mean CUQ Question Scores, per chatbot, per round
(a) *WoeBot* (b) *Weight Loss Bot* (c) *FlyBot*. Error Bars represent SD.

6.4.3.3 Reliability of CUQ Questions

Mean CUQ question scores were analysed to determine the percentage differences between scores for each round, and these are presented in **Figure 6.5.** Two questions (Q7 & Q9) showed percentage differences > 5% for one chatbot, nine questions (Q4-Q6, Q8, Q10, Q11, Q13, Q14 and Q16) showed percentage differences > 5% for at least two chatbots, and two questions (Q6 and Q16) showed percentage differences > 5% for all three chatbots. Percentage differences were > 10% for seven questions (Q4, Q6, Q8, Q10, Q11, Q13, Q16) and > 15% for five questions (Q4, Q6, Q8, Q10, Q16). The largest percentage differences were found in Question 8 (*Weight Loss Bot*, 21.4%), Q10 (*Weight Loss Bot*, 17.3%) and Q4 (*Weight Loss Bot*, 16.1%). *Weight Loss Bot* and *FlyBot* had the most questions with percentage differences > 5% (*Weight Loss Bot* 9, *FlyBot* 8).

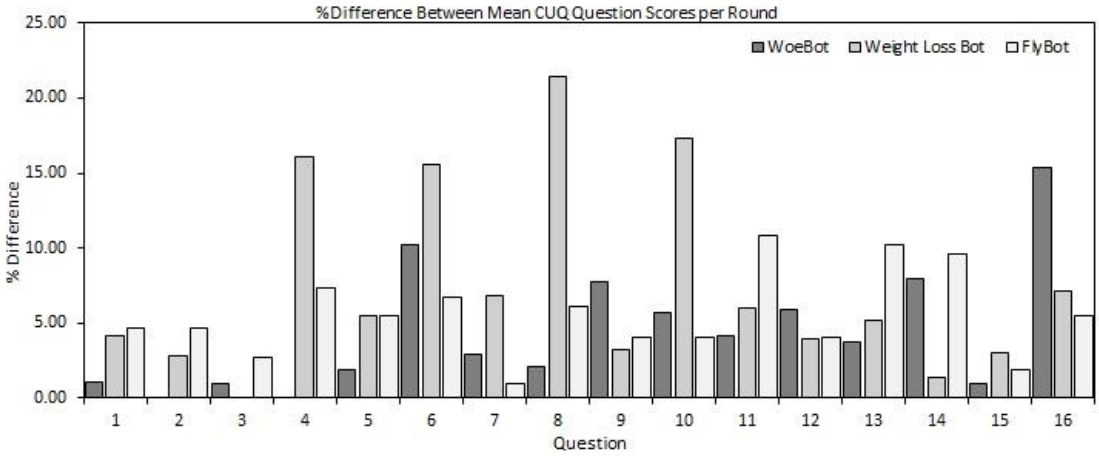


Figure 6.5: Percentage differences between CUQ Question scores per round

6.4.4 Construct Validity

6.4.4.1 Questionnaire Level

CUQ scores per chatbot were compared as discussed in section 3.5.3. A boxplot of CUQ scores per chatbot is presented in **Figure 6.6**. T-test p-values suggest a significant difference between chatbot scores (p was less than 0.05) providing construct validity. Hence, users completing the CUQ provide scores that discriminate between poor, moderate and optimal chatbot usability.

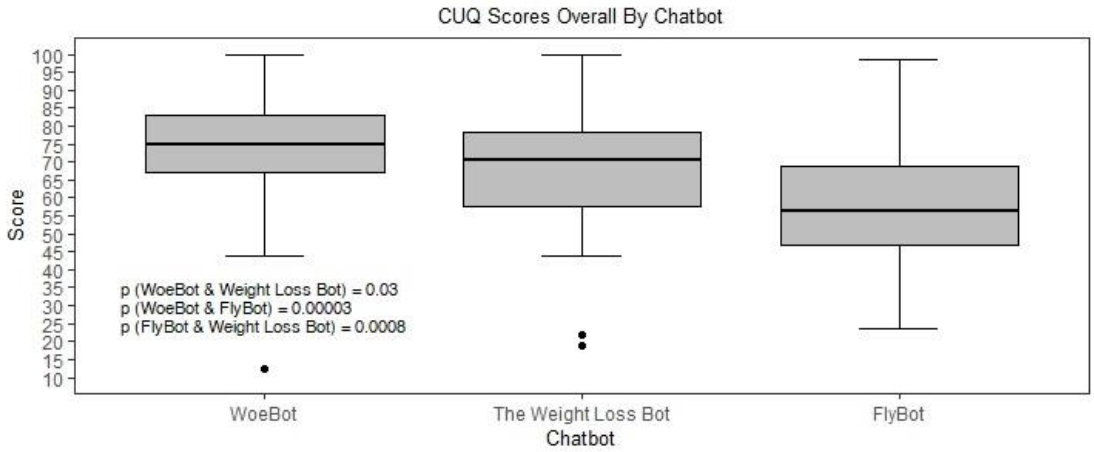
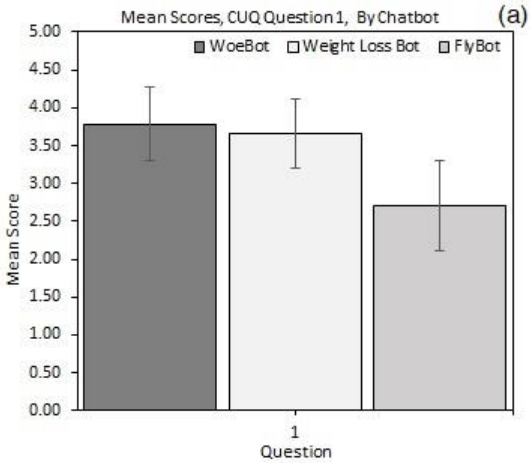


Figure 6.6: Chatbot CUQ scores, per chatbot
Error Bars represent SD.

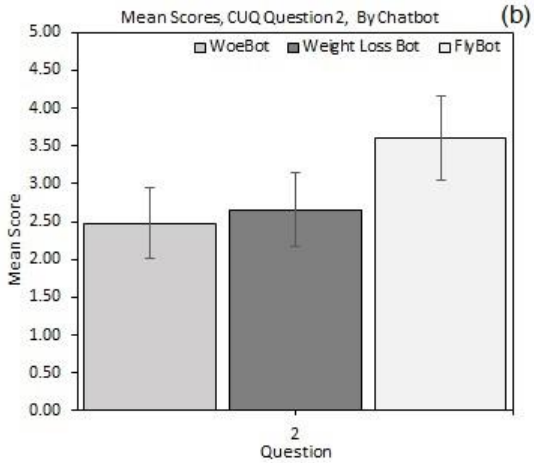
6.4.4.2 Question Level

CUQ question responses per chatbot were compared as discussed in section 3.5.3. Bar plots of CUQ question responses are found in **Figure 6.7**. T-test p-value significance and correlation strengths are summarised in **Table 6.7**. There was generally significant difference between *WoeBot* and *FlyBot* question scores and *Weight Loss Bot* and *FlyBot*

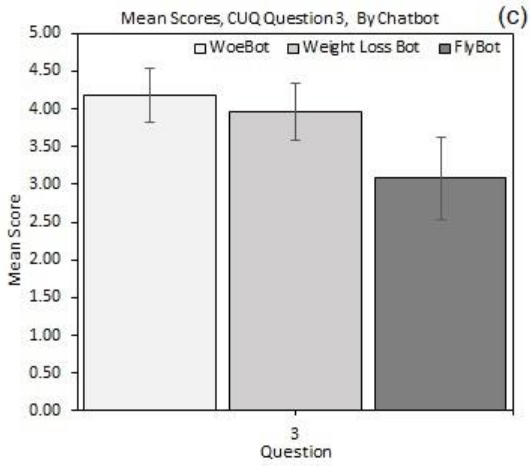
question scores, but generally no significant difference between *WoeBot* and *Weight Loss Bot* question scores. Correlation between chatbot question scores was generally weak.



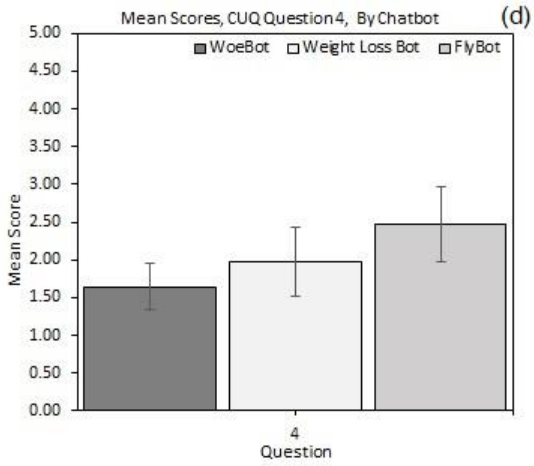
WB&WLB: t-test = 0.47, cor = 0.17
WB&FB: t-test = 0.00002 cor = -0.09
WLB&FB: t-test = 0.000004 cor = 0.15



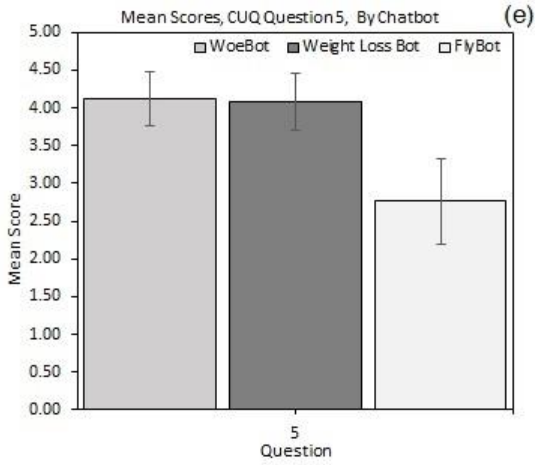
WB&WLB: t-test = 0.34, cor = 0.09
WB&FB: t-test = 0.000006, cor = -0.24
WLB&FB: t-test = 0.000004, cor = 0.15



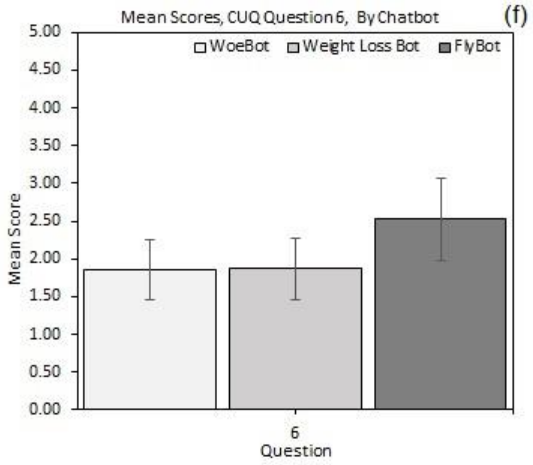
WB&WLB: t-test = 0.08, cor = 0.40
WB&FB: t-test = 0.0000007, cor = -0.17
WLB&FB: t-test = 0.0000004, cor = 0.29



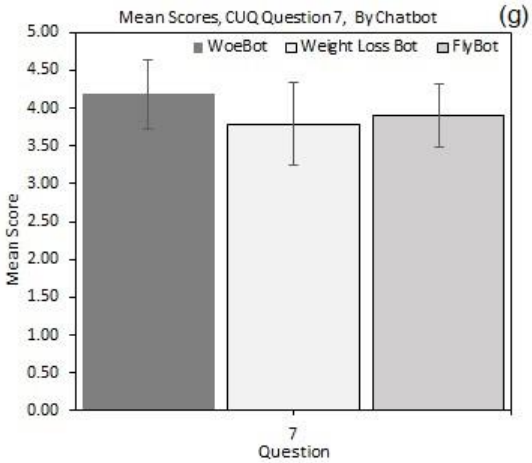
WB&WLB: t-test = 0.02, cor = 0.30
WB&FB: t-test = 0.000009, cor = -0.09
WLB&FB: t-test = 0.006, cor = 0.23



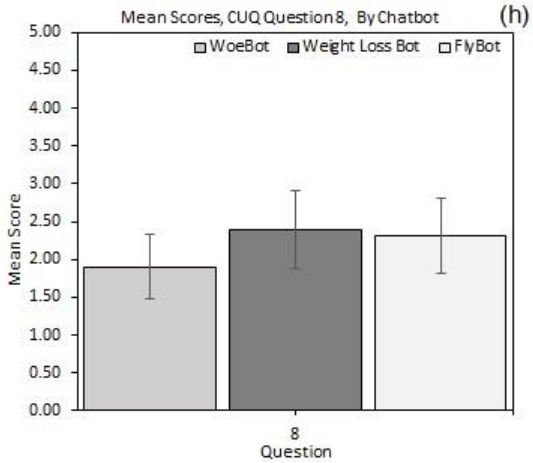
WB&WLB: t-test = 0.74, cor = 0.40
WB&FB: t-test = 0.000000005, cor = -0.09
WLB&FB: t-test = 0.00000000004, cor = 0.29



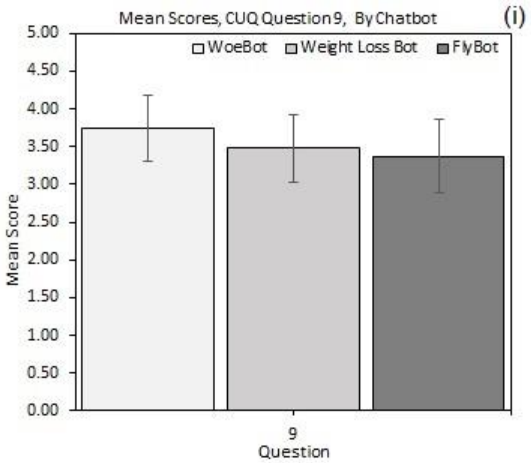
WB&WLB: t-test = 0.97, cor = 0.30
WB&FB: t-test = 0.0009, cor = -0.03
WLB&FB: t-test = 0.00008, cor = 0.34



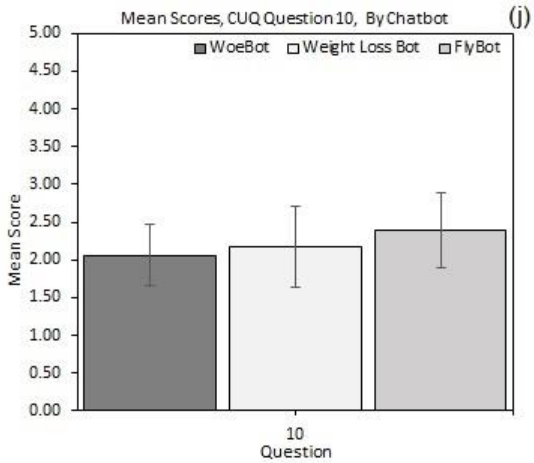
WB&WLB: t-test = 0.05, cor = 0.12
WB&FB: t-test = 0.15, cor = -0.15
WLB&FB: t-test = 0.55, cor = 0.046



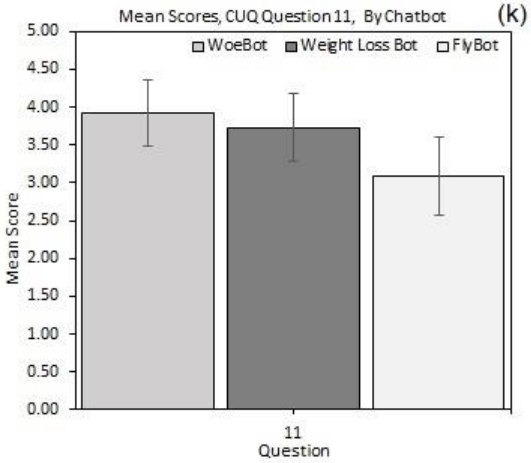
WB&WLB: t-test = 0.009, cor = 0.16
WB&FB: t-test = 0.03, cor = 0.003
WLB&FB: t-test = 0.70, cor = 0.22



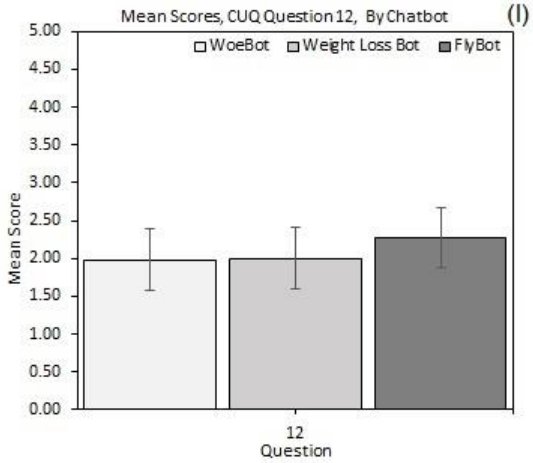
WB&WLB: t-test = 0.05, cor = 0.52
WB&FB: t-test = 0.04, cor = 0.15
WLB&FB: t-test = 0.46, cor = 0.33



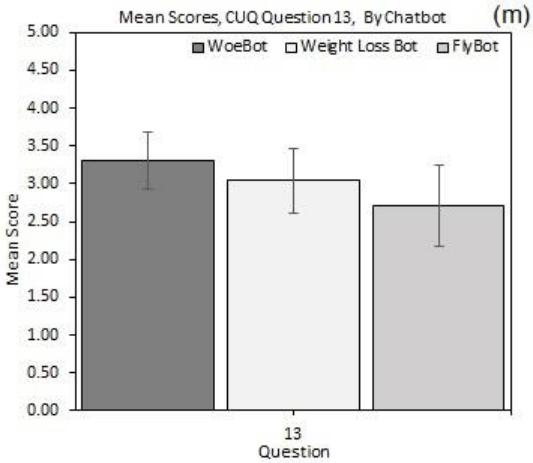
WB&WLB: t-test = 0.43, cor = 0.45
WB&FB: t-test = 0.07, cor = -0.026
WLB&FB: t-test = 0.28, cor = 0.05



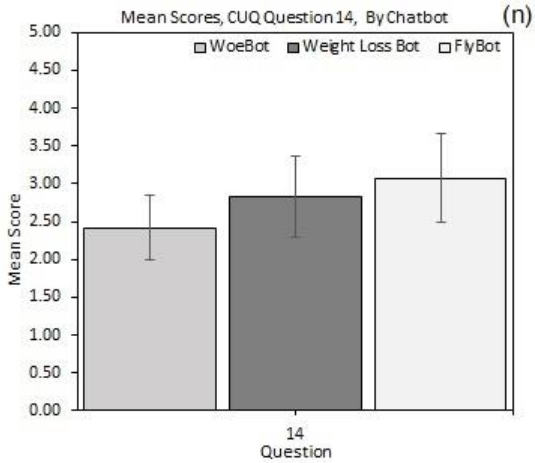
WB&WLB: t-test = 0.22, cor = 0.32
WB&FB: t-test = 0.0001, cor = -0.10
WLB&FB: t-test = 0.0007, cor = 0.17



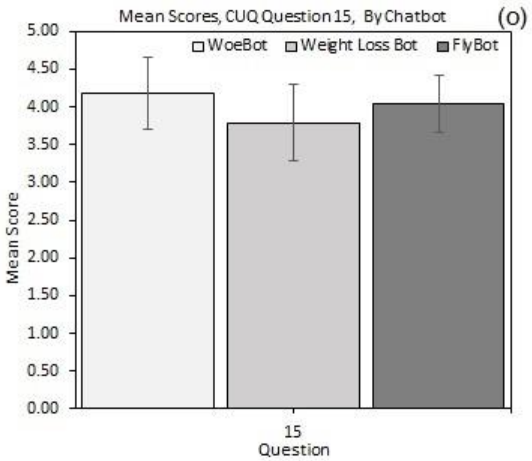
WB&WLB: t-test = 0.85, cor = 0.56
WB&FB: t-test = 0.06, cor = 0.09
WLB&FB: t-test = 0.08, cor = 0.12



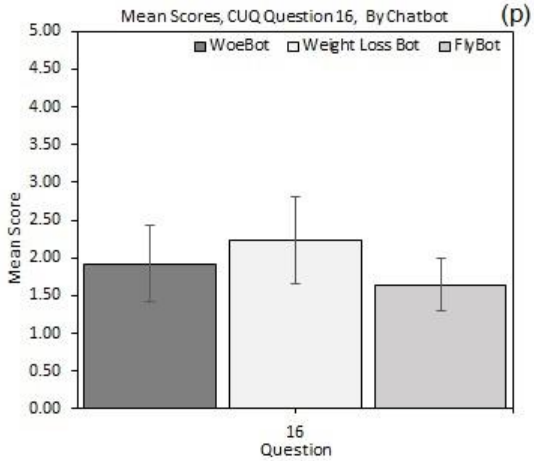
WB&WLB: t-test = 0.03, cor = 0.45
WB&FB: t-test = 0.0005, cor = 0.17
WLB&FB: t-test = 0.01, cor = 0.54



WB&WLB: t-test = 0.02, cor = 0.19
WB&FB: t-test = 0.002, cor = 0.031
WLB&FB: t-test = 0.17, cor = 0.31



WB&WLB: t-test = 0.03, cor = 0.28
WB&FB: t-test = 0.40, cor = 0.03
WLB&FB: t-test = 0.15, cor = 0.15



WB&WLB: t-test = 0.06, cor = 0.55
WB&FB: t-test = 0.10, cor = 0.13
WLB&FB: t-test = 0.001, cor = 0.26

Figure 6.7: Mean CUQ Question Scores, by Chatbot
Error Bars represent SD.

Table 6.7: Summary of p-value significance and Correlations, CUQ Questions

Chatbots	**WB&WLB**		**WB&FB**		**WLB&FB**	
Question	**Significant**	**Correlation**	**Significant**	**Correlation**	**Significant**	**Correlation**
Q1	No	Weak	Yes	Weak	Yes	Weak
Q2	No	Weak	Yes	Weak	Yes	Weak
Q3	No	Weak	Yes	Weak	Yes	Weak
Q4	Yes	Weak	Yes	Weak	Yes	Weak
Q5	No	Weak	Yes	Weak	Yes	Weak
Q6	No	Weak	Yes	Weak	Yes	Weak
Q7	No	Weak	No	Weak	No	Weak
Q8	Yes	Weak	Yes	Weak	No	Weak

Chatbots	WB&WLB		WB&FB		WLB&FB	
Question	**Significant**	**Correlation**	**Significant**	**Correlation**	**Significant**	**Correlation**
Q9	Yes	Moderate	Yes	Weak	No	Weak
Q10	No	Weak	No	Weak	No	Weak
Q11	No	Weak	Yes	Weak	Yes	Weak
Q12	No	Moderate	No	Weak	No	Weak
Q13	Yes	Weak	Yes	Weak	Yes	Moderate
Q14	Yes	Weak	Yes	Weak	No	Weak
Q15	Yes	Weak	No	Weak	No	Weak
Q16	No	Moderate	No	Weak	Yes	Weak

6.4.5 Factor Analysis

6.4.5.1 Principal Component Analysis

The scree plot for the CUQ Principal Component Analysis (**Figure 6.8**) suggested a maximum of five factors in the CUQ and perhaps a minimum of three.

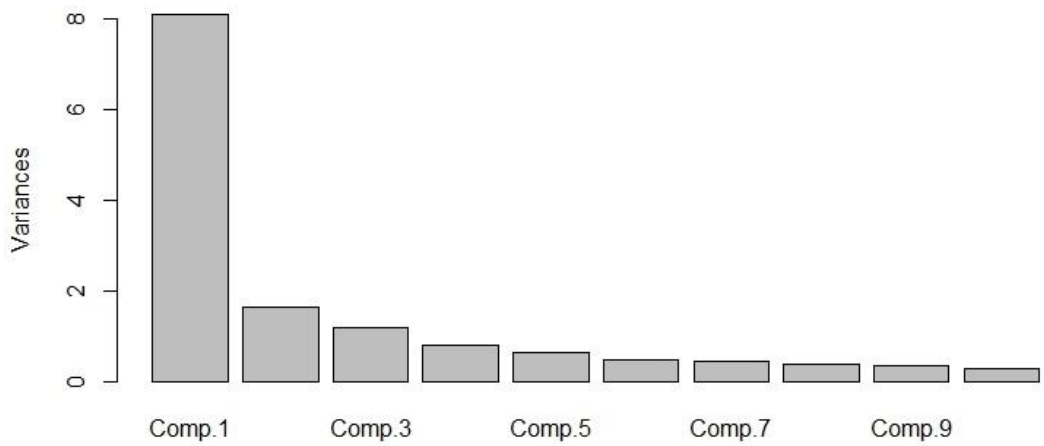


Figure 6.8: Scree Plot for CUQ Principal Component Analysis

6.4.5.2 Five Factor Analysis

Factor Analysis using five factors suggested that the Sum of Square loadings of four factors were greater than 1.0, thus were significantly influencing the variance, while the Sum of Square loading of the fifth factor was less than 1.0, thus was not significant. as shown in **Table 6.8**.

Table 6.8: Sum of Square Loadings: 5-Factor Analysis

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
SS Loadings	3.09	3.06	2.19	1.98	0.97

6.4.5.3 Four Factor Analysis

Factor Analysis using four factors suggested that the Sum of Square loadings of four factors were significant (greater than 1.0), as shown in **Table 6.9**.

Table 6.9: Sum of Square Loadings: 4-Factor Analysis

	Factor 1	Factor 2	Factor 3	Factor 4
SS Loadings	3.33	3.25	2.20	1.97

6.4.5.4 Identifying Factors

Using a cut-off point of 0.5, factor loadings were analysed to identify which questions were closely correlated with each factor. Factors and their correlated questions are summarised in **Table 6.10**. Factors were identified based on correlated questions to determine which aspects of chatbot usability were being measured by the CUQ. **Table 6.10** shows correlations between questions and the identified factors within the CUQ. Factors were named based on their correlated questions; these are shown in **Table 6.11.** From this table it may be seen that there are four factors being measured by the CUQ, Personality, User Experience, Error Handling and Onboarding. The only question that did not suitably match the suggested factor name was Q9, correlated with Factor 1 (Personality).

Table 6.10: Factors and correlated questions, with loadings

Factor	Question	Loading
1	Q1	0.74
	Q2	-0.78
	Q3	0.65
	Q4	-0.52
	Q9	0.57
2	Q7	0.82
	Q8	-0.70
	Q15	0.86
	Q16	-0.60

Factor	Question	Loading
3	Q13	0.82
	Q14	-0.78
4	Q5	0.88
	Q6	-0.59

Table 6.11: Suggested Factor Names, Based on Correlated Question Text

Factor	Question	Text	Suggested Factor Name
1	Q1	The chatbot's personality was realistic and engaging	Personality
	Q2	The chatbot seemed too robotic	
	Q3	The chatbot was welcoming during initial setup	
	Q4	The chatbot seemed very unfriendly	
	Q9	*The chatbot understood me well*	
2	Q7	The chatbot was easy to navigate	User Experience
	Q8	It would be easy to get confused when using the chatbot	
	Q15	The chatbot was easy to use	
	Q16	The chatbot was very complex	
3	Q13	The chatbot coped well with any errors or mistakes	Error Handling
	Q14	The chatbot seemed unable to cope with any errors	
4	Q5	The chatbot explained its scope and purpose well	Onboarding
	Q6	The chatbot gave no indication as to its purpose	

6.4.5.5 Exclusion of Questions

The CUQ is designed to be comparable to SUS in both size and function (i.e. calculation of scores). The CUQ is functionally comparable to the SUS because both questionnaires use balanced questions i.e. pairs of questions related to positive and negative aspects of system usability, and both use a similar method for score calculation (see sections 5.3.6.1 and 5.3.6.3 of Chapter 5 for discussion of calculation of scores for SUS and CUQ). However as a 16-item questionnaire, the CUQ is not comparable to SUS in terms of size, as SUS has only ten items. Therefore, it is desirable to reduce the size of the CUQ as much as possible, ideally to no more than ten questions, to improve comparability with SUS. Following factor analysis, a heuristic for exclusion of questions was developed based on identified factors and correlated questions. The heuristic is a simple three-step process, summarised in **Table 6.12**.

Table 6.12: Heuristic for selection of questions for exclusion from the CUQ

Step	Description
1	Identify and remove any questions which do not correlate with any factor
2	Identify and remove any questions which do not relate to their correlated factor
3	Identify and remove, as appropriate, questions from factors which are correlated with more than two questions

Based on the above heuristic, six questions were identified as potentially excludable from the CUQ, and these are summarised in **Table 6.13**. Removal of questions based on **Table 6.13** reduces the CUQ to either 10 or 12 items. These new questionnaires are found in **Table 6.14** and **Table 6.15**. There were no questions which did not correlate with any factor. One question (Q9) was excluded because it correlated with a factor (personality) which it was not related to (i.e. did not measure). Three questions (Q10, Q11 and Q12) were excluded because they were not correlated with any of the four factors. Two factors (Personality and User Experience) were correlated with more than two questions, and in order to maintain the balance between positive aspect (odd-numbered) and negative aspect (even-numbered) questions, it was determined that two questions should be excluded from one of these two factors. The "User Experience" factor was selected; however it would have been equally justified to exclude questions from the "Personality" factor instead.

Table 6.13: Questions to exclude from CUQ

Question to Exclude	Rationale	New Questionnaire Size
Q9	Q9 only correlates with "Personality" factor (Factor 1), but it does not measure personality.	12-item
Q10	Not correlated with any factor.	
Q11		
Q12		
Q15	Correlated with "User Experience" (Factor 2), which is already correlated with Q7 and Q8.	10-item
Q16	Correlated with only User Experience/Factor 2, which is already correlated with Q7 and Q8.	

6.4.5.6 Proposed new 10- and 12-item CUQs

Table 6.14: Proposed 12-item CUQ

Question		Question Text	Linked Factor	
Old Number	**New Number**		**Number**	**Name**
	Q1	The chatbot's personality was realistic and engaging	1	Personality
	Q2	The chatbot seemed too robotic		
	Q3	The chatbot was welcoming during initial setup		
	Q4	The chatbot seemed very unfriendly		
	Q5	The chatbot explained its scope and purpose well	4	Onboarding
	Q6	The chatbot gave no indication as to its purpose		
	Q7	The chatbot was easy to navigate	2	User Experience
	Q8	It would be easy to get confused when using the chatbot		
Q13	Q9	The chatbot coped well with any errors or mistakes	3	Error Management
Q14	Q10	The chatbot seemed unable to cope with any errors		
Q15	Q11	The chatbot was easy to use	2	User Experience
Q16	Q12	The chatbot was very complex		

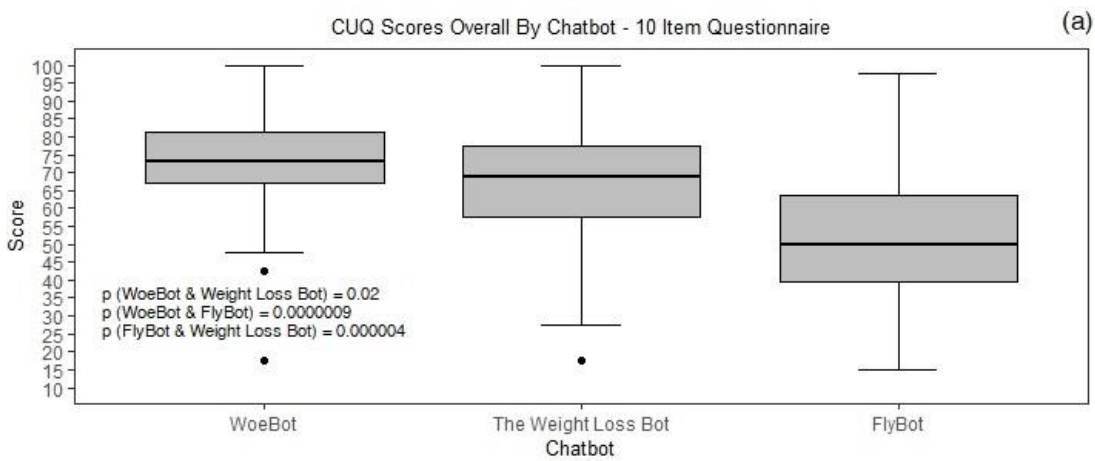
Table 6.15: Proposed 10-item CUQ

Question		Question Text	Linked Factor	
Old Number	**New Number**		**Number**	**Name**
	Q1	The chatbot's personality was realistic and engaging	1	Personality
	Q2	The chatbot seemed too robotic		
	Q3	The chatbot was welcoming during initial setup		
	Q4	The chatbot seemed very unfriendly		

Question		Question Text	Linked Factor	
Old Number	New Number		Number	Name
	Q5	The chatbot explained its scope and purpose well	4	Onboarding
	Q6	The chatbot gave no indication as to its purpose		
	Q7	The chatbot was easy to navigate	2	User Experience
	Q8	It would be easy to get confused when using the chatbot		
Q13	Q9	The chatbot coped well with any errors or mistakes	3	Error Management
Q14	Q10	The chatbot seemed unable to cope with any errors		

6.4.5.7 Construct Validity of 10- and 12-item CUQs (Questionnaire Level)

Using the same data for all three chatbots, the construct validity of the 10- and 12-item questionnaires proposed above was measured as discussed in section 3.5.4. Boxplots of CUQ scores for 10- and 12-item questionnaires are found in **Figure 6.9**. T-test p-values suggest a significant difference between chatbot scores (p was less than 0.05). Validity of shorter CUQs (less than ten questions) was not measured as the CUQ is designed to be comparable to SUS in terms of size and function, and it was considered unnecessary to measure validity of the CUQ with fewer than ten questions.



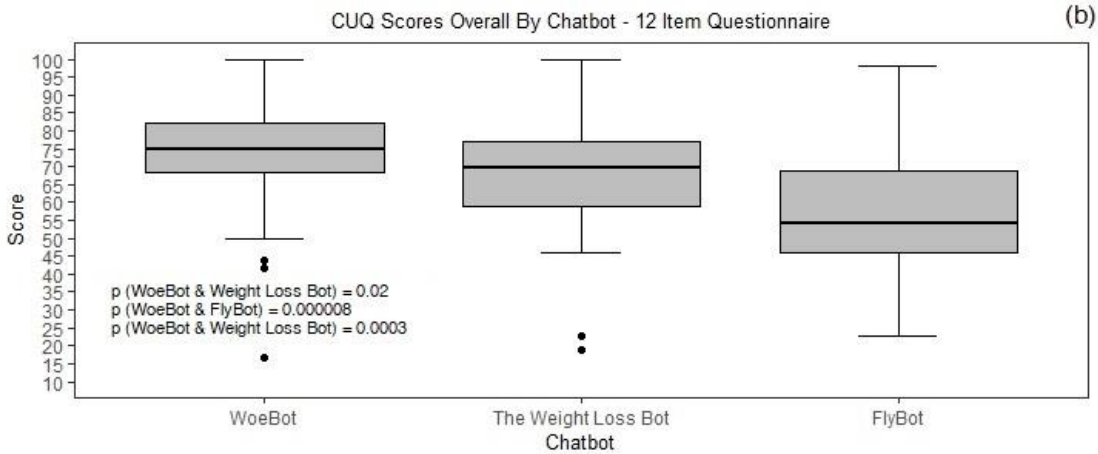


Figure 6.9: Chatbot CUQ scores, per chatbot
(a) 10-item CUQ, (b) 12-item CUQ. Error Bars represent SD.

6.4.6 Comparison of CUQ Validation scores with *WeightMentor* CUQ Score

Overall CUQ scores for each of the three chatbots (across two rounds) are compared with *WeightMentor's* CUQ score in **Table 6.16.** A boxplot of scores is shown in **Figure 6.10**. Mean *WeightMentor* score was lower than *WoeBot* and higher than *Weight Loss Bot*, and variation (Standard Deviation) of *WeightMentor* scores across participants was lower than for each of the three chatbots. Results from T-tests are shown in **Table 6.17** and suggest no significant difference ($p > 0.05$) between scores for *WeightMentor* and *WoeBot* and *WeightMentor* and *Weight Loss Bot*, and significant difference ($p < 0.05$) between scores for *WeightMentor* and *FlyBot*. These p-values indicate that CUQ score distribution for *WeightMentor* lies between *WoeBot* and *Weight Loss Bot* score distributions and is closest to the score distribution for *WoeBot*, i.e. the "good" category.

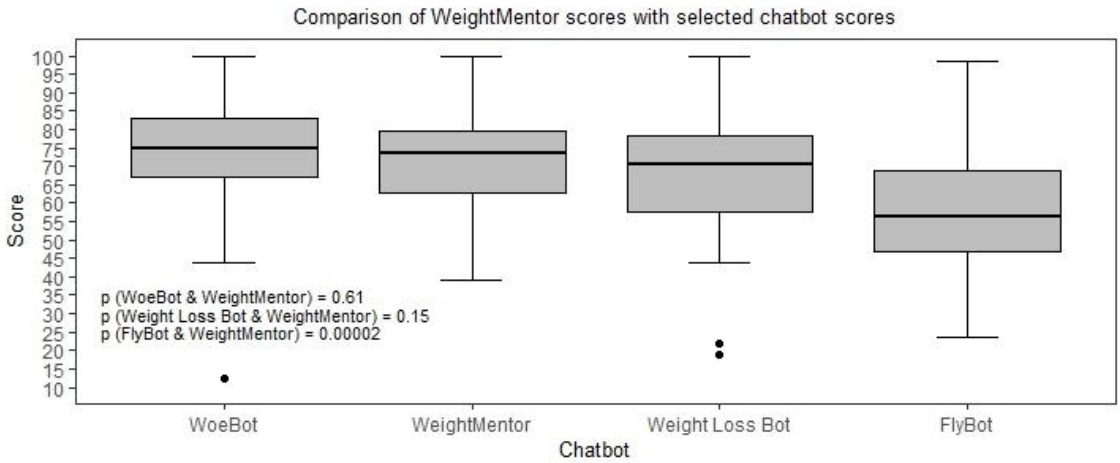


Figure 6.10: Comparison of WeightMentor scores with selected chatbot scores.
Error bars represent SD.

Table 6.16: Comparison of *WeightMentor's* CUQ score with scores from selected chatbots

Chatbot	Score
WoeBot	73.7±15.8
WeightMentor	*72.2±13.9*
Weight Loss Bot	67.8±16.5
FlyBot	58.4±17.2

Table 6.17: T-Tests of WeightMentor scores against selected chatbot scores

Test	p-value
WoeBot & WeightMentor	0.61
Weight Loss Bot & WeightMentor	0.15
FlyBot & WeightMentor	0.00002

6.5 Discussion

There was no significant difference between chatbot scores between rounds, and correlation coefficients are all strong (greater than 0.7), as shown in **Table 6.6** and **Figures 6.1 – 6.3**, which suggests that test-retest reliability exists at the questionnaire level. It may be observed that while correlation coefficients for *Weight Loss Bot* and *FlyBot* are greater than 0.8, the correlation coefficient for *WoeBot* is less than 0.8 but greater than 0.7. This is due to the presence of an outlier, visible on **Figure 6.1(b)**. For this participant, the *WoeBot* CUQ score for Round 1 was 12.5, but was 62.5 for Round 2. It is unknown why this participant scored the chatbot more highly during Round 2, however the difference in scores has reduced the correlation coefficient slightly. Test-retest reliability also exists at the question level, based on the bar plots in **Figure 6.4**, which show only a small difference between mean question scores per round for each chatbot. Additionally, t-test p-values were all greater than 0.05 for mean question scores and thus were not significant.

It was intended that participants would complete round 2 questionnaires approximately 2 weeks after Round 1, however it was not always possible to do this. As the research study ran during summer 2019, some participants were on holiday and were unable to complete round 2 until they returned. Other participants forgot to complete round 2 questionnaires and were sent a reminder. Despite participants not completing both rounds within 2

weeks of each other, CUQ scores and mean question scores were not significantly different. The suggestion is therefore that test-retest reliability is not influenced by time.

Based on **Figure 6.6,** it is possible to tell the difference between each of the three chatbots *WoeBot*, *Weight Loss Bot,* and *FlyBot*. *WoeBot* scores are highest, followed by *Weight Loss Bot* and then *FlyBot*. T-test p-values for chatbot scores demonstrate statistical significance. These scores are consistent with the decision of the expert panel (see section 3.3) and demonstrate that the CUQ has construct validity at questionnaire level. Based on the bar plots in **Figure 6.4** and **Figure 6.7** a difference may be observed between mean question scores for each chatbot, suggesting that construct validity exists at the question level. Results from the correlation analysis of individual question scores across chatbots suggested that correlation between individual question scores per chatbot was generally weak, which was expected. For any given question, it is unlikely that the mean score for a supposedly *good* quality chatbot (such as *WoeBot*) would be the same as the mean score for a supposedly *average* or *poor* quality chatbot (such as *Weight Loss Bot* and *FlyBot*), thus would be unlikely that specific question means would correlate across chatbots. Although this correlation analysis of individual question scores across chatbots was largely superfluous, the results support the existing evidence that construct validity exists at the question level.

Twelve questions showed percentage differences of > 5% between mean scores at each round of questionnaires. The most unreliable questions were Q8, Q10 and Q4 in order from greatest percentage difference to least percentage difference. Q8 was “It would be easy to get confused when using the chatbot, Q10 was “The chatbot failed to recognise a lot of my inputs”, and Q4 was “The chatbot seemed very unfriendly”. All three of these questions relate to negative aspects of the chatbot User Experience (UX). Differences between question round scores for Q6 and Q16 were > 5% for all three chatbots. Q6 was “The chatbot gave no indication as to its purpose”, and Q16 was “The chatbot was very complex”. Both questions are also related to negative aspects of chatbot UX. Most (n=9) of the unreliable CUQ questions showed greatest percentage differences in scores between rounds. This may suggest that negative aspects of chatbot UX are more difficult to measure consistently than positive aspects.

Principal Component Analysis of the CUQ identified a maximum of five possible factors. Factor Analysis of five factors showed that the Sum of Square loading of the fifth factor

was less than 1.0. Based on McCaffrey, 2017, it may be assumed that a factor is only significant if its Sum of Square loading is greater than 1.0, thus it was determined that there were at most four factors in the CUQ. This was confirmed using Factor Analysis of four factors. Based on CUQ question topics, these four factors were identified as *Personality, User Experience, Error Management* and *Onboarding*. Excluding questions from the CUQ based on these identified factors (see sections 4.4.5 & 4.4.6) meant the CUQ could potentially be reduced to either 12 or 10 items. It was desired to reduce the CUQ to 10 items if possible, as the questionnaire is intended to be comparable to the System Usability Scale (SUS) (see chapter 5), which is also a 10-item scale (Brooke, 1996). Following exclusion of questions based on the heuristic in **Table 6.12**, the new 10- and 12-item questionnaires were tested for construct validity using data for each of the three chatbots and significant difference was observed between chatbot scores. Thus, it may be determined that the CUQ exhibits construct validity with no more than 10 items.

As discussed in the introduction to this chapter, Martín et al. (2017), suggest that chatbot UX may be measured using seven different aspects. The CUQ measures only three of the aspects proposed by Martín et al. (Personality, Error Management and Onboarding). However, the "User Experience" factor of the CUQ covers questions relating to navigation and ease of use, thus it is reasonable to suggest that the "Navigation" aspect is being measured by the CUQ.

The CUQ score from the *WeightMentor* chatbot usability tests (see chapter 5) was compared with CUQ scores for each of the three chatbots used in this validation study. *WeightMentor* scored 4.4 and 13.8 points higher than *Weight Loss Bot* and *FlyBot*, but only 1.5 points lower than *WoeBot*. This suggests that *WeightMentor* may potentially be rated "good" quality. Variance (Standard Deviation) was lower for *WeightMentor's* CUQ score than for CUQ scores of the other chatbot. This may be explained by the difference in sample sizes between the CUQ validation study (n=26) and usability tests (n=50) and also by the fact that participants in the CUQ validation study were repeating their assessment of each chatbot approximately two weeks after the initial assessment.

6.6 Study Limitations

There were specific limitations to this CUQ validation study. Firstly, the number of participants was quite small (n=26). The *WeightMentor* usability testing study (chapter

5), which was comparable in design and data analysis, included 50 participants. It may be argued however that in this CUQ validation study, statistical significance was already achieved with n=26 participants and would be retained with a greater number of participants. Other limitations concern the number and type of chatbots. Three chatbots were selected for this research study and categorised into "good", "average" and "poor" quality, and two of these were health related – *WoeBot* is for mental health, and *Weight Loss Bot* is for weight loss. Although differences in chatbot scores were significant across the three chatbots, it would have been useful to select multiple types of chatbot for each of the three categories, in order to compare differences in scores across different types of chatbot.

6.7 Future Work

Results from this preliminary validation of the CUQ suggests that the questionnaire demonstrates construct validity and test-retest reliability (also known as *intra-rater* reliability). However as discussed at the beginning of the chapter, there are several types of reliability and validity, summarised in **Table 6.18.** The CUQ will be more reliable as a validated instrument for measuring chatbot usability if other types of reliability and validity can be demonstrated. Inter-rater reliability of CUQ scores may be measured using Fleiss' Kappa, and internal consistency may be measured using Cronbach's Alpha.

Table 6.18: Types of Reliability and Validity

	Type	Description
Reliability	Test-Retest Reliability/ Intra-Rater Reliability	Consistency of scores over time/consistency of repeat scores from the same "rater" (participant)
	Internal Consistency	Strong correlation between items that measure the same construct
Validity	Inter-Rater Reliability	Consistency of scores from different "raters" (participants)
	Content Validity	Extent to which the items in the questionnaire relate to what is being measured
	Criterion-Related Validity/ Criterion Validity	Extent of correlation between a new (unvalidated) instrument and an existing validated instrument
	Construct Validity	Extent to which the questionnaire measures the "construct" (e.g. usability, personality) that it is supposed to be measuring

6.7.1 CUQ Benchmarking

SUS and UEQ have a clear advantage over the present CUQ in that their reliability and validity are based on *benchmarks*. The SUS average score of 68.0 (see **section 5.3.4.1**) is based on a benchmark of around 500 tests involving more than 5000 users (Sauro, 2011). The UEQ benchmark (**see section 5.3.4.2**) is based on data from 20,190 participants in 452 evaluations as of 2019 (Schrepp 2019b). As discussed in **section 6.2**, CUQ validity is currently limited by the number and type of chatbots used. Only three chatbots were used during the study, one for each quality category (good, average and poor). To improve the reliability and validity of the CUQ a benchmark could be established using many chatbots for each quality category, with different chatbot types included in each category. The CUQ validation study is also limited by the number of users (n=26) who completed the study, for an effective benchmark to be established it would be necessary to include larger numbers of participants in order to be comparable with SUS and UEQ.

6.8 Conclusion

Chatbot usability testing does not easily follow conventional principles, thus it is necessary to adapt conventional tools for use with chatbots. The CUQ validated in this study is based on the ALMA chatbot test tool (Martín et al. 2017) and designed to be comparable or equivalent to SUS (Brooke, 1996). Findings from this validation study suggest that the CUQ possesses construct validity and test-retest reliability in its native 16-item form but may also be reduced to a 10-item questionnaire akin to SUS without sacrificing validity, and thus may be deployed in chatbot usability testing scenarios either in conjunction with or as an alternative to SUS. Further validation would determine the extent of content and criterion validity and determine inter-rater validity. Establishment of a benchmark would improve the reliability and validity of the CUQ, and this may be achieved by conducting extensive tests of large numbers of chatbots using large numbers of participants. In comparison with the three chatbots in this validation study, *WeightMentor's* CUQ score places it between *WoeBot* and *Weight Loss Bot*, but closer to *WoeBot*, which suggests that it may be rated as "good" quality. As a validated instrument comparable in procedure to the most popular system usability metric (SUS), the CUQ will be highly desirable to researchers and business organisations wishing to conduct credible usability tests on chatbots.

Chapter 7: Discussion, Conclusion and Future Work

7.1 Introduction

This research focused on the use of technology to manage weight maintenance in the management and prevention of obesity. The aim of this PhD was to design and develop *WeightMentor*, a self-help personalised chatbot for weight loss maintenance. This aim was fulfilled through five main objectives, each of which progressed as its own individual research study.

During the first objective, the literature review, the key electronic databases *EMBASE, PubMed, and Medline* were searched for articles dated between 2006 and 2018 and relating to the general search terms "weight loss maintenance", "personalised messaging" and "digital technologies". Findings from this literature review are discussed in section 2.1. The purpose of this literature review was to identify gaps in the literature and identify opportunities for further study. Following the literature review, it was determined that a chatbot may be a suitable solution for motivational messaging during weight loss maintenance. This was in response to findings by Donaldson et al. (2014) and Fjeldsoe et al. (2016) who reported that text-message-based systems were effective for weight loss maintenance when used for self-reporting and motivational feedback, however it was suggested that text-message-based systems present limitations as they do not offer 24/7 availability since they require a human operator to read and respond to text messages (Donaldson et al. 2014). As automated conversation-driven systems, chatbots can personalise their interactions with individual users and are also available 24/7, thus providing a viable solution to the limitations identified by Donaldson et al. (2014).

In order to determine how to best serve the target demographic of individuals who are maintaining weight loss (or indeed who have previously lost weight but have since regained it), individual interviews were conducted with participants (n=15) from this demographic, who identified five key themes relating to challenges, social support, app use, personalised messaging and chatbots. Findings from these interviews are discussed in section 2.2. Findings from these interviews were used in conjunction with a recent review of popular nutrition apps (Franco et al. 2016) to inform development of the *WeightMentor* chatbot for weight loss maintenance. The *WeightMentor* chatbot was

designed using contemporary tools for chatbot development and its functions were carefully designed to provide an engaging, easy to use environment. Facebook messenger was selected as the user interface to provide users with a potentially familiar interface and to eliminate the need for installing and learning a new smartphone app.

The *WeightMentor* chatbot was usability tested using structured tests (discussed in section 2.4) in which participants (n=50) were video and audio recorded attempting a series of four chatbot tasks. Baki Kocabali et al. (2018) recommended the use of multiple metrics for chatbot usability testing to provide a more comprehensive picture of the chatbot user experience (UX), thus usability test participants completed three usability questionnaires. Two of these were existing instruments, the System Usability Scale (SUS) (Brooke 1996), currently the most popular usability metric, and the User Experience Questionnaire (UEQ) (Laugwitz et al. 2008), which is designed to provide a more thorough evaluation of UX. The third questionnaire, the Chatbot Usability Questionnaire (CUQ), was designed as part of this PhD specifically for assessing chatbot usability. This questionnaire was unvalidated during the *WeightMentor* usability tests but was validated in a later research study. Participants (n=26) in the CUQ validation study (discussed in section 2.5) evaluated three chatbots identified as “good”, “average” and “poor” quality and completed two rounds of CUQs for each one, with approximately two weeks delay between each round of questionnaires.

7.2 Review of literature related to technology for Weight Loss Maintenance (Objective 1, Chapter 2)

During the literature review a total of 61 research studies were selected for inclusion, seven of which were randomised controlled trials (RCTs), which were reviewed systematically and published in the European Journal of Public Health (Holmes et al. 2018, Holmes et al. 2019). The systematic literature review paper concluded that digital technologies are effective for weight loss maintenance in the short term- less than 24 months, but the longer-term effect requires further research. Thus, it is clear from previous research that digital technologies can aid weight loss maintenance up to 24 months. However, this finding is based on seven studies mainly in USA, UK and Europe. The technologies used were older tools such as text messaging, e-mail and websites, which are gradually being superseded by newer technologies such as social media and

chatbots (UKOM, 2018). The longer-term effectiveness is currently not known and further research is required.

From considering all relevant literature using all study designs, the conclusion supported the systematic literature review findings that current technologies for weight loss maintenance are mainly traditional tools with limited research using complementary technologies such as smartphone apps (Brindal et al. 2016) and chatbots (Fadil & Gabrielli, 2017). Digital technologies are constantly evolving. More technological innovations are required in the area of weight loss maintenance and these new technologies need to be researched to provide the evidence base.

Traditional technologies are useful for weight loss maintenance. Text-message-based systems are particularly effective because they are convenient and permit direct delivery of messages and information to recipients without needing to log in to a website or check an email. (Donaldson et al. 2014). Donaldson et al. (2014) and Fjeldsoe et al. (2016) reported that text-message-based systems were effective for facilitating self-reporting and motivational feedback. The reviewed literature also suggested that personalised communication leads to improved intervention outcomes (Fry & Neff, 2009) and reduced participant attrition (Fjeldsoe et al. 2009). Generic messages are widely reported to be less desirable for recipients, and are often ignored (Pollard et al. 2016, Smith et al. 2014). Opportunities for further research identified in the literature included investigating the effectiveness of weight loss maintenance technology interventions in the longer term (over 24 months) and investigating how chatbots may be used to effectively automate personalised messages and deliver weight loss maintenance interventions. Text-message-based systems present limitations as they do not offer 24/7 availability since they require a human operator to read and respond to text messages (Donaldson et al. 2014). Additionally, while it is theoretically possible to automate replies to text messages, doing so results in messages that are highly generic and seem machine-generated (Donaldson et al. 2013). As automated conversation-driven systems, chatbots can personalise their interactions with individual users and are also available 24/7, thus they provide a viable solution to these limitations. Text-messaging is also limited by the decline in its use in favour of newer technologies such as WhatsApp and Facebook Messenger. Although a tried and tested communication method that has existed for many years, recent statistics suggest that text-messaging use is declining steadily (Ofcom, 2018, UKOM, 2017) in favour of social media and platforms such as Skype and WhatsApp. As chatbots may be

easily integrated into platforms such as Facebook Messenger it is thus possible to present them as an attractive option for contemporary smartphone users. As obesity continues to increase worldwide to the point of being considered an epidemic (World Health Organisation (WHO), 2018) and as successful weight loss remains a challenge for many individuals (National Health Service (NHS), 2019), it makes sense to present these individuals with a source of support and motivation that is convenient, friendly and simple.

Robust clinical trials of technology for weight loss maintenance have been effective in the short term and used older forms of technology. These interventions have been shown to be less effective than face-to-face coaching techniques (Wing et al. 2006, Svetkey et al. 2008). It is recommended that health researchers test innovative technology interventions, particularly those which employ contemporary technologies. Future research should investigate the effectiveness of technology for weight loss maintenance in the long term (over 24 months), how chatbots may be used to effectively automate personalised feedback and messages and the extent to which chatbots may be used as a delivery method for weight loss maintenance interventions. Based on the findings of this review, the *WeightMentor* chatbot was designed to utilise contemporary technologies, support behaviour change through self-monitoring and motivation and provide personalised feedback in line with current wisdom regarding personalised message tones and types (Pollard et al. 2016, Smith et al. 2014).

7.3 Needs of individuals maintaining weight loss (Objective 2, Chapter 3)

The needs of individuals maintaining weight loss were obtained. A total of 15 adults participated in the needs analysis interviews, and the data were analysed using thematic analysis. From this analysis, five key themes were identified associated with weight loss maintenance and the use of technology, these were, “Weight loss maintenance is a challenge”, “Social contact can be a double-edged sword”, “Apps are popular”, “Personalised messages are more useful” and “Chatbots have potential for weight loss maintenance”. In key theme 1, “*Weight loss maintenance is a challenge*”, participants acknowledged that weight loss and maintenance were challenging, but generally seemed capable of identifying strategies and solutions that worked for them, such as avoiding snack foods (especially when tired or stressed), or taking regular exercise, whether on

their own or as part of a group such as Parkrun. These findings are consistent with research by Wing et al. (2006), who propose that maintenance is the greatest challenge to obesity management. Wing & Hill, (2001) suggested that regular self-monitoring, high levels of physical activity and a reduction in high-fat foods all contribute positively to successful weight loss maintenance, findings which are echoed by needs analysis participants. Key theme 2, "*Social contact can be a double-edged sword*" discussed the importance of social contact, but also looked at how friends and family could encourage unhealthy habits unless they were sympathetic to the individual who was trying to maintain weight loss. The positive influence of social support is echoed by Cussler et al. (2008), who reported no significant difference in weight gain between control and intervention groups in the *Healthy Weight for Life* study. The authors suggest that this resulted from control group participants organising their own support groups during the intervention to practice techniques they had learnt. Participants discussed the use of smartphone apps in key theme 3, "*Apps are popular*" and suggested that convenience was important, as was the ability to record and track progress. Previous research has suggested that self-reporting and monitoring can lead to positive outcomes (Donaldson et al. 2014, Fjeldsoe et al. 2016, Korotitsch & Nelson-Grey, 1999) and improve self-awareness (Yeager et al. 2008) and findings from key theme 3 are consistent with this. Key theme 4, "*Personalised messages are more useful*", discussed participant attitudes to messages, and how they should be personalised, if at all. Findings in this theme were generally consistent with work by Pollard et al. 2016, Smith et al. 2014, Stephens et al. 2015 and Woolford et al. 2011, which suggested that "generic" messages tend to be ignored and that personally relevant messages are more useful. Smith et al. (2014) reported that adolescent participants considered overly formal messages to be "nagging", a sentiment that seemed to be echoed by needs analysis participants, who stated that personal messages should be subtle and not "bossy". In key theme 5, "*Chatbots have potential for weight loss maintenance*", participants generally reacted positively to the prototype *WeightMentor* chatbot, reporting that the chatbot was engaging, natural and not too generic. It was concluded by participants that with its range of functions and personality, *WeightMentor* would be a useful weight loss maintenance tool provided it required only minimal interactions and users were not expected to use it daily. These findings are consistent with what has been reported by Fadil & Gabrielli (2017) that chatbots can improve participant engagement with interventions.

A motivational chatbot could potentially be a viable solution for assisting individuals with weight loss maintenance. It was determined during needs analysis that a weight loss maintenance chatbot should support users at times when they are most vulnerable (e.g. tired or stressed) and thus should be available 24/7, should permit progress tracking with appropriate praise or encouragement, should conveniently fit into the user's daily life with minimal effort required for interaction, and should provide feedback that is subtle, positive and encouraging without sounding patronising or bossy.

7.4 Design and development of a chatbot for Weight Loss Maintenance, WeightMentor (Objective 3, Chapter 4)

Chatbots differ from conventional graphical user interface (GUI) systems in functionality and interface, thus it is not always practical to apply conventional design principles to their development (Cameron et al. 2018a). During the *WeightMentor* chatbot design process, conversation analysis studies (Moore et al. 2017) were used to develop a set of chatbot design principles, based on Shneiderman's Eight Golden Rules and Nielsen's Ten Usability Heuristics (both discussed in chapter 4). The *WeightMentor* chatbot was designed to offer functions that would be most useful to users, based on findings from the Needs Analysis study (chapter 5), the Literature Review (chapter 2) and a recent review of popular nutrition apps (Franco et al. 2016). Self-reporting, progress tracking and motivation were selected as the main functions as these have all been shown to benefit weight loss maintenance (Korotitsch & Nelson-Grey, 1999, Donaldson et al. 2014, Fjeldsoe et al. 2016). User profile creation was included to permit the chatbot to remember the user's name and self-reported data. The chatbot integrated with Facebook messenger to provide a simple, familiar user interface based on one of the most popular social media messaging tools (UKOM 2018, Ofcom 2018). The basic conversation flow was designed using DialogFlow and additional functionality was implemented through an application developed in NodeJS, permitting storage of data (user profile information and self-reporting history) in a database, variability of chatbot responses, motivational feedback and animated GIF images which are used as part of welcome and goodbye messages.

WeightMentor's primary functions will potentially be useful for individuals who are maintaining weight loss, and the chatbot will also stimulate user interest and potentially

maintain engagement using varied motivational feedback and conversational responses. By varying and personalising its messages to the user, *WeightMentor* potentially solves the problems of automated text messaging – lack of personalisation, limited empathy, and sounding too robotic (Donaldson et al. 2014). *WeightMentor* will permit users to be open about their diet and physical activity habits without fear of being judged or talking to another human.

7.5 Usability of the *WeightMentor* chatbot for Weight Loss Maintenance (Objective 4, Chapter 5)

The *WeightMentor* chatbot was usability tested using structured tests with 50 participants. The chatbot generally scored highly across the three usability questionnaires. *WeightMentor's* mean SUS score was 78.0±16.2, 10 points above the SUS benchmark score of 68.0, and UEQ scores for each of the six scales were all above +0.8, which represents a positive evaluation. The CUQ score for *WeightMentor* was 72.2±13.9, but the CUQ was not validated at the time of usability tests. These scores suggest that *WeightMentor* is more highly usable than most conventional systems, although it is unknown if any chatbots were used for SUS and UEQ benchmarking and not possible therefore to compare *WeightMentor* to other chatbots. These scores are consistent with recent work which suggests that chatbots are generally received positively (Budiu et al. 2018, Fitzpatrick et al. 2017). The chatbot generally seemed to exceed participants' expectations as post-task Single Ease Question (SEQ) scores per task were higher (suggesting an easier task) than pre-task SEQ scores. It may be determined that participants perceived *WeightMentor* tasks as easier than traditional GUI-based tasks, as mean scores were higher than the benchmark SEQ score of 5.3 – 5.5 (Sauro, 2012). Time taken to complete repeated tasks decreased with each repetition as expected, however it was determined that optimum performance at tasks was reached after just one repetition of the task, as differences in time between repetitions were not significant. Analysis of usability issues identified by participants during best case, worst case and chronological testing scenarios suggested that 39 users were required to identify most of the *WeightMentor* usability issues, a finding that challenges work by Nielsen & Landauer (1993) which suggests a maximum of 5 – 8 users for identifying 80% of usability issues (Nielsen & Landauer, 1993).

Findings from the *WeightMentor* usability tests suggests that chatbots are perceived as more usable than conventional systems, that users may become proficient in tasks after just one attempt, and perhaps most significantly, that it is more difficult to identify usability issues in chatbots than in conventional systems.

7.6 Validation of the Chatbot Usability Questionnaire (CUQ) for measuring chatbot usability (Objective 5, Chapter 6)

The CUQ was validated to measure construct validity and test-retest reliability (also known as intra-rater reliability) and factor analysis was used to identify the factors within the CUQ. It was determined that the construct validity of the CUQ should be measured because "usability" is a highly subjective concept, based on the opinions of individual users. Preedy & Watson (2009) suggested that intangible concepts (such as usability) are difficult to measure, yet often need to be measured, thus it was considered a priority for the CUQ to demonstrate construct validity. Three chatbots were identified by an expert panel and categorised as "good", "average" or "poor" quality, and 26 participants conducted two evaluations of each chatbot (within approximately two weeks of each other) and completed a CUQ for each chatbot at each evaluation. Analysis of results suggested that test-retest reliability and construct validity exist in the CUQ at both questionnaire and question level. Correlation between CUQ scores at each round of questionnaires was strong, and differences between scores were not significant (suggesting the presence of test-retest reliability at the questionnaire level) and differences between mean question scores at each round were not significant (suggesting the presence of test-retest reliability at the question level). Differences between mean CUQ scores for each chatbot were significant (suggesting construct validity at the questionnaire level), and differences between mean individual question scores for each chatbot were generally also significant (suggesting construct validity at the question level). Factor analysis identified four factors within the CUQ, and analysis of correlations between questions and factors suggested that the CUQ could potentially be reduced in length from 16 items to 12 or 10 items. Both the proposed 12- and 10-item CUQs demonstrated construct validity.

Findings from the CUQ validation study suggest that the CUQ demonstrates construct validity and test-retest reliability and that the questionnaire may be reduced to 10 items

without sacrificing construct validity, which would create a usability instrument that is akin to SUS, which is also a 10-item questionnaire (Brooke 1996).

7.7 Research Limitations

All research has limitations, and there were a few within this PhD. One limitation was the selection of Facebook Messenger for the *WeightMentor* user interface during the design and development phase. It was suggested by usability testing participants that it could be useful to provide an app version of the chatbot as not everyone uses Facebook messenger and an app version would make the chatbot available to a greater number of people. While it may be the case that not everyone uses Facebook Messenger, it does remain the most popular social media platform in the UK (Ofcom, 2018, UKOM, 2017) and is a familiar user interface for those who do use it. However, it is certainly worth considering implementing an app version of the *WeightMentor* chatbot to extend its availability beyond users of Facebook Messenger.

A second limitation concerned the number of participants in the CUQ validation study (as discussed in chapter 6). The study included data from 26 participants, however the *WeightMentor* usability tests, which were comparable in design and data analysis, included 50 participants. It is worth arguing however that the required statistical significance of results was achieved at 26 participants, and there may be limited value to increasing this number when statistical significance has already been achieved. Other limitations within the CUQ validation study included the number and type of chatbots (as discussed in chapter 6). Two of the three chatbots evaluated in the study were health related, and it would perhaps be of use to select multiple types of chatbot from each category in order to compare scores across different chatbot types.

Purposive sampling was used to recruit participants for the Needs Analysis interviews, and convenience sampling was used during recruitment for Usability Testing and CUQ Validation studies. Dudovskiy (2018) suggested that purposive sampling and convenience sampling may present limitations such as reduced reliability and the potential for increased bias. It is also suggested that purposive sampling is vulnerable to errors in judgement and given that only a very specific sample of participants were selected from the population it is difficult to generalise findings in relation to the whole population (Dudovskiy, 2018). However, these two sampling types were determined to

be a cost-effective and time-effective means of quickly recruiting participants, and the risk of bias associated with each type was mitigated through the use of carefully determined selection criteria which were strictly adhered to when selecting participants.

7.8 Relevance of the *WeightMentor* chatbot to the *Transtheoretical Model of Health Behaviour Change*

It is important that all research is underpinned with theory to ensure that decisions made, and actions taken can be justified and are contributing to the knowledge base. This research was underpinned by the Transtheoretical Model of Health Behaviour Change (Prochaska & Velicer, 1997). As discussed in chapter 2, section 3.1, the Transtheoretical Model of Health Behaviour Change is a seven-stage model based on work by Albert Bandura (Prochaska & Velicer, 1997). The model suggests that behaviour change begins with precontemplation, (where the individual is uninterested in behaviour change) and progresses through preparation and action (where positive behaviour changes are actively implemented, to maintenance, where the individual reinforces the positive behaviours learnt during the action stage. Maintenance may be followed either by termination, where the individual achieves self-efficacy, or relapse, where the individual returns to their old habits and either begins the cycle again or abandons it. The individual is most vulnerable during the maintenance stage, which it is suggested may last for up to five years (Prochaska & Velicer, 1997).

As has been previously discussed, the *WeightMentor* chatbot will support individuals during their vulnerable maintenance stage by reinforcing six of the ten behaviour change processes (Prochaska & Velicer, 1997). The main *WeightMentor* functions have been carefully selected to facilitate this phase. *Consciousness raising* will be reinforced using the self-reporting functionality which, if used regularly will place a user's diet and physical activity in the forefront of their mind, raising their awareness of unhealthy habits and hopefully encouraging them to think about causes of unhealthy behaviours and ways they can prevent these. The personalised feedback functionality, which provides positive feedback in response to positive behaviours and encouragement in response to negative behaviours will help reinforce *self-re-evaluation* and *self-liberation*, as the user comes to realise that failure is not 'the end of the world', and that positive behaviours will be rewarded, and re-evaluates their self-image accordingly. This will also reinforce *contingency management*, as the user begins to associate positive behaviours with

positive feedback. Appropriate feedback will also stimulate *counter conditioning* by providing users with practical, easy to achieve strategies for reinforcing positive behaviour change (such as doing a small amount of exercise every day or eating a few small healthy snacks if they feel hungry during the day). Motivation elements of the chatbot will reinforce *stimulus control* any time the user calls on the motivation function, and as the *WeightMentor* chatbot is available 24/7, users may do so at times when they are most likely to give in to temptation. Motivation also influences *self-liberation*. This research provides another example how the Transtheoretical Model of Health Behaviour (Prochaska & Velicer, 1997) can be applied in practice, i.e. weight loss maintenance.

7.9 Recommended Future Work

7.9.1 Extended Chatbot Content Analysis

A preliminary analysis of chatbot content was conducted in summer 2019. The *WeightMentor* chatbot was compared with six other chatbots related to the general area of health. Chatbots were analysed for purpose, user interface platform, response generation, dialogue initiative, input modality, target, personality/role, gender, use of multimedia, free text entry, reading age, introduction, range of functions, motivation, use of external resources, use of social support and ratio of text to multimedia per 20 messages (based on the first 20 messages). Findings from the content analysis are discussed in **chapter 4, section 4**. This content analysis study could be developed further by comparing *WeightMentor* to other types of chatbot than those related to healthcare, and by conducting more in-depth analysis into types of motivation used, conversation flows and conversation type analysis (based on Moore et al. 2017).

7.9.2 Feasibility of the *WeightMentor* chatbot for Weight Loss Maintenance

The next stage of this research is to determine the feasibility of the *WeightMentor* chatbot for weight loss maintenance. A proposed future study would assess the feasibility of delivering weight loss maintenance interventions using the *WeightMentor* chatbot. This proposed study would run for four weeks and would determine if the intervention is workable, evaluate the recruitment, participant engagement (or non-engagement) and attrition rates, report on technical difficulties and provide participant feedback on the chatbot. Although three months is a common duration for intervention studies

(Donaldson et al. 2014), recent successful studies have been conducted in four weeks, such as the Starlight trial (Volkova et al. 2014) and the *WoeBot* randomised controlled trial (Fitzpatrick et al. 2017). As no definitive results will be collected, four weeks should be long enough to gather enough data to plan further trials.

Participants could be adults (aged 18+ years) who are willing to lose weight or maintain weight loss. No specific timescale or amount of weight loss would be applied to the criteria as the chatbot is available to anyone who has lost weight, and this study would determine the feasibility of using *WeightMentor* on a day-to-day basis. Participant selection criteria are summarised in **Table 7.1**. A minimum of 50 participants would be recruited for the feasibility study in order to account for approximately 10% attrition by study completion.

Table 7.1: Participant selection criteria for the proposed *WeightMentor* feasibility study

Inclusion Criteria	Exclusion Criteria
Aged 18+ years	Not providing informed consent
Willing to lose weight or maintain weight loss	Not resident in Northern Ireland for the duration of the study
Willing to use the chatbot, WeightMentor	
Own a smartphone (e.g. Android device or iPhone)	
Able to attend three in-person sessions at the start of the study, after two weeks, and at the conclusion of the study	

Participants could be recruited using convenience sampling from staff and students at the Ulster University, through local weight management groups such as Unislim Northern Ireland and through local community groups (e.g. churches). Recruitment emails could be circulated to staff and students at the Ulster University via the mailing list and to local weight management groups. Recruitment posters and flyers could also be distributed to community groups. All recruitment materials would invite interested individuals to contact the researcher via supplied contact details to request a copy of the Participant Information Sheet (PIS). Interested individuals would be screened based on the selection criteria and selected participants would provide written informed consent prior to baseline measurements during the initial in-person session.

Multiple in-person sessions could be conducted based on participant locations and would take place either on one of the university campuses (for staff and students of the university

or for community group participants) or in normal meeting places of local weight management groups. During the initial (baseline) session, the researcher would introduce himself and welcome the participants, who would all be given an opportunity to read the PIS again if they needed to do so. The researcher would explain the purpose of the study and demonstrate the *WeightMentor* chatbot, answer any initial questions and then assist participants with setting up their own mobile devices to use the *WeightMentor* chatbot and then create their user profiles. Participants' height and baseline weight would be recorded to calculate baseline BMI, and they would also be asked to rate their current motivation for losing weight, from 1 (Extremely low) to 5 (Totally motivated). Participants would complete a demographic fact sheet during this session. Following the baseline session, participants would use the chatbot for motivation as required and would also self-report their perceived levels of physical activity and food intake, ideally daily, but for a minimum of three times a week.

An in-person follow-up session could take place two weeks after baseline. As for the initial session, multiple follow-ups would be conducted based on participant locations, or could be conducted via telephone if participants are unable to attend in person. During this session, the researcher would record each participant's weight, which would be used to calculate BMI. Participants would be able to ask questions and get assistance with any technical issues from the researcher. Participants would also complete a short evaluation questionnaire.

A final in-person follow-up session could be conducted four weeks after baseline, either in one of the four university campuses or in normal locations of local weight management group meetings. The researcher would record each participant's final weight, which would be used to calculate final BMI, and the participants would complete a final study evaluation questionnaire.

Study procedures are illustrated in **Figure 7.1**.

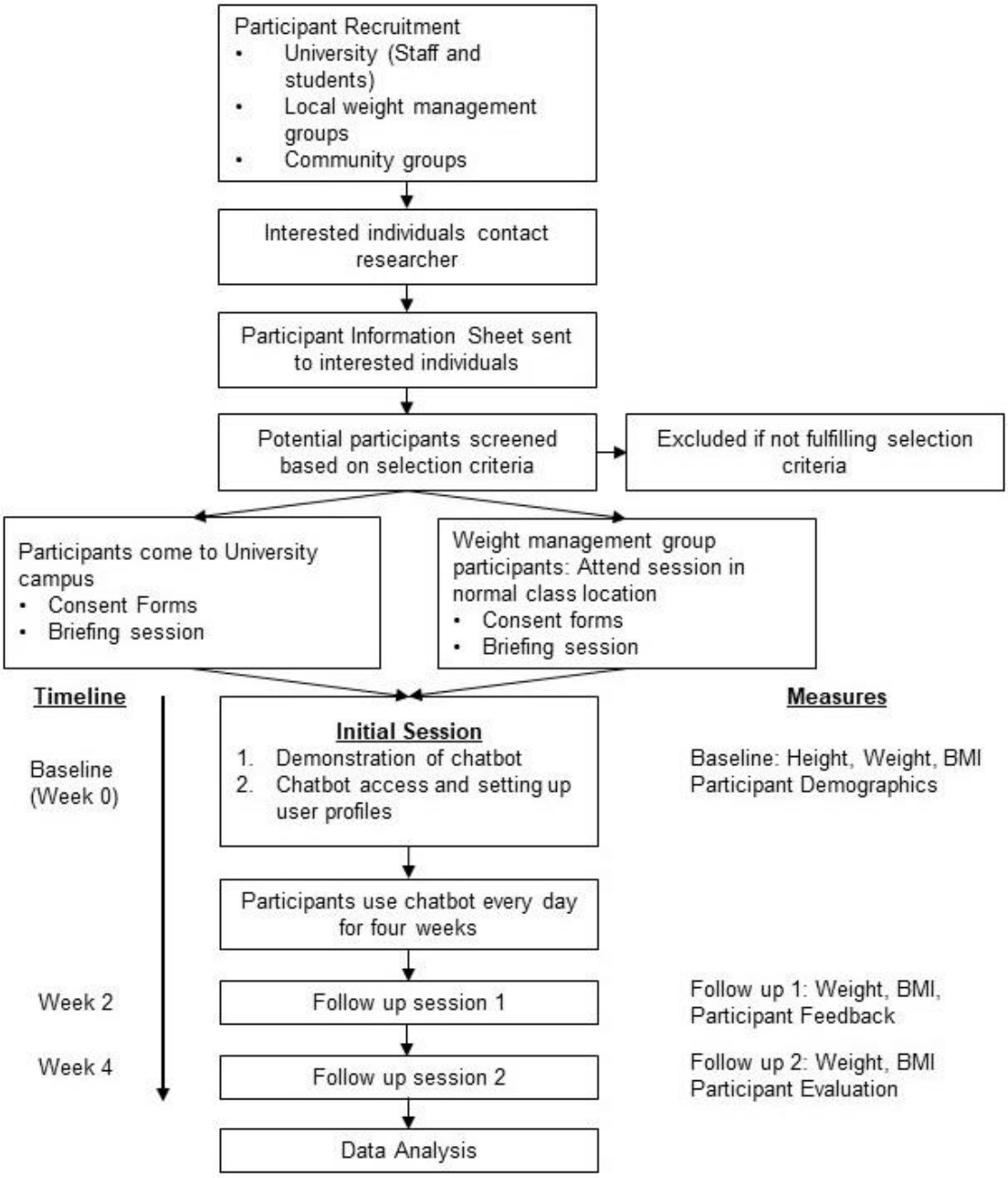


Figure 7.1: Feasibility study procedures

Weight maintenance, the primary outcome measure, would be measured based on percentage body weight gained or lost during the study, at 1, 2, and 4 weeks. Weight, BMI, participants' evaluation of the study, and chatbot usage statistics could be secondary outcome measures. BMI and weight are common outcome measures for similar interventions (Donaldson et al. 2014, Fjeldsoe et al. 2016).

Weight would be measured using a calibrated scale. Height would be measured once using a stadiometer. Chatbot usage statistics would be logged by *WeightMentor* in an internal database and extracted at the end of the study. Participant demographics (user ID, age range, gender), self-reported food intake and physical activity would all be

recorded by the chatbot. A short feedback questionnaire could be used during follow-up session one to record participants' experiences of using WeightMentor and a questionnaire evaluating both the chatbot and the study could be used during follow-up session 2.

This project may be considered low risk; there could be a slight potential risk to participant wellbeing as some individuals may be sensitive about their weight, and thus could be reluctant to submit for weight checks during the study. To mitigate this and reduce potential distress to participants, height and weight would be checked individually in a private room and weight records be treated with complete confidentiality. Participants would be informed on the PIS that the weight checks would be conducted by a male researcher, but they could be given the option to request a female researcher to conduct weight checks if they wish. Weight loss group participants would be able to request that a representative from their group conduct their weight checks. Participants may be inconvenienced by the need to attend the three in-person sessions, however this inconvenience could be mitigated by conducting the sessions at a location that is convenient for participants, such as their nearest campus of the Ulster University or their regular weight management club or community group meeting place. Participants who were genuinely unable to attend the first follow-up session (at two weeks after baseline) could request a follow-up by telephone.

7.10 Contributions to Literature

This PhD makes several key contributions to literature. This research has determined that there is a clear need for longer-term research into the use of technological solutions for weight loss maintenance, particularly those which use newer technologies such as chatbots. Based on the needs analysis, it has been determined that chatbots have potential for weight loss maintenance, because they are perceived as engaging, friendly and simple to use. Simple interventions, such as those using text messaging have been shown to be the most effective, and it is suggested that the success of these interventions is because of the convenience and ease of use of the system and the ability to access motivational materials without needing to log into a website or download an app (Donaldson et al. 2014).

Usability testing results confirmed that chatbots present much higher usability than conventional systems, suggested that users generally perceive chatbot-based tasks as easier than conventional tasks, and indicated that chatbot users may become proficient in their use after just one attempt at a task. It is worth noting however that findings from these usability tests have challenged conventional wisdom, as proposed by Nielsen & Landauer (1994) that no more than 5 – 8 users are required to identify 80% of usability issues. During the *WeightMentor* usability tests it was determined that 39 users were required to identify most of the usability issues, and the implication of this, with regard to Nielsen & Landauer's work, is that the nature of chatbots makes it potentially more difficult to identify usability issues during testing and thus, this conventional wisdom does not easily apply.

Findings from the CUQ validation study suggests that the questionnaire demonstrates construct validity and intra-rater reliability, and that construct validity is preserved when the questionnaire is scaled down to ten items. As a ten-item questionnaire the CUQ will be equivalent to the SUS and may therefore be used in chatbot usability testing scenarios where it is desirable to have a score that may be reliably compared with SUS.

7.10.1 Implications for Weight Loss Maintenance

Based on this research project, there are implications for weight loss maintenance, which are as follows:

1. For individuals to consider using chatbots such as WeightMentor for weight loss maintenance

Findings from recent weight loss maintenance interventions (Donaldson et al. 2014, Fjeldsoe et al. 2016) have suggested that self-reporting and appropriate, personalised feedback are beneficial for weight loss maintenance. These findings were reinforced from the needs analysis **(chapter 3)** that suggested chatbots are viewed positively by individuals who are maintaining weight loss because they are engaging and can provide appropriate personalised feedback. Chatbots have been shown to improve participant engagement in therapeutic interventions (Barak et al. 2009) and are suited to this type of intervention because they offer improved usability over conventional systems (Fadil & Gabrielli, 2017). It is important for individuals to consider chatbots for weight loss maintenance because they may potentially provide useful and important features without

compromising on usability or convenience. The needs analysis interviews revealed that convenience is an important consideration when choosing apps, and that chatbots may be convenient if their interactions are minimal and they fit easily into users' daily lives. A chatbot such as *WeightMentor* has the potential to improve recording, monitoring and communication of food intake and physical activity for effective weight loss maintenance. Using a chatbot with a very simple, conversation-driven user interface simplifies this process and improves convenience. Although the *WeightMentor* chatbot merely uses submitted data to provide the user with personalised feedback, a future development could provide the user with the option of connecting the chatbot to their social media accounts, making it possible to share progress. This would introduce a social contact element, which could benefit the user as their friends and family could provide positive encouragement on their progress. Additionally, data could be shared (with the user's consent) with health professionals such as the user's General Practitioner, who could provide guidance and support based on their self-reporting history.

2. For health professionals to make their patients aware of chatbots such as WeightMentor for weight loss maintenance

This present research clearly shows that chatbots have potential to aid weight loss maintenance. This research supports previous work from the *Weight Care Project* (Moorhead et al. 2013) that suggested that although effective communication is crucial to successful weight loss maintenance, the effectiveness of communication is reduced by confusing and often conflicting nutrition information and advice and by a reluctance among healthcare professionals to engage in frank and open discussion about obesity with their patients (Moorhead et al. 2013). As has been suggested by findings from the *WeightMentor* needs analysis, individuals who are investigating tools for weight loss maintenance will be keen to explore those which are convenient and have been shown to be effective for weight loss maintenance. Given the sheer number of smartphone-based tools which potentially promise more than they can realistically deliver, it is important for individuals to be able to identify those which have been demonstrated to be effective and those which are worth ignoring. Health professionals who have earned the trust of their patients will be a useful resource in recommending evidence-based chatbots for weight loss maintenance. As discussed in implication #1 above, a chatbot may potentially improve the monitoring, recording and communication of food intake and physical activity data, and will simplify the process of doing so. Health professionals may be keen to recommend or promote evidence-based chatbots for this purpose.

3. For health professionals to consider chatbots such as WeightMentor for weight loss maintenance in support as part of management and prevention strategies

Recent research has shown that technology-based weight loss maintenance tools improve weight loss maintenance outcomes in the short term (Donaldson et al. 2014, Fjeldsoe et al. 2016, Thomas et al. 2012), and that specifically chatbots improve participant engagement (which is itself linked with intervention success) (Barak et al. 2009). The *WeightMentor* needs analysis suggested that chatbots may be an engaging and convenient means of delivering weight loss and maintenance interventions, and usability testing results showed that the chatbot is highly usable and easy to master. It is likely that individuals who are considering chatbots for weight loss maintenance will be attracted to high usability and simplicity. It is important for health professionals to consider recommending chatbots either as a primary means of managing weight loss maintenance or as an adjunct to conventional means in order to maximise intervention effectiveness and participant engagement while reducing costs.

7.10.2 Implications for Computer Science

Based on this research project, there are implications for computer science, which are as follows:

1. To consider using the validated Chatbot Usability Questionnaire as an alternative or in addition to traditional metrics

Chatbots are conversation-driven interfaces (Knowledge@Wharton, 2016) and do not easily conform to traditional design principles (Cameron et al. 2018). Similarly, conventional usability metrics such as the System Usability Scale (SUS) (Brooke, 1996) may not be best suited to chatbot usability testing as it is not known how many chatbots (if any) were included in benchmarking and so these metrics may not accurately reflect chatbot usability. Baki Kocaballi et al. (2018) suggested that multiple metrics should be used for chatbot usability testing in order to provide a comprehensive picture. It is important to use tools that will adequately measure chatbot usability and will provide a means of comparison with both other chatbots and conventional systems of similar function. A validated questionnaire that has been designed specifically to measure chatbot usability will provide a useful comparison with other metrics in addition to a tangible measure of chatbot usability. The new Chatbot Usability Questionnaire,

designed as part of this PhD, has been shown during preliminary validation to possess construct validity and test-retest reliability at questionnaire and question level, even when reduced in size to ten items (comparable in size to the SUS). This questionnaire has been shown to measure aspects of usability that are particularly relevant to chatbots, and so may be an effective tool for measuring chatbot usability, either on its own or in combination with other validated usability questionnaires such as SUS or the User Experience Questionnaire (UEQ).

2. To plan to use a minimum of 26 participants for chatbot usability testing

The current literature states that conventional wisdom argues that a maximum of 5 to 8 users will identify 80% of usability issues and thus it is unnecessary to conduct usability tests with more than this with many users (Nielsen & Landauer 1993). However, during usability testing of the *WeightMentor* chatbot it was determined that most usability issues will be identified by 39 participants, and this is most likely the case because chatbots are by their very nature easier to use than conventional systems (Knowledge@Wharton, 2016) and thus it may be more difficult to identify usability issues with a small number of users. It is important for chatbot testers to plan as many tests as possible (ideally a minimum of 39) in order to ensure that most usability issues are identified during chatbot usability testing.

3. To consider deploying chatbots in circumstances where ease of use or quick proficiency are priorities

During the *WeightMentor* usability testing no significant difference was observed in mean completion time between attempts 2 to 4 of task 2, or between any attempts at task 3. Task 3 (self-reporting food intake) was very similar in procedure to task 2 (self-reporting physical activity, and the implication is that *WeightMentor* users will become proficient in its use after only one attempt at a task. These findings are consistent with current wisdom, which suggests that chatbots are easier to use than conventional systems (Knowledge@Wharton 2016) thus it is reasonable to expect that users should quickly become proficient in chatbot use. It is important to consider using chatbots where a simple interface that is easy to learn and to master is required, and where there is limited scope for providing user training.

7.10.3 Implications for Policy

Based on this research project, the implication for policy is as follows:

1. To encourage the use of chatbots such as WeightMentor for weight loss maintenance as part of prevention and management strategies for weight management and obesity, particularly as an element of social prescribing-based strategies.

As discussed in implication 3 for weight loss maintenance (**section 7.10.1**), recent research has shown that technology-based weight loss maintenance present short-term benefits for weight loss maintenance (Donaldson et al. 2014, Fjeldsoe et al. 2016, Thomas et al. 2012), and that participant engagement (which is itself linked with intervention success) is improved during interventions where chatbots are used (Barak et al. 2009). It is important for chatbots to be considered as part of weight loss prevention and management strategies as they have the potential to improve intervention success while reducing costs. As discussed in **section 7.10.1** chatbots such as *WeightMentor* have potential to simplify and improve the process of recording, monitoring and communication of food intake and physical activity data and may also permit social sharing of these data, permitting health professionals or the user's friends and family to be involved in the weight loss maintenance process by providing support and encouragement. Thus, chatbots may be of relevance or interest to health professionals seeking to recommend tools for weight loss maintenance.

In recent years, the UK's National Health Service has moved towards the use of *social prescribing*, which enables healthcare professionals to refer patients to non-clinical services (Kings Fund, 2017), and a chatbot such as *WeightMentor* may potentially form part of a social prescribing-based strategy for weight loss maintenance. Health professionals may be reluctant to recommend chatbots for weight loss maintenance unless they met certain criteria. Any chatbot would need to be evidence-based, supported by a body of strong research from feasibility and pilot studies and a Randomised Controlled Trial. If a weight loss maintenance chatbot offered nutrition or health advice would need to ensure that the advice being offered was based on evidence or official guidance from health agencies such as the UK's National Health Service or World Health Organisation. However, as discussed in **section 2.6.8.4.1** of chapter 2, Natural Language Agents (NLAs) such as Siri, Alexa and Google Assistant may not be suitable for providing medical advice without input from a medical professional (Bickmore et al. 2018), thus

health agencies may prefer chatbots which do not specifically offer advice and instead provide only positive encouragement (such as *WeightMentor*).

2. To promote the use of chatbots as a means of encouraging individuals to take a more active role in maintaining their own positive health and wellbeing.

The King's Fund suggest that *patient activation* is linked with positive health maintenance behaviours (Hibbard & Gilburt, 2014). Patient activation is described as the capability of an individual to actively manage their own health and it is suggested that individuals who have low levels of patient activation are less active in maintaining their own health, often ignore warning signs and medical advice, and tend to avoid thinking about their health. Conversely, highly activated individuals actively engage in positive health behaviours and are less likely to require hospital treatment.

Patient activation may be measured using the Patient Activation Measure (PAM), which has been shown to be valid and reliable. Research has shown that PAM scores can predict health behaviours. Given the link between patient activation and positive health behaviour it is important for health policymakers to stimulate patient activation through the promotion of positive health behaviours and raised awareness of health. As personal digital technologies such as smartphones are popular and chatbots have potential for supporting weight loss maintenance it is logical to recommend these tools for positive health behaviour promotion.

7.10.4 Implications for Research

1. To make use of interdisciplinary teams during obesity research or research of technology for weight loss maintenance

This PhD was an interdisciplinary research project supervised by senior academics from a range of disciplines at the Ulster University. Dr Anne Moorhead, the lead supervisor, is a researcher in the School of Communication & Media and Institute of Nursing & Health Research. She is an expert in the field of Health Communication and has a background in nutrition and psychology. Dr Raymond Bond is a researcher in the School of Computing, with expertise in the field of Conversational Artificial Intelligence (AI) Data Analytics, and Health Informatics. Professor Huiru Zheng from the School of Computing has expertise in Health Informatics and Data Mining. Professor Vivien Coates from the School of Nursing and Institute of Nursing and Health Research has

expertise in the field of Nursing Practice Research and planning/designing clinical trials. In addition to the expert supervision team, the research was supported by Emeritus Professor Michael McTear from the Ulster University, who is a leading expert in chatbot design and research. Interdisciplinary research is beneficial for healthcare because the changing face of human health presents new challenges that often require a multi-faceted approach that reaches beyond the scope of traditional medicine (Kivits et al. 2019). Additionally, new challenges to human health (such as health inequality and rising obesity) may not be as easily solved by the health sector on its own and may require intervention from social and technological research sectors (Kivits et al. 2019).

The *WeightMentor* chatbot was an interdisciplinary research project, drawing on knowledge from the Health Communication, Computer Science, and Nursing research areas. It was necessary to adopt this interdisciplinary approach because the issue of weight loss maintenance and the difficulties associated with it require a variety of strategies to effectively solve. It was determined that a technological solution should be adopted – i.e. a chatbot – thus it was necessary to involve experts in the Computer Science field, who understood the nature of chatbots and how they may be developed. A core part of this intervention is Health Communication, as the *WeightMentor* chatbot collects data from users and provides feedback based on these data, thus it was important to work with experts who understood Health Communication and how messages may be presented to users most efficiently. It was necessary to involve experts from the Nursing research field as most of the research studies involved human participants, and it was important to work with experts who understood the ethical considerations associated with human-based research and who could provide insight into the most appropriate way to design and deliver human-based research studies.

7.10.5 Research Outputs

Work from this PhD was published in one scientific journal and presented in seven conferences. These are summarised in **Table 7.2**. Conference abstracts are in **Appendix 31**.

Table 7.2: Publications and Conferences

Title	Type	Publication/Conference	Date/Issue/Location
Impact of digital technologies for communicating messages on weight loss maintenance: a systematic literature review	Journal	European Journal of Public Health	Volume 29, Issue 2, April 2019.
Digital communication tools in support of weight loss maintenance: A systematic literature review	Conference	3rd Annual Public Health PhD Symposium	Liverpool, UK, 13th – 14th July 2017
Digital technologies for communicating messages on weight loss maintenance: A systematic literature review	Conference	EACH (International Association for Communication in Healthcare) Summer Event	London, UK, 4th – 6th September 2017
WeightMentor: A new automated chatbot for weight loss maintenance	Conference	32nd Human Computing Interaction (HCI) conference	Belfast, N. Ireland, UK, 2nd – 6th July 2018
Digital technologies for communicating on weight loss maintenance: A systematic literature review	Conference	EACH (International Association for Communication in Healthcare) International Conference on Communication in Healthcare (ICCH)	Porto, Portugal, 1st – 4th September 2018
Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?	Conference	ECCE (European Conference on Cognitive Ergonomics)	Belfast, N. Ireland, UK, 10th - 13th September 2019
Communication tool for weight loss maintenance: Chatbot, WeightMentor - needs analysis & development	Conference	EACH (International Association for Communication in Healthcare) Forum on Healthcare Communication	Leiden, Netherlands, 16th - 18th September 2019
WeightMentor: A bespoke chatbot for Weight Loss Maintenance	Conference	13th European Nutrition Conference, Federation of European Nutrition Societies (FENS)	Dublin, Ireland, 15th - 18th October 2019

7.10.6 Future Research Papers

Proposed future publications are summarised in **Table 7.3**. In total, this PhD has the data for 4 future research papers.

Table 7.3: Proposed Research Papers

Working Title	Type	Publication/Conference	Date/Issue/Location
WeightMentor, bespoke chatbot for weight loss maintenance: Needs assessment & Development	Conference	ACBH 2019: 4th Workshop on Affective Computing in Biomedicine and Healthcare	November 2019
Does user technical ability influence chatbot usability testing results?	Conference	32nd International BCS Human-Computer Interaction Conference	July 2020
Validating the CUQ: a bespoke questionnaire for chatbot usability measurement	Journal	Behaviour & Information Technology	Winter 2019
Designing *WeightMentor*, a chatbot for weight loss maintenance	Journal	International Journal of Human Computer Studies	Spring 2020

7.11 Conclusion

Although health technology is popular and its role in weight loss is a popular contemporary research topic, there is very limited research into the use of technology for weight loss maintenance. Current research has indicated that technology-based interventions may be effective in the short term, but significant limitations of this research include short study durations (up to 3 months) and the use of older technologies which are gradually being superseded by newer advances such as social media and instant messaging systems. Text messaging-based systems, reported to be particularly effective, are unable to offer 24-hour support or automated personalised feedback. Chatbots, an emergent conversation driven technology are available 24/7 while also providing personalised automated feedback and present users with a casual, easy to use interface. Chatbots have already been shown to influence positive outcomes in therapeutic health interventions. Further study involving the *WeightMentor* chatbot will assess its feasibility as a weight loss maintenance tool, and lay groundwork for a full-scale randomised trial.

WeightMentor was designed as a personalised motivational tool for weight loss maintenance using contemporary design technologies and frameworks and built on existing research which suggested that motivation and self-reporting are important for weight loss maintenance and that personalised messages should be relevant and encouraging without sounding patronising or bossy. Usability tests suggested that

WeightMentor is highly usable and that it should be possible for users to become competent in its use quickly. Using a chatbot such as *WeightMentor* may potentially improve the process of reporting, monitoring and communication of data related to weight loss maintenance and thus it is beneficial if such a tool is highly usable. The Chatbot Usability Questionnaire, trialled during the usability tests, has been shown to demonstrate validity and inter-rater reliability and with further testing may be deployed as a validated tool for chatbot usability tests. The potential construct validity of this tool when reduced from its native 16 items to 10 items makes it equivalent and comparable to the System Usability Questionnaire (SUS), which, although currently the leading measure for system usability is not specifically designed for chatbots, and use of a tool such as CUQ which is not only comparable to SUS but also specifically designed for chatbots will potentially improve the accuracy of chatbot usability testing research. During usability testing it was observed that 39 users were required to identify 80% of unique usability issues, which is contrary to what has been suggested by Nielsen and Landauer 1993, who propose a maximum of 5 to 8 users. However as has been discussed previously, Nielsen & Landauer's work predominantly involved conventional GUI (Graphical User Interface) systems, and the fact that 39 users are required to identify 80% of usability issues may suggest that the nature of chatbots as Conversational Artificial Intelligence (AI) systems makes it more difficult to identify usability issues and thus a greater number of users are required.

This PhD research project used an interdisciplinary approach to develop and test the *WeightMentor* chatbot, and to develop, trial and validate the Chatbot Usability Questionnaire. This project highlights the importance of interdisciplinary research for healthcare projects, which ensures that a broad spectrum of expertise may be utilised to fulfil requirements of the project.

Appendices

Appendix 1: EMBASE Database: Description & Search Strategy

Appendix 1.1 EMBASE Database Description

Produced by Elsevier, the EMBASE database contains over 32 million records from over 8,500 biomedical and pharmacological journals from more than 95 countries, published from 1974 to the present (Elsevier 2019). More than 1.5 million records are added annually. EMBASE includes all MEDLINE journals, although it also includes 6 million records from 2,900 unique titles. EMBASE also contains entries for more than 2.4 million conference abstracts, published since 2009. All journals and papers are fully indexed using Emtree, Elsevier's life science thesaurus (Elsevier 2019).

Appendix 1.2 EMBASE Search Strategy

1. exp mobile phone/
2. cellphone*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
3. cell phone*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
4. cellular phone*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
5. mobile phone*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
6. 1 or 2 or 3 or 4 or 5
7. exp computer/
8. exp personal digital assistant/
9. *computer/
10. computer*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
11. ipad*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
12. tablet*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
13. handheld*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]

14. hand held*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
15. 7 or 8
16. 9 or 10
17. 13 or 14
18. 15 and 16
19. 12 or 17
20. 18 and 19
21. 11 or 20
22. exp health/
23. exp medical informatics/
24. health*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
25. app*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
26. exp mobile application/
27. mobile*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
28. device*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
29. 22 or 24
30. 25 and 29
31. 25 and 27
32. 26 or 31
33. 27 and 28
34. medical*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
35. 25 and 27 and 34
36. 23 or 35
37. exp smartphone/
38. smartphone*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
39. smart phone*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
40. 38 or 39
41. 37 or 40
42. 6 or 41
43. exp telemedicine/
44. telemedicine.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
45. ehealth.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
46. mhealth.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
47. mobile health.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
48. exp telehealth/

49. telehealth.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
50. 48 or 49
51. exp personalized medicine/
52. phealth.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
53. 43 or 44
54. 51 or 52
55. 45 or 46 or 47 or 50 or 53 or 54
56. exp Internet/
57. internet*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
58. web based.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
59. web-based.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
60. website*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
61. web site*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
62. 56 or 57
63. 58 or 59 or 60 or 61 or 62
64. exp online system/
65. online.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
66. on-line.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
67. 64 or 65 or 66
68. exp text messaging/
69. messag*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
70. text*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
71. 68 or 69 or 70
72. personal*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
73. tailor*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
74. 72 or 73
75. notification*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
76. 71 or 75
77. 74 and 76
78. exp body mass/
79. body mass index.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]

80. bmi.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
81. 78 or 79 or 80
82. exp body weight/
83. body mass.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
84. body weight.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
85. 82 or 84
86. 83 or 85
87. exp diet/
88. diet*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
89. 87 or 88
90. exp nutrition/
91. nutrition*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
92. exp obesity/
93. obes*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
94. 90 or 91
95. 92 or 93
96. exp secondary prevention/ or exp prevention/ or exp primary prevention/ or exp "prevention and control"/
97. prevention.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
98. control.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
99. 96 or 97
100. 98 or 99
101. 95 or 100
102. overweight.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
103. over weight.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
104. 102 or 103
105. exp weight reduction/
106. weight loss.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
107. maintenance.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
108. 105 or 106
109. 107 and 108

110. weight loss maintenance.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
111. 21 or 30 or 32 or 36 or 42 or 55 or 63
112. 81 or 86 or 89 or 94 or 95 or 101 or 104
113. 109 or 112
114. 110 or 112
115. 77 and 111 and 113
116. 77 and 111 and 114
117. 116 and 2006:2017. (sa_year).

Appendix 2: MEDLINE Database: Description & Search Strategy

Appendix 2.1 MEDLINE Database Description

MEDLINE (MEDical Literature Analysis and Retrieval System (MEDLARS) OnLINE) is the premier database of the U.S National Library of Medicine (NLM) (NLM 2019). It contains more than 25 million references to life science (particularly biomedicine) articles from over 5,200 international journals in 40 languages. All records are indexed using NLM Medical Subject Headings (MeSH). MEDLINE predominantly covers literature published between 1966 and the present, with limited coverage of specific journals published prior to 1966 (NLM 2019).

Appendix 2.2 MEDLINE Search Strategy

1. exp Cell Phones/
2. cellphone*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
3. cell phone*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
4. cellular phone*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
5. mobile phone*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
6. 1 or 2 or 3 or 4 or 5
7. exp Computers/
8. exp Computers, Handheld/
9. *computer/
10. computer*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
11. ipad*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
12. tablet*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
13. handheld*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
14. hand held*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
15. 7 or 8
16. 9 or 10

17. 13 or 14
18. 15 and 16
19. 12 or 17
20. 18 and 19
21. 11 or 20
22. exp Health/
23. exp Medical Informatics Applications/
24. health*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
25. app*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
26. exp Mobile Applications/
27. mobile*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
28. device*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
29. 22 or 24
30. 25 and 29
31. 25 and 27
32. 26 or 31
33. 27 and 28
34. medical*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
35. 25 and 27 and 34
36. 23 or 35
37. exp Smartphone/
38. smartphone*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
39. smart phone*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
40. 38 or 39
41. 37 or 40
42. 6 or 41
43. exp Telemedicine/
44. telemedicine.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
45. ehealth.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
46. mhealth.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
47. mobile health.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
48. telehealth.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]

49. telecare.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
50. exp Precision Medicine/
51. phealth.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
52. 43 or 44
53. 50 or 51
54. 45 or 46 or 47 or 48 or 49 or 52 or 53
55. exp Internet/
56. internet*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
57. web based.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
58. web-based.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
59. website*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
60. web site*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
61. 55 or 56
62. 57 or 58 or 59 or 60 or 61
63. exp Text Messaging/
64. messag*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
65. text*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
66. 63 or 64 or 65
67. personal*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
68. tailor*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
69. 67 or 68
70. notification*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
71. 66 or 70
72. 69 and 71
73. exp Body Mass Index/
74. body mass index.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
75. bmi.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
76. 73 or 74 or 75
77. exp Body Weight/

78. body mass.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
79. body weight.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
80. 77 or 79
81. 78 or 80
82. Diet/
83. diet*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
84. 82 or 83
85. nutrition*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
86. Obesity/
87. obes*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
88. 86 or 87
89. exp Primary Prevention/ or exp Secondary Prevention/
90. prevention.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
91. control.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
92. 89 or 90
93. 91 or 92
94. 88 or 93
95. exp Overweight/
96. overweight.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
97. over weight.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
98. 95 or 96
99. 97 or 98
100. Weight Loss/
101. weight loss.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
102. maintenance.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
103. 100 or 101
104. 102 and 103
105. weight loss maintenance.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
106. 21 or 30 or 32 or 36 or 42 or 54 or 62
107. 76 or 81 or 84 or 85 or 88 or 94 or 99
108. 104 or 107
109. 105 or 107
110. 72 and 106 and 108

111. 72 and 106 and 109
112. 111 and 2006:2017. (sa_year).

Appendix 3: PubMed Database Description & Search Strategy

Appendix 3.1 PubMed Database Description

PubMed is a free search tool for peer-reviewed biomedical and life sciences literature. PubMed's database comprises more than 30 million entries (citations and abstracts) (National Library of Medicine (NLM) 2019a). Unlike EMBASE and MEDLINE it does not include full-text articles, which, when available, may be accessed through PubMed Central or directly from the publisher's website. The database is maintained by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) and includes resources from both MEDLINE and PubMed Central (NLM 2019a).

Appendix 3.2 PubMed Search Strategy

((cellphone* OR cell phone* OR cellular phone) OR ((computer* AND (handheld OR hand held OR tablet*)) OR iPad*) OR (health* AND app*) OR (mobile* AND (app* OR medical app* OR device* OR phone*)) OR (smartphone* OR smart phone*) OR (telemedicine OR telecare OR telehealth OR ehealth OR mhealth OR mobile health OR phealth) OR (web-based OR web based OR website* OR web site* OR internet OR online OR on-line)) AND ((personal* OR tailor*) AND (messag* OR text* OR push notification*)) AND ((body mass index OR bmi) OR (body AND (weight OR mass)) OR diet* OR nutrition* OR obes* OR (obesity AND (prevention OR control)) OR (weight loss maintenance) OR overweight OR over weight)

Appendix 4: Participant Recruitment Email, Needs Analysis Interviews

Dear Staff and Students,

You are being invited to participate in a research study titled "Using technology to maintain weight loss" conducted by Samuel Holmes, first year PhD student and supervised by an interdisciplinary team from the Ulster University.

The purpose of this research study is to identify the needs of adults who are seeking support for maintaining weight loss, identify how technology could help them, and determine the key features of a chatbot for weight loss maintenance.

Participants will be invited to take part in one interview which will be held in the participant's preferred campus of the Ulster University.

Participant eligibility criteria are summarised below:

Aged 18 years or over
EITHER previously lost weight and now maintaining it
OR previously lost weight, but have since regained it
Reasonable understanding of the English language

Participant ineligibility criteria are as follows:

Unwilling or unable to provide consent

Your participation in this study is entirely anonymous, voluntary and you can withdraw at any time.

If you have would like to find out more about this study or you are interested in participating, please contact Sam Holmes at holmes-w@ulster.ac.uk and you will be sent a Participant Information Sheet containing detailed information about the research and what is involved.

PhD Student details:
Samuel Holmes
School of Communication
(Contact Details Removed)

Project Chief Investigator:
Dr Anne Moorhead
School of Communication
(Contact Details Removed)

This research study has been approved by the School of Communication Filter Committee.

Appendix 5: Participant Information Sheet, Needs Analysis Interviews

Information Sheet - Interviews: Using technology to maintain weight loss

Background

We are inviting adults (18+ years) who have previously lost weight to take part in a research project. A multidisciplinary team from the Schools of Communication and Media, Computing and Nursing is conducting research into the area of communication in healthcare. The aim of this project is to design a chatbot that will support individuals in weight loss maintenance, through the delivery of personalised messages. As part of this project we are inviting adults who have previously lost weight to participate in interviews. During the interviews, participants will discuss which features to include in the chatbot, to investigate attitudes to tone and content of messages, and to decide how messages should be personalised.

Who can participate in this project?

In order to be eligible, participants must have recently lost weight and be actively maintaining this weight loss. Alternatively, participants will be eligible if they have previously lost weight but have since regained it.

What is required?

If you decide to participate in this project, you will be asked to participate in a short interview with a member of the research team. This interview will last approximately 30-60 maximum minutes and will take place on one of the University campuses, of your choice. You will be required to give written informed consent.

What are the interviews?

Interview participants will discuss weight loss maintenance and the use of technology. The purpose of the interviews is to determine how to support adults who are maintaining weight loss, to determine how technology can help them, and to identify the key features of such technology. We will also be investigating attitudes to the wording and personalisation of messages that may be sent to you by the system. You will have the opportunity to discuss as a group the features you think should be included in the system, and which should be left out.

Confidentiality

Please be assured that all information will be anonymous and treated as confidential. Your participation in this research project is voluntary and you will be free to withdraw at any time without giving any reason.

What will happen to the data collected?

Interviews will be audio recorded for analysis purposes only and accessed only by the research team. The findings will be used to publish research papers that will contribute to the existing body of literature by the addition of this new knowledge and will also be used to inform further research into this area.

What are the benefits of taking part in this research project?

You will be part of a project that will contribute to new knowledge on the communication of personalised messages in support of weight loss maintenance. This will contribute to further research with the aim of improving support for weight loss maintenance.
This research project has received ethical approval from the School of Communication Filter Committee.

What happens if I decide that I want to withdraw?

You have the right to withdraw from the project at any time without giving a reason for doing so. When you withdraw from the project, any data that has been collected from you will not be used in the research or published as part of the findings. All data that has been collected about you will be destroyed in compliance with the Data Protection Act (1998).

Contact Details

If you would like further information, or to discuss participating in this project, please do not hesitate to contact either the Chief Investigator or the PhD researcher.

Chief Investigator	PhD Researcher
Dr Anne Moorhead	Samuel Holmes
School of Communication & Media	School of Communication & Media
(Contact Details Removed)	(Contact Details Removed)

Project Team

Samuel Holmes (School of Communication & Media)
Dr Anne Moorhead (School of Communication & Media/Institute of Nursing & Health Research)
Dr Raymond Bond (School of Computing)
Professor Huiru (Jane) Zheng (School of Computing)
Professor Vivien Coates (School of Nursing/Institute of Nursing and Health Research)

Appendix 6: Consent Form, Needs Analysis Interviews

Consent Form

Study Title:	Interviews: Using technology to maintain weight loss

Chief Investigator:	Dr Anne Moorhead	PhD Student: Samuel Holmes

Please initial the boxes to confirm you have read and agree with the following statements:

1. I have been given and have read and understood the relevant information sheet for the above project and have asked and received answers to any questions raised. ☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without my rights being affected in any way. ☐
3. I understand that the researchers will hold all information and data collected during the project securely and in confidence and that all efforts will be made to ensure that I cannot be identified as a participant in the project (except as might be required by law) and I give permission for the researchers to hold relevant personal data. ☐
4. I agree to take part in this study and for data relating to me to be collected for analysis purposes only. ☐

Participant Name (Print)	Signature	Date
Name of person taking consent (If different from researcher)	Signature	Date
Researcher Name	Signature	Date

Appendix 7: Schedule, Needs Analysis Interviews

Briefing

The following briefing will be read to the participant before starting the interview:

1. Purpose: "The purpose of this interview is to investigate the needs of adults who need support for maintaining weight loss, to determine how technology can help them, and to identify the key features of such technology."
2. Participant role: "You will be asked to discuss what you think you will need to do to maintain your weight, the difficulties you think you might face, if you think technology will help you and how, and what functions you think this technology should have."
3. Duration: "This interview will last approximately 30 minutes."
4. Confidentiality: "All information is confidential and will be treated as such. "
5. Participation: "Participation is voluntary. You may withdraw at any point, and do not have to justify your decision for doing so."
6. "Findings from the interviews will be used to design, develop, and test a digital personalised messaging system to assist individuals weight loss maintenance. Findings may also be published in relevant scientific journals."
7. "An audio recording will be made of the interview. This is solely for analysis purposes. The recording will be accessed solely by the research team."

- *Ask participant if they have read the participant information sheet. Give them a copy if they need it.*
- *Ask for and answer any initial questions.*
- *Ask participant to read and sign the consent form.*
- *Ask participant to read and complete participant fact sheet.*
- *Switch on the recorder:* "I have switched on the audio recorder, and now the interview session is being recorded."

Questions

1. "Congratulations on your weight loss! Do you have any plans for maintaining this weight loss?"
    - Are you going to maintain it?
    - Do you want to maintain it long term?
    - How long for?
    - Do you have any weight loss maintenance goals and why?
    - Do you want to lose further weight?
    - Are you trying to keep the weight off for a special event or function?

2. "What do you think you will need to do to maintain your weight loss?"
    - Do you need to make any further changes to your diet?
    - Do you need to look at your physical activity?

3. "What challenges do you consider you will experience in maintaining weight loss?"
    - How easily tempted are you?
    - When are you most likely to be at risk?
    - Do things like stress, tiredness, or emotions affect your behaviour?

4. "Are you doing anything to overcome these challenges, and if so, what?"
    - Could you ask a friend to help you?
    - Do you need to look at your behaviour and habits?
    - Does social support help?
    - What kind of social support?

5. "Have you ever considered using apps to help you maintain weight loss, and why?"
    - Do you know what sort of apps are available?
    - Do existing apps have functions you find useful?
    - Are the existing apps easy to use?
    - Do you find technology easy to use in general?
    - Would you find it too time consuming to use an app?
    - If yes, what would you like this app to do?
    - If yes, how would you like it to help you?
    - If yes, what features should it have?

6. "Do you consider that a personalised message would motivate you to try and maintain weight loss?"
    - How would you personalise the message?
    - Who would the message come from?
    - If you knew you were getting a message from a virtual person, how would that make you feel?

Show the handout about messages (part b) to the user

7. "Please have a look at the types of message described on the handout. How would it make you feel to receive a message like this?"
    - Which type of message would motivate you and why?
    - Which type would you find least helpful and why?

8. "Please have a look at the message tones described on the handout. How would it make you feel to receive a message like this?"
    - Which message tone would motivate you and why?
    - Would any of these tones have the opposite effect, and why?

"We're going to look at something called a "chatbot". A chatbot is a computer program that simulates human conversation."

Demonstrate the WeightMentor chatbot.

9. "What did you think of the interactions we had with this chatbot?"
    - Did it seem realistic?
    - Did you find it unnerving?

10. "If the app we talked about earlier was the WeightMentor chatbot, would you use it and why?"
    - Would you feel uncomfortable interacting with it?
    - Would you find it more natural to interact with a chatbot?
    - How would you feel if it was humorous?
    - Would you prefer the app to take its role seriously?

"Thank you for participating in this interview."

Turn off the recorder

Appendix 8: Participant Fact Sheet, Needs Analysis Interviews

Please complete the following short questionnaire about yourself. Remember that all data collected will be confidential.

Question 1
Please indicate which of the following age groups you fit into:

18-25	26-30	31-35	36-40	41-45	45-50	Over 50

Question 2
Please indicate your gender:

Male ☐ Female ☐ Other ☐ Prefer not to say ☐

Question 3

What is your height?

_____

Question 4

What is your current weight?

_____

Question 5

What was your weight before your weight loss?

_____

(For researcher use only)

Amount of weight lost:

_____

Question 6

Over how long did you lose this weight? (e.g. 3 months)

_____

(Please Turn Over) →

Question 7

Please indicate how you lost your weight:

Weight loss programme (e.g. Weight Watchers)	
Medical advice	
NHS obesity clinic	
Diet plan (e.g. 5/2, Atkins etc.)	
Self-regulation	
Other (please specify below)	

Appendix 9: Participant Handout, Needs Analysis Interviews

There are several types of message that may be sent, these are listed in the table below.

Message Type	Description
Request for feedback	Asking participants to provide feedback on their progress. E.g. "How physically active have you been today, on a scale of 0 to 5?"
Reflection	Asking participants to think more deeply about situations or behaviours. E.g. "Can you think of things that discourage you from exercising? How do you think you could overcome those challenges?"
Generic messages	Messages that just convey a general theme E.g. "Nibbling during the day can contribute to weight regain!"
Tailored messages	Messages that are specifically tailored towards the individual user. E.g. "I see you have been working hard to reduce the amount of nibbling during the day. Well done, that's going to help control your weight!"
Tips	Tips and advice. E.g. "If you are tempted to snack in between meals, try drinking water. Sometimes it's easy to mistake thirst for hunger."

Additionally, there are different tones that may be conveyed by these messages. These are listed below.

Tone	Description
Authoritative	Telling the user what they should or should not do. E.g. "Don't nibble in between meals, you will find it harder to maintain your weight!"
Empathetic	Creating empathy with the user. E.g. "It can be difficult to resist the urge to go snacking. I understand how you feel!"
Congratulatory	Congratulating the user on a job well done. E.g. "Oh, you went out for a walk after work? Well done, I bet you feel great now!"
Commiserating	Messages that commiserate the user if they have been unable to achieve a goal. E.g. "Don't feel bad, these things happen. It could have been worse. Just hang in there!"
Encouragement	Messages that encourage the user to keep going. E.g. "Do you realise this is the third day in a row that you've been physically active? Keep going, you're doing well."

Appendix 10: Coding Table, Needs Analysis Interviews

Topics	Sub-themes	Key themes	Quotes
further weight loss maintenance enforcing diet maintaining physical activity stress emotion boredom temptation snacking negative social influences encouragement accountability “doing it together” managing diet managing exercise slimming/fitness clubs personalised app feedback app convenience personalised recommendations progress tracking linking different apps myfitnesspal fitbit weight watchers app slimming world app timely messages tailored messages don’t nag don’t be bossy don’t overdo sentiment who is the message from? chatbots are engaging chatbots are friendly	Weight loss goals Maintenance strategies Challenges to weight loss Solutions to challenges Negative social influences Positive social influences Popular apps Desirable app functions Personalised messages Chatbots	Weight loss maintenance: the challenge Social contact: a double-edged sword Apps are popular Tailored messages are more useful Chatbots have potential for weight loss maintenance	Theme 1: “Sometimes you don’t want to do any cooking when you get home, so you order a pizza” “I find that exercise helps, I’m less likely to be stressed” “If you’re stressed it’s easy to [eat out] or buy a high-calorie meal”. “I comfort eat” “If I’m hungry I sometimes make the wrong choices” “I don’t drink, I don’t smoke but I feel I have as much addiction to say crisps, or chocolate” “unless I am out doing my activity or whatever, there is a tendency to go to the cupboard and maybe have a few snacks, where during the day is fine, keeping active” “when I had drink at the weekends, you had quite a few tots, mine was whisky and coke, then you get the munchies and then you are on the phone and it is just a progressive cycle, whereas now I caravan quite a bit, I bought a caravan to study, I know it sounds crazy but it got me out of the house and I went away and studied...temptation used to be a problem for me, but now it isn't” “if I don’t have my breakfast in the morning, by the time I come to college and I go through the mall and I smell the bacon, I want bacon!” “I would comfort eat, yes, I am an emotional eater so that is probably the biggest issue is to try and change that behaviour” “I think that exercise helps...and because I have to do some physical activity every day, or at least every other day, I have made that part of my routine, I find it probably helps with the stress because I think that when you stop and you start getting, eating badly it can then become a downward spiral quite quickly”

Topics	Sub-themes	Key themes	Quotes
chatbot personality straightforward minimise chatbot interactions simplicity of chatbots			Theme 2: "If someone knows [you're losing weight] they'll invite you out less, but sometimes people say "It's only one drink, it's only one chocolate bar. That's not the point!" "It's not really an issue for [friends who aren't overweight], but at the other end of the scale, people who are overweight look at you and think you have nothing to worry about." "[at a slimming club] you have the weigh-in, I think it helps because I think everybody is friendly and there is different people with different weights, so in some ways you feel you are not the worst off, but with some people you have aspirations too" "With exercising, sometimes I'll go to the gym with a friend and we'll do the same activity together, having that social support definitely helps." "If you are [at work events] they are always throwing everything at you to get you drunk" Theme 3: "Sometimes it's just, can you be bothered inputting the data…" "It takes a lot more time to input [into a specific app] manually and stuff. It was just laziness and I got out of the way of it" "It is the fact that [the Fitbit] is connected to everyday living, whereas the phone is a little bit remote" "I liked [a weight loss app] because it runs in the background, you can kind of set it and forget about it, yeah it's good." "a lot of these apps aren't really intrusive or make me you know think, oh they are difficult to operate, or they take a long time to get working properly" "Sometimes [a weight loss app] can be quite time consuming because if you make a meal at home you have to log every single part of it, but if you have one you

Topics	Sub-themes	Key themes	Quotes
			can scan the barcode which is quite easy to use" "I didn't like [a smartphone app] because you had to say how much you were eating, how much calories they ask, yeah. So many things you had to do." Theme 4: “there is nothing wrong with generic messages, but I just think that it is...yes it’s similar to a tip but to me if it is not a tip then it is not as useful” "I think tailored messages are best because it feels like you are engaging with it…generic messages would be less useful" "Generic messages would be a definite no…if I sent someone a message, I'd expect them to come back to me with something specific to me and not a textbook answer” "Having a tailored message boosts motivation, I wouldn't want any generic messages" "Tailored messages...feel like they relate to what you are doing" "I would find tailored messages most useful" "Tailored messages need to be very specific to me” "I think tailored messages are quite good" "I think that [tailored messages make people] think of the good things that they are doing, that’s a good thing” "Tailored messages are very important" "why would I want to go and download an app just to get patronizing or random messages that have nothing to do with me. If you are going to get me to download an app you need to make it specific for me" "I like messages that are specifically tailored to me" "Tailored messages would have to be tailored to me not tailored to

Topics	Sub-themes	Key themes	Quotes
			my problem, but tailored to me and my behaviours" "specific tailored messages, they are referenced, well done, or you are doing something that is helpful, so that is encouraging" "If they know what you are doing is not a bad thing, that is positive reinforcement" "I think if it just, if an authoritative one just keeps telling you stuff and you are just like, leave me alone, stop being annoying" "I would actually take [authoritative messages] as, who do you think you are" "It is because [Authoritative] is punitive isn't it" "So authoritative, the thing is, this is a tricky one because some people I am sure like being told things softly" "I guess it depends on your audience member or the person who is receiving that message, how they like being talked to" "It's just telling me what to do" "people who are adults who struggle with their weight don't need that, that's the voice in their own head" "Authoritative is just trying to tell you what to do" "I don't like being told what to do" "[Authoritative messages are] a bit bossy" "Authoritative messages are just like bossing me about". "If it was too cheesy, I'd just throw it away!" "If you make people become too comfortable there's not much point…" "I would feel a bit patronised by an insincere message…" Theme 5: "I would use it because it is engaging and to me it seems a lot more fluent how to enter data"

Topics	Sub-themes	Key themes	Quotes
			"I would use it because it is natural and not too generic" "I like the graph function and the motivational aspect; I think it could be quite useful" "I liked it because it keeps me accountable and would be a good way to keep people motivated" "I thought it was natural, and if it could be linked to post to your Facebook feed it would be really useful!" “I would want to use it when I want to, without having to use it every day, or whatever.” “I think there are potentially too many questions [in the activity logging] …I think someone could potentially get frustrated.” “I would like to be able to choose the personality. It would be nice to have an upbeat personality one day and more serious the next.” “I liked the humour of the hello and goodbye…I think as well if you could create some kind of ‘character’, that might be quite useful.

Appendix 11: *WeightMentor* Code, selfreporting.js

```
'use strict';
const charts = require('./charts');
const common = require('./common');
const database = require('./database');
const messaging = require('./messaging');
const strings = require('./strings');
const user = require('./user');
module.exports = {
    getTrendChart: function(type, senderID, callback) {
        let dbRow;
        let chartData = [], chartURL;
        let trendSQL;
        trendSQL = "SELECT value FROM self_reporting WHERE fbid='"
+ senderID + "' AND srtype='" + type + "' ORDER BY srid ASC";
        database.query(trendSQL, function(result) {    // Query
database to pull out self-report history
            for (let i = 0; i < result.length; i++) { // Loop
through each row in result
                dbRow = result[i];
                chartData.push(dbRow.value);    // Add the 'i'th
value to chartData
            }
            chartURL = charts.plot(senderID, type, chartData);
// Plot chart and return URL
            callback(chartURL);
        });
    },

    handleReport: function(senderID, type, responseText, callback)
{
        let dbRow, delay;
        let sqlCountQuery = "SELECT COUNT(srid) FROM self_reporting
WHERE fbid='" + senderID + "' AND srtype='" + type + "'";
        let txtResponse;
        database.query(sqlCountQuery, function(result) { // Check
if user has previously reported
            dbRow = result[0];
            if (Number(dbRow.count) === 0) {  // No history
                console.log("No previous self-reports found for
user '%s'.", senderID);
                writeData(senderID, type, function() {  // Write to
database
                    if (responseText.includes("%text-response%")) {
                        getResponse(senderID, type,
function(response) {
                            responseText =
responseText.replace("%text-response%",
common.insertUserID(senderID, response));
                            delay = common.setDelay(responseText);
                            setTimeout(function() {
                                messaging.sendTextMessage(senderID,
responseText); // Send message to user
                                callback(responseText);
                            }, delay);
                        });
                    }
                });
```

```
            } else if (Number(dbRow.count) < 5) { // Fewer than
five previous self-reports
                let largerSR;
                console.log("%s self-reports found for user '%s'",
dbRow.count, senderID);
                getLastSR(senderID, type, function(prev) {
                    writeData(senderID, type, function() { // Write
to database
                        largerSR = compareSRValues(senderID, type,
prev);
                        if (largerSR === "current") {
                            if (type === 1) {
                                // Activity
                                txtResponse = "You were more active
today than last time you reported, %USER.";
                            } else if (type === 2) {
                                // Food
                                txtResponse = "You ate more today
than the last time you reported, %USER.";
                            }
                        } else if (largerSR === "previous") {
                            if (type === 1) {
                                // Activity
                                txtResponse = "You were less active
today than the last time you reported, %USER.";
                            } else if (type === 2) {
                                // Food
                                txtResponse = "You ate less today
than the last time you reported, %USER.";
                            }
                        } else {
                            if (type === 1) {
                                // Activity
                                txtResponse = "Your physical
activity today was the same as the last time you reported, %USER.";
                            } else if (type === 2) {
                                // Food
                                txtResponse = "Your food intake
today was the same as the last time you reported, %USER.";
                            }
                        }
                        if (responseText.includes("%text-
response%")) {
                            getResponse(senderID, type, function
(response) {
                                responseText =
responseText.replace("%text-response%",
common.insertUserID(senderID, txtResponse));
                                delay =
common.setDelay(responseText);
                                setTimeout(function () {

messaging.sendTextMessage(senderID, responseText); // Send message
to user
                                    responseText =
common.insertUserID(senderID, response);
                                    delay =
common.setDelay(responseText);
                                    setTimeout(function () {
```

```
messaging.sendTextMessage(senderID, responseText);
                                        callback(responseText);
                                    }, delay);
                                }, delay);
                            });
                        }
                    });
                });
            } else { // Five or more self-reports
                console.log("%s self-reports found for user '%s'",
dbRow.count, senderID);
                writeData(senderID, type, function() {  // Write to
database
                    getPreviousReports(senderID, type, function
(prevReports) { // Get chart of previous reports
                        if (responseText.includes("%text-
response%")) {
                            getResponse(senderID, type,
function(response) {
                                responseText =
responseText.replace("%text-response%",
                                    common.insertUserID(senderID,
strings.prevReportsMessage));
                                // Replace "%text-response%" with
the previous reports message
                                delay =
common.setDelay(responseText);   // Calculate delay based on
response
                                messaging.sendTextMessage(senderID,
responseText);  // Send previous reports message
                                setTimeout(function() {

messaging.sendImageMessage(senderID, prevReports);    // Display
image
                                }, delay);
                                delay = common.setDelay("_IMAGE_");
// Create a delay of 10 seconds
                                responseText =
common.insertUserID(senderID, response);
                                // Get a motivational response
based on self-reported data
                                setTimeout(function() { // Timeout
for 10 seconds to allow user to look at chart

messaging.sendTextMessage(senderID, responseText);
                                    callback(responseText);
                                }, delay);
                            });
                        }
                    });
                });
            }
        });
    },
};



function compareSRValues(senderID, type, previous) {
    let current, larger;
```

```
    current = user.getSelfReport(senderID, type);
    if (Number(current) < Number(previous)) { // Current is lower
        larger = "previous";
        return larger;
    } else if (Number(current) > Number(previous)) { // Previous is
lower
        larger = "current";
        return larger;
    } else {    // Current and previous are equal
        larger = "none";
        return larger;
    }
}

function getLastSR(senderID, type, callback) {
    let dbSelection = "value";
    let dbTable = "self_reporting";
    let dbCondition = "srtype=" + type + " AND fbid='" + senderID
+"' ORDER BY date DESC, time DESC";
    let limit = 1;
    let dbRow;
    let prevSR;
    database.select(dbSelection, dbTable, dbCondition,
function(result) {
        dbRow = result[0];
        prevSR = dbRow.value;
        callback(prevSR);
    }, limit);
}

function getPreviousReports(senderID, type, callback) {
    let chartData = [], chartURL;
    let dbRow;
    let sqlAverageQuery = "SELECT CAST(AVG(value) AS INT) FROM
self_reporting WHERE fbid='" + senderID + "' AND srtype='" +
        type + "'";
    let sqlLastFiveQuery = "SELECT * FROM (SELECT * FROM
self_reporting WHERE fbid='" + senderID + "' AND srtype='" + type +
        "' ORDER BY srid DESC LIMIT 5) AS lastfive ORDER BY date
ASC, time ASC";
    database.query(sqlLastFiveQuery, function(result) { // Get last
five sr values
        for (let i = 0; i < result.length; i++) { // Loop through
each row in result
            dbRow = result[i];
            chartData.push(dbRow.value);    // Add the 'i'th value
to chartData
        }
        database.query(sqlAverageQuery, function (result) {  // Get
average sr value
            dbRow = result[0];
            user.setAvg(senderID, type, dbRow.avg);   /**
@namespace dbRow.avg */
            chartURL = charts.plotPrevious(senderID, type,
Number(user.getAvg(senderID,type)), chartData);
            callback(chartURL);
        });
    });
}
```

```
function getResponse(senderID, type, callback) {
    // Function to get responses based on sr value
    let response;
    let value;
    let dbSelection = "rtext";
    let dbTable = "sr_responses";
    let dbCondition = "srtype=_TYPE_ AND srvalue=_VALUE_";
    let dbRow;
    value = user.getSelfReport(senderID, type);
    dbCondition = dbCondition.replace("_TYPE_",
type).replace("_VALUE_", value);    // Build condition statement
based
    database.select(dbSelection, dbTable, dbCondition,
function(result) {                   // on type and value
        // Select from database, and store as response
        dbRow = result[0];
        response = dbRow.rtext;    /** @namespace result.rtext */
        callback(response); // Send response as callback
    });
}

function writeData(senderID, type, callback) {
    let dbColumns, dbRow, dbTable, dbValues;
    let inserted;
    let sqlInsertQuery = "INSERT INTO self_reporting (srtype, fbid,
value, date, time) VALUES ('" + type + "', '" + senderID +
            "', '" + user.getSelfReport(senderID, type) + "', '" +
user.getTimeStamp(senderID, type)[1] + "', '" +
                user.getTimeStamp(senderID, type)[0] + "')
RETURNING srid";
    if (type === 1) {   // Activity
        if (user.getDetails(senderID, type)=== strings.NO) { // No
additional details supplied
            database.query(sqlInsertQuery, function(result) { //
Insert sr value and callback true
                callback(result);
            });
        } else {    // Additional details supplied
            database.query(sqlInsertQuery, function(result) {   //
Insert sr value and get id of inserted value
                dbRow = result[0];
                sqlInsertQuery = "INSERT INTO sr_activity_details
(srid, resting, sitting, standing, moving, exercise) " +
                    "VALUES ('" + dbRow.srid + "', '" +
user.getDetailInfo(senderID, type, 1) + "', '" +
                        user.getDetailInfo(senderID, type, 2) + "',
'" + user.getDetailInfo(senderID, type, 3) +
                            "', '" + user.getDetailInfo(senderID,
type, 4) + "', '" +
                                user.getDetailInfo(senderID, type,
5) + "') RETURNING id";    /** @namespace dbRow.srid */
                if (user.getDetailInfo(senderID, type, 5) ===
strings.NO) { // No exercise details
                    database.query(sqlInsertQuery, function(result)
{ // Insert additional details and callback true
                        callback (result);
                    });
                } else {    // Exercise details supplied
```

```
                    database.query(sqlInsertQuery, function(result)
{   // Insert additional details and get id
                        dbRow = result[0];
                        dbTable = "sr_activity_exercise";
                        dbColumns = ["did", "type", "duration"];
                        sqlInsertQuery = "INSERT INTO
sr_activity_exercise (did, type, duration) VALUES ('";
                        for (let key in user.getExercise(senderID))
{   // For each key in exercise object...
                            // noinspection JSUnfilteredForInLoop
                            if
(user.getExercise(senderID).hasOwnProperty(key)) {
                                // noinspection
JSUnfilteredForInLoop
                                dbValues = [`'${dbRow.id}'`,
`'${key}'`, `'${user.getExercise(senderID)[key]}'`];
                                // noinspection
JSUnfilteredForInLoop
                                sqlInsertQuery += dbRow.id + "', '"
+ key + "', '" + user.getExercise(senderID)[key] + "')";
                                database.insert(dbTable, dbColumns,
dbValues);  //...insert into database
                            }
                        }
                        inserted = true;
                        callback(inserted);
                    });
                }
            });
        }
    } else {    // Food
        if (user.getDetails(senderID, type) === strings.NO) {
            database.query(sqlInsertQuery, function (result) {
                callback(result);
            });
        } else {
            database.query(sqlInsertQuery, function(result) {
                dbRow = result[0];
                sqlInsertQuery = "INSERT INTO sr_food_details
(srid, breakfast, lunch, dinner, supper, snacks) VALUES ('" +
                    dbRow.srid + "', '" +
user.getDetailInfo(senderID, type, 1) + "', '" +
                        user.getDetailInfo(senderID, type, 2) + "',
'" + user.getDetailInfo(senderID, type, 3) + "', '" +
                            user.getDetailInfo(senderID, type, 4) +
"', '" + user.getDetailInfo(senderID, type, 5) +
                                "') RETURNING fid";
                database.query(sqlInsertQuery, function (result) {
                    callback(result);
                });
            });
        }
    }
}
```

Appendix 12: *WeightMentor* Code, charts.js

```
/*
* charts.js - Module for plotting charts using ImageCharts
* */
'use strict';
const common = require('./common');
const strings = require('./strings');
// Constants for chart creation
module.exports = {
    plot: function(senderID, type, data) {
        const ACT_COLOUR = "424FB2";
        const FOOD_COLOUR = "CE1D1D";
        const HOSTNAME = "https://image-charts.com/chart?";
        const OPTIONS = "cht=ls" + // Chart type (line with hidden
axes by default)
            "&chf=bg,s,ffffff" + // Chart fill (background, solid,
white)
            "&chg=1,1,2,5" +    // Chart grid lines (steps of 1 on
each axis, dashed 2px with 5px spacing)
            "&chxt=y" + // Enable y axis
            "&chxr=0,0,6,1" + // Set y axis range (0 to 6) with a
step of 1
            "&chs=999x300" +    // Chart size (999x300)
            "&chma=5,10,5,5" +  // Chart margins (left=5px,
right=10px, top=5px, bottom=5px)
            "&chtt=_title_" +   // Chart title
            "&chts=000000,20" + //Chart title colour and size
(black, 20pt)
            "&chls=2" + // Chart line thickness (2px)
            "&chco=_colour_";   // Chart colour
        let chartColour, chartCustomOpts, chartData, chartTitle,
chartURL;  // Chart variables
        if(type === 1) {      // Activity Trend
            chartColour = ACT_COLOUR;
            chartTitle = common.insertUserID(senderID,
strings.activityChart);
        } else {   // Food Trend
            chartColour = FOOD_COLOUR;
            chartTitle = common.insertUserID(senderID,
strings.foodChart)
        }
        if (data.length <= 5) {
            chartData = encodeText(data);   // Use text encoding
for five or fewer data elements
        } else {
            chartData = encodeSimple(data);   // Use simple
encoding for more than five data elements
        }
        chartCustomOpts = OPTIONS.replace("_title_",
chartTitle).replace("_colour_", chartColour);  // Add title and
colour codes to chart options
        chartURL = HOSTNAME + chartCustomOpts + chartData;  //
Build chart URL
        console.log("Chart URL is " + chartURL);
        return chartURL;
    },

    plotPrevious: function(senderID, type, average, data) {
```

```
        const ACT_COLOUR = "424FB2";
        const FOOD_COLOUR = "CE1D1D";
        const ACT_AVG_COLOUR = "F1C40F";
        const FOOD_AVG_COLOUR = "196F3D";
        const HOSTNAME = "https://image-charts.com/chart?";
        const OPTIONS = "cht=ls" + // Chart type (line with hidden
axes by default)
            "&chf=bg,s,ffffff" + // Chart fill (background, solid,
white)
            "&chg=1,1,2,5" +    // Chart grid lines (steps of 1 on
each axis, dashed 2px with 5px spacing)
            "&chxt=y" + // Enable y axis
            "&chxr=0,0,6,1" + // Set y axis range (0 to 6) with a
step of 1
            "&chs=999x300" +    // Chart size (999x300)
            "&chma=5,10,5,5" +  // Chart margins (left=5px,
right=10px, top=5px, bottom=5px)
            "&chtt=_title_" +   // Chart title
            "&chts=000000,20" + //Chart title colour and size
(black, 20pt)
            "&chls=2|2" + // Chart line thickness (2px)
            "&chdl=_type_|Average" +     // Legend labels
            "&chdls=000000,12" +    // Legend text colour and size
            "&chco=_colour_,_avgcolour_";   // Chart colour
        let avgColour, chartColour, chartCustomOpts, chartData,
chartTitle, chartURL, legendTitle;  // Chart variables
        if(type === 1) {      // Activity Trend
            chartColour = ACT_COLOUR;
            avgColour = ACT_AVG_COLOUR;
            chartTitle = common.insertUserID(senderID,
strings.previousActivityChart);
            legendTitle = "Activity";
        } else {   // Food Trend
            chartColour = FOOD_COLOUR;
            avgColour = FOOD_AVG_COLOUR;
            chartTitle = common.insertUserID(senderID,
strings.previousFoodChart);
            legendTitle = "Food";
        }
        chartData = buildPreviousSeries(data, average);
        chartCustomOpts = OPTIONS.replace("_title_",
chartTitle).replace("_colour_", chartColour)
            .replace("_avgcolour_", avgColour).replace("_type_",
legendTitle);
        // Add titles and colour codes to chart options
        chartURL = HOSTNAME + chartCustomOpts + chartData;  //
Build chart URL
        console.log("Chart URL is " + chartURL);
        return chartURL;
    },
};

function buildPreviousSeries(data, average) {
    const SERIES_SEPARATOR = "|";
    const DELIMITER = ",";
    let chartData;
    chartData = encodeText(data) + SERIES_SEPARATOR;
    for (let i = 0; i < data.length; i++) {
        if (i === (data.length - 1)) {  // Last element in array
            chartData += average;   // Add last element to string
```

```
without delimiter
        } else {    // Not the last element
            chartData += average + DELIMITER;   // Add last element
to string, followed by delimiter
        }
    }
    return chartData;
}

function encodeSimple(data) {   // Function to encode chart data
using simple encoding
    const DATA_SIMPLE = "&chd=s:";  // Beginning of simple data
segment
    const SIMPLE_ENCODING = "ABCDEF";  // String for encoding
values
    let encodedData = DATA_SIMPLE;
    let encodedElement;
    for (let i = 0; i < data.length; i++) {  // Loop through the
array
        encodedElement = SIMPLE_ENCODING.charAt(data[i]);    //
Return a character based on the current element
        encodedData += encodedElement;  // Add encoded element to
encoded data string
    }
    return encodedData;
}

function encodeText(data) { // Function to encode chart data using
text encoding
    const DATA_TEXT = "&chd=t:";    // Beginning of text data
segment
    let encodedData = DATA_TEXT;
    encodedData += data.join(); // Convert data array to string and
add to encoded data string
    return encodedData;
}
```

Appendix 13: *WeightMentor* Code, gifs.js

```
'use strict';
module.exports = {
    getCongratulationsGIF: function(callback) {
        getGIF(2, function(gif) {
            callback(gif);
        });
    },

    getGoodbyeGIF: function(callback) {
        getGIF(3, function(gif) {
            callback(gif);
        });
    },

    getGreetingGIF: function(callback) {
        getGIF(1, function(gif) {
            callback(gif);
        });
    },
};

function getGIF(type, callback) {
    const EXTENSION = ".gif";
    const NAME_CONGRATULATIONS = "congratulations";
    const NAME_GOODBYE = "goodbye";
    const NAME_GREETING = "greeting";
    const MAX = 20;
    let gifName;
    let randomGIF;
    if (type === 1) {   // Greeting
        gifName = NAME_GREETING;
    } else if (type === 2) {    // Congratulations
        gifName = NAME_CONGRATULATIONS;
    } else {    // Goodbye
        gifName = NAME_GOODBYE;
    }
    randomGIF = Math.floor(Math.random() * MAX) + 1;
    if (randomGIF < 10) {
        gifName += "0" + randomGIF;
    } else {
        gifName += randomGIF;
    }
    gifName += EXTENSION;
    callback(gifName);
}
```

Appendix 14: *WeightMentor* Code, motivation.js

```
/*
* Module for motivating the user
* */
'use strict';
const database = require('./database');
const user = require('./user');
const self = module.exports = {
    motivateUser: function(senderID, responseText, callback) {
        // Function to motivate the user by sending a quote, tip,
or image
        let dbRow;
        let maxIDQuery = "SELECT MAX(id) FROM mm_inspire";
        let mmQuery;
        let randomMax, randomRow;
        database.query(maxIDQuery, function(result) { // Query
database to get maximum mmid - maximum number of records
            dbRow = result[0];
            randomMax = dbRow.max;
            randomRow = Math.floor((Math.random()*randomMax)+1);
// Generate a random number corresponding to a row
            while
(user.getMotivationsUsed(senderID).includes(randomRow)) {
                console.log("Motivation number %s has already been
used.", randomRow);
                randomRow =
Math.floor((Math.random()*randomMax)+1);    // Get another random
row
            }
            user.setMotivation(senderID, randomRow);
            console.log("Motivation %s added to recently used list.
List contents: %s", randomRow,
                user.getMotivationsUsed(senderID));
            mmQuery = `SELECT type, data FROM mm_inspire WHERE id
<=${randomRow} ORDER BY id DESC LIMIT 1;`;
            database.query(mmQuery, function(result) { // Query
database to retrieve random row
                dbRow = result[0];
                self.updateImpressions(randomRow);  // Update the
'impressions' field for this response
                callback (dbRow);
            });
        });
    },

    updateLikes: function(type, id) {
        // Function to update the number of likes for a specific
motivational response
        let table = 'mm_inspire';
        let condition = `id = '${id}'`;
        let updates;
        if (type === "like") {
            updates = 'likes = likes + 1';  // User liked the
motivation, add a like
        } else {
            updates = 'likes = likes - 1';  // User didn't like the
motivation, remove a like
        }
```

```
        database.update(table, updates, condition);   // Update the
database
    },

    updateImpressions: function(id) {
        // Function to update the number of impressions i.e. number
of times a response has been used anywhere
        let table = 'mm_inspire';
        let condition = `id = '${id}'`;
        let updates = 'impressions = impressions + 1';  //
Increment number of impressions
        database.update(table, updates, condition);   // Update the
database
    }
};
```

Appendix 15: Participant Recruitment Email, *WeightMentor* Usability Testing

Dear Staff & Students,

You are being invited to participate in a research study as part of a PhD. This study will be conducted by Samuel Holmes, second year PhD Researcher and supervised by an interdisciplinary team.

The purpose of this study is to test the usability of *WeightMentor*, a Facebook **M**essenger-based chatbot designed for helping with weight loss maintenance.

As part of this research study, participants will take part in a usability test. These tests will involve attempting several tasks using the chatbot. After the test, participants will provide feedback on the chatbot by completing three questionnaires.

Eligible participants will be aged 18+, have access to Facebook, and be willing to use a chatbot. Note: This study will be testing the usability of the chatbot, rather than its usefulness for weight loss and weight loss maintenance. It is not necessary for participants to have an interest in weight loss and weight loss maintenance in order to participate.

Your participation in this study is entirely anonymous, voluntary and you can withdraw at any time.

This research study has been approved by the School of Communication & Media Ethics Filter Committee.

If you have would like to find out more or you are interested in participating, please contact Sam Holmes at holmes-w@ulster.ac.uk and you will be sent a Participant Information Sheet containing detailed information about the research and what is involved.

PhD Researcher details:
Samuel Holmes
School of Communication and Media
(Contact Details Removed)

Project Chief Investigator:
Dr Anne Moorhead
School of Communication and Media
(Contact Details Removed)

Appendix 16: Participant Information Sheets, *WeightMentor* Usability Testing

Personalised Chatbot for Weight Loss Maintenance: Usability Testing Individual Sessions Participant Information Sheet

Background

You are being invited to participate in a research study as part of a PhD. The aim of this study is to conduct usability tests to assess the usability of WeightMentor, a custom designed chatbot for helping with weight loss maintenance. We are recruiting participants for these tests.

What is usability testing?

Usability testing is a way of finding out how easy a computer system is to use. The "computer system" in this case is the custom designed chatbot called WeightMentor. Chatbots are intelligent systems that simulate human conversation. Participants will be observed completing different tasks using this chatbot, to finds out how easy it is to use. Results of the tests can be used to make recommendations for improving the chatbot.

Who can participate in these tests?

We are inviting adults (age 18+) who are willing to use the chatbot, WeightMentor. Participants are advised that we will be testing the usability of this chatbot, rather than its usefulness for weight loss and weight loss maintenance. It is not necessary for participants to have an interest in weight loss and weight loss maintenance in order to participate. Participants do not need to have lost or maintaining weight loss to participate in this study but if they have previously lost weight they can also be included.

What is involved?

The individual usability tests will take place within a meeting room on one of the four campuses of the Ulster University. Participants will be invited to attend a test on their home campus. Participants will access the WeightMentor chatbot through Facebook Messenger on a desktop PC and complete several tasks to assess usability, such as creating a user profile or sending self-reported physical activity and food consumption data to the chatbot. During the tasks, participants' interactions with the chatbot will be recorded using a camera that will record the screen of the mobile device and the participant's hands. Note that no faces will be recorded. During tests, participants will "think aloud", describing what they can see on the screen, what they are going to do, and how the chatbot is interacting with them. The audio from this "thinking aloud" will be recorded. After the usability tests, participants will be asked to complete three questionnaires to assess the usability of the chatbot.

Note: We will be testing the system, not the participants. There is no chance of "failure". Participants will not be penalised if they are unable to complete a task.

This research project has received ethical approval from the Communication Ethics Filter Committee.

What are the benefits of taking part?

You will be part of a project that will contribute to new knowledge on the communication of personalised messages in support of weight loss maintenance. This will contribute to further research with the aim of improving support for weight loss maintenance.

What happens if I decide that I want to withdraw?

You have the right to withdraw from the study at any time without giving a reason for doing so. If you withdraw from the study, any data that has been collected from you will not be used in the research or published as part of the findings. All data that will be collected about you will be managed, stored and destroyed in compliance with the General Data Protection Regulation 2018.

What will happen to the data collected?

Please be reassured that all personal data will be stored securely in a separate location from Facebook and will be deleted as soon as usability tests have completed. The WeightMentor chatbot has been designed to collect data in such a way as to make it difficult to identify the associated users. Personal data will be collected as follows:

1. User ID – this will only be used by the chatbot as a means of identifying the user. Users may use their first name, a nickname, or something generic (e.g. User26).
2. Age – age will be stored as a range rather than an actual age (e.g. 18-25). Users may decline to supply this information if they wish by choosing "Prefer not to say".
3. Gender – users may decline to supply this information if they wish by choosing "Prefer not to say".

Data collected will only be accessed by the research team. Please be assured that all information will be treated as confidential.

Any usability issues identified from the tests will be used to improve the chatbot. Findings may be published in research papers that will contribute to the existing body of literature by the addition of this new knowledge and will also be used to inform further research into this area.

Contact Details

If you would like further information, or to discuss participating in this project, please do not hesitate to contact either the Chief Investigator or the PhD researcher.

Chief Investigator
Dr Anne Moorhead
School of Communication & Media
(Contact Details Removed)

PhD Researcher
Samuel Holmes
School of Communication & Media
(Contact Details Removed)

Personalised Chatbot for Weight Loss Maintenance: Usability Testing Group Session
Participant Information Sheet

Background
You are being invited to participate in a research study as part of a PhD. The aim of this study is to conduct usability tests to assess the usability of WeightMentor, a custom designed chatbot for helping with weight loss maintenance. We are recruiting participants for these tests.

What is usability testing?
Usability testing is a way of finding out how easy a computer system is to use. The "computer system" in this case is the custom designed chatbot called WeightMentor. Chatbots are intelligent systems that simulate human conversation. Participants will be observed completing different tasks using this chatbot, to finds out how easy it is to use. Results of the tests can be used to make recommendations for improving the chatbot.

Who can participate in these tests?
We are inviting adults (age 18+) who are willing to use the chatbot, WeightMentor. Participants are advised that we will be testing the usability of this chatbot, rather than its usefulness for weight loss and weight loss maintenance. It is not necessary for participants to have an interest in weight loss and weight loss maintenance in order to participate. Participants do not need to have lost or maintaining weight loss to participate in this study but if they have previously lost weight they can also be included.

What is involved?
The group usability test will take place in one of the Block 16 Computer Labs at Ulster University, Jordanstown campus. Participants within a group, i.e. class setting will access the WeightMentor chatbot through Facebook Messenger on a desktop PC and complete a number of tasks to assess usability, such as creating a user profile or sending self-reported physical activity and food consumption data to the chatbot. During the tasks, participants' interactions with the chatbot will be recorded using screen capture software. Note that no faces or audio will be recorded, only the computer screen.

After the usability tests, participants will be asked to complete three questionnaires to assess the usability of the chatbot.

Note: We will be testing the system, not the participants. There is no chance of "failure". Participants will not be penalised if they are unable to complete a task.

This research project has received ethical approval from the Communication Ethics Filter Committee.

What are the benefits of taking part?
You will be part of a study that will contribute to new knowledge on the communication of personalised messages in support of weight loss maintenance. This will contribute to further research with the aim of improving support for weight loss maintenance.

What happens if I decide that I want to withdraw?

You have the right to withdraw from the study at any time without giving a reason for doing so. If you withdraw from the study, any data that has been collected from you will not be used in the research or published as part of the findings. All data that will be collected about you will be managed, stored and destroyed in compliance with the General Data Protection Regulation 2018.

What will happen to the data collected?

Please be reassured that all personal data will be stored securely in a separate location from Facebook and will be deleted as soon as usability tests have completed. The WeightMentor chatbot has been designed to collect data in such a way as to make it difficult to identify the associated users. Personal data will be collected as follows:

1. User ID – this will only be used by the chatbot as a means of identifying the user. Users may use their first name, a nickname, or something generic (e.g. User26).
2. Age – age will be stored as a range rather than an actual age (e.g. 18-25). Users may decline to supply this information if they wish by choosing "Prefer not to say".
3. Gender – users may decline to supply this information if they wish by choosing "Prefer not to say".

Data collected will only be accessed by the research team. Please be assured that all information will be treated as confidential.

Any usability issues identified from the tests will be used to improve the chatbot. Findings may be published in research papers that will contribute to the existing body of literature by the addition of this new knowledge and will also be used to inform further research into this area.

Contact Details
If you would like further information, or to discuss participating in this project, please do not hesitate to contact either the Chief Investigator or the PhD researcher.

Chief Investigator
Dr Anne Moorhead
School of Communication & Media
(Contact Details Removed)

PhD Researcher
Samuel Holmes
School of Communication & Media
(Contact Details Removed)

Appendix 17: Participant Consent Forms, *WeightMentor* Usability Testing

Consent Form

Study Title:	Personalised Chatbot for Weight Loss Maintenance: Usability Testing Individual Sessions

Chief Investigator:	Dr Anne Moorhead	PhD Student: Samuel Holmes

Please initial the boxes to confirm you have read and agree with the following statements:

1. I have been given and have read and understood the relevant information sheet for the above project and have asked and received answers to any questions raised. ☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without my rights being affected in any way. ☐
3. I understand that the researchers will hold all information and data collected during the project securely and in confidence and that all efforts will be made to ensure that I cannot be identified as a participant in the project (except as might be required by law) and I give permission for the researchers to hold relevant personal data. ☐
4. I agree to take part in this study and for data relating to me to be collected for analysis purposes only. ☐

Participant Name (Print)	Signature	Date
Name of person taking consent (If different from researcher)	Signature	Date
Researcher Name	Signature	Date

Consent Form

Study Title:	Chatbot for Weight Loss Maintenance: Usability Testing Group Session

Chief Investigator:	Dr Anne Moorhead	PhD Student: Samuel Holmes

Please initial the boxes to confirm you have read and agree with the following statements:

1. I have been given and have read and understood the relevant information sheet for the above project and have asked and received answers to any questions raised. ☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without my rights being affected in any way. ☐
3. I understand that the researchers will hold all information and data collected during the project securely and in confidence and that all efforts will be made to ensure that I cannot be identified as a participant in the project (except as might be required by law) and I give permission for the researchers to hold relevant personal data. ☐
4. I agree to take part in this study and for data relating to me to be collected for analysis purposes only. ☐

Participant Name (Print)	Signature	Date
Name of person taking consent (If different from researcher)	Signature	Date
Researcher Name	Signature	Date

Appendix 18: Pre-Test Questionnaire, *WeightMentor* Usability Testing

Please complete the following short background questionnaire. Remember that all data collected will be confidential. Please read the following questions and select your answer.

1. Please indicate your **age group:**

18 - 25 years	26 - 30 years	31 - 35 years	36 - 40 years	41 - 45 years	46 - 50 years	Over 50 years

2. Please indicate your **gender:**

Male ☐ Female ☐ Other ☐ Prefer not to say ☐

3. What is your **occupation?**

_____

4. What is your **first language?**

_____

5. Please rate your level of **technical ability using mobile devices:**

Low	1	2	3	4	5	High

6. Please rate your level of **computer literacy:**

Novice	1	2	3	4	5	Expert

7. Have you previously **used software for weight loss or weight loss maintenance**?

Yes ☐ No ☐

If yes, please give details of the software or apps you have used:

_____

8. Please specify the type of mobile devices you normally use (please list all devices used):

Apple iPhone	☐
Apple iPad	☐
Android Phone	☐
Android Tablet	☐
Windows Phone	☐
Windows Tablet	☐
Other (Please specify)	☐

_____

9. Please indicate what device you are using for this usability test?

☐	Desktop PC
☐	Mobile Device

Appendix 19: Chatbot Tasks, *WeightMentor* Usability Testing

Task 1: You have in front of you a chatbot that is designed to support weight loss maintenance. It requires initial setup before it can be used. Please interact with the chatbot to complete the setup.

Please complete this task using the chatbot and follow the prompts.

Task 2: The chatbot requires you to enter your self-reported food and activity data once per day. Interact with the chatbot and enter some self-reported data of your physical activity (enter a few examples from yesterday or today).

Please complete this task using the chatbot and follow the prompts. Please repeat this task four times.

Task 3: Interact with the chatbot and enter some self-reported data of your food consumption (enter a few examples from yesterday or today).

Please complete this task using the chatbot and follow the prompts. Please repeat this task four times.

Task 4: Scenario - Let us say it is 6pm, you are tired and hungry after a stressful day. You are tempted to indulge yourself. Interact with the chatbot and ask for motivation.

Please complete this task using the chatbot and follow the prompts. Please ask for as much motivation as you feel you would like.

Appendix 20: Single Ease Question Recording Form, *WeightMentor* Usability Testing

Participant No:

Pre-Task Question: On a scale of 1 to 7 (where 1 is very hard and 7 is very easy), how difficult do you think it will be to complete this task?

Task	Rating (1 – 7)
1	
2	
3	
4	

Post-Task Question: On a scale of 1 to 7, how difficult was this task to complete?

Task	Task Completed (Y/N)	Rating (1 – 7)
1		
2		
3		
4		

Appendix 21: Post-Test Questionnaire, *WeightMentor* Usability Testing

Please complete the following short questionnaire. Remember that all data collected will be confidential. Please read the following questions and select your answer.

How useful do you think this chatbot would be for helping to maintain weight loss?

Not useful	1	2	3	4	5	Very useful

The following 10 validated questions are the **System Usability Scale (SUS)** (Brooke, 1996)

Statements	**Rating**				
	1 **Strongly Disagree**	**2** **Disagree**	**3** **Neutral**	**4** **Agree**	**5** **Strongly Agree**
1. I think that I would like to use this system frequently					
2. I found the system unnecessarily complex					
3. I thought the system was easy to use					
4. I think that I would need the support of a technical person to be able to use this system					
5. I found the various functions in this system were well integrated					
6. I thought there was too much inconsistency in this system					
7. I would imagine that most people would learn to use this system very quickly					
8. I found the system very cumbersome to use					
9. I felt very confident using the system					
10. I needed to learn a lot of things before I could get going with this system					

Do you have any other comments you wish to add?

Thank you for completing this questionnaire.

Appendix 22: User Experience Questionnaire

For the assessment of the chatbot, please complete the following questionnaire. The questionnaire consists of pairs of contrasting attributes that may apply to the chatbot. The circles between the attributes represent gradations between the opposites. You can express your agreement with the attributes by ticking the circle that most closely reflects your impression.

Example:

attractive	○	⊗	○	○	○	○	○	unattractive

This response would mean that you rate the chatbot as more attractive than unattractive. Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression.

Sometimes you may not be completely sure about your agreement with an attribute or you may find that the attribute does not apply completely to the chatbot. Nevertheless, please tick a circle in every line.

It is your opinion that counts. Please remember there is no wrong or right answer!

Please assess the chatbot, *WeightMentor*, by selecting the appropriate rating by one line at a time.

	1	2	3	4	5	6	7	
annoying	○	○	○	○	○	○	○	enjoyable
not understandable	○	○	○	○	○	○	○	understandable
creative	○	○	○	○	○	○	○	dull
easy to learn	○	○	○	○	○	○	○	difficult to learn
valuable	○	○	○	○	○	○	○	inferior
boring	○	○	○	○	○	○	○	exciting
not interesting	○	○	○	○	○	○	○	interesting
unpredictable	○	○	○	○	○	○	○	predictable
fast	○	○	○	○	○	○	○	slow
inventive	○	○	○	○	○	○	○	conventional
obstructive	○	○	○	○	○	○	○	supportive
good	○	○	○	○	○	○	○	bad
complicated	○	○	○	○	○	○	○	easy
unlikeable	○	○	○	○	○	○	○	pleasing

	1	2	3	4	5	6	7	
usual	○	○	○	○	○	○	○	leading edge
unpleasant	○	○	○	○	○	○	○	pleasant
secure	○	○	○	○	○	○	○	not secure
motivating	○	○	○	○	○	○	○	demotivating
meets expectations	○	○	○	○	○	○	○	does not meet expectations
inefficient	○	○	○	○	○	○	○	efficient
clear	○	○	○	○	○	○	○	confusing
impractical	○	○	○	○	○	○	○	practical
organized	○	○	○	○	○	○	○	cluttered
attractive	○	○	○	○	○	○	○	unattractive
friendly	○	○	○	○	○	○	○	unfriendly
conservative	○	○	○	○	○	○	○	innovative

(Laugwitz et al, 2008)

Appendix 23: Chatbot Usability Questionnaire

Please complete this questionnaire by reading each statement carefully and placing a tick (✓) or a cross (✗) in the circle that best matches how you feel about the statement. Remember that there are no right or wrong answers!

	Strongly Disagree 1	**Disagree 2**	**Neutral 3**	**Agree 4**	**Strongly Agree 5**
The chatbot's personality was realistic and engaging	○	○	○	○	○
The chatbot seemed too robotic	○	○	○	○	○
The chatbot was welcoming during initial setup	○	○	○	○	○
The chatbot seemed very unfriendly	○	○	○	○	○
The chatbot explained its scope and purpose well	○	○	○	○	○
The chatbot gave no indication as to its purpose	○	○	○	○	○
The chatbot was easy to navigate	○	○	○	○	○
It would be easy to get confused when using the chatbot	○	○	○	○	○
The chatbot understood me well	○	○	○	○	○
The chatbot failed to recognise a lot of my inputs	○	○	○	○	○
Chatbot responses were useful, appropriate and informative	○	○	○	○	○
Chatbot responses were irrelevant	○	○	○	○	○
The chatbot coped well with any errors or mistakes	○	○	○	○	○
The chatbot seemed unable to handle any errors	○	○	○	○	○
The chatbot was very easy to use	○	○	○	○	○
The chatbot was very complex	○	○	○	○	○

Appendix 24: Individual SUS Scores per Participant, *WeightMentor* Usability Testing

Participant	Question										SUS Score
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
1	3	1	5	1	4	2	4	2	5	1	85.00
2	3	4	3	2	4	3	4	3	2	1	57.50
3	2	2	4	1	4	3	4	3	3	1	67.50
4	2	1	5	1	3	4	5	1	5	1	80.00
5	4	1	5	1	5	1	5	2	5	1	95.00
6	4	2	5	1	5	3	5	2	5	1	87.50
7	4	1	5	1	5	1	5	1	5	1	97.50
8	4	1	5	1	5	1	5	1	5	1	97.50
9	4	1	5	1	4	1	5	1	5	1	95.00
10	2	1	5	1	3	1	3	1	5	1	82.50
11	4	1	5	1	4	1	4	2	5	1	90.00
12	4	2	5	2	5	2	5	1	5	1	90.00
13	4	2	5	5	4	2	4	1	1	1	67.50
14	3	1	5	1	4	1	5	1	5	1	92.50
15	5	1	5	1	5	1	5	1	4	2	95.00
16	5	1	5	1	5	1	4	1	5	1	97.50
17	2	2	4	1	3	2	3	2	4	2	67.50
18	4	1	4	1	5	1	5	1	5	1	95.00
19	4	1	5	1	5	1	5	1	5	1	97.50
20	4	1	5	1	3	1	5	2	4	2	85.00
21	5	1	5	1	5	1	5	1	5	1	100.00
22	4	2	5	2	3	2	4	2	2	2	70.00
23	4	2	4	1	3	2	5	2	5	1	82.50
24	4	2	3	1	4	2	4	2	4	1	77.50
25	4	2	4	1	3	1	4	2	3	1	77.50
26	4	1	4	1	4	1	5	2	5	1	90.00
27	3	2	5	1	4	1	4	1	4	3	80.00
28	2	2	4	2	3	3	4	2	3	2	62.50

Participant	Question										SUS Score
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
29	5	1	5	1	5	1	5	1	5	1	100.00
30	4	2	4	1	4	2	4	1	4	1	82.50
31	2	2	3	1	3	5	3	4	2	1	50.00
32	3	2	4	1	4	2	5	3	4	1	77.50
33	4	2	4	3	3	1	4	3	4	3	67.50
34	2	1	5	3	3	1	3	3	5	5	62.50
35	2	4	4	1	3	4	3	3	3	1	55.00
36	2	3	4	4	3	4	2	3	3	3	42.50
37	1	1	5	1	5	1	5	3	5	1	85.00
38	4	3	3	4	3	2	4	2	2	3	55.00
39	3	2	4	1	4	2	5	4	2	2	67.50
40	3	1	5	1	3	2	5	4	4	1	77.50
41	3	1	5	1	5	1	5	1	5	1	95.00
42	4	2	5	1	3	2	5	3	2	2	72.50
43	3	2	5	1	4	3	5	4	5	1	77.50
44	2	2	4	3	3	2	5	2	4	2	67.50
45	3	2	4	4	3	2	2	3	3	4	50.00
46	3	1	5	1	5	1	5	1	5	1	95.00
47	2	2	5	1	4	2	5	1	4	1	82.50
48	2	4	3	2	2	4	3	3	3	2	45.00
49	3	1	5	1	3	3	5	2	5	1	82.50
50	3	4	4	3	2	3	5	5	5	5	47.50

Appendix 25: Individual CUQ Scores per Participant, *WeightMentor* Usability Testing

P	Question																CUQ Score
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	
1	4	4	5	1	4	4	5	1	4	1	4	1	3	2	5	1	79.7
2	4	2	3	1	2	4	4	4	3	2	5	1	2	3	4	2	62.5
3	4	3	4	2	3	4	4	3	4	2	4	2	3	3	4	3	62.5
4	4	4	5	1	2	2	5	2	3	2	3	2	2	3	5	1	68.8
5	4	2	5	1	5	1	5	2	5	1	5	1	3	3	5	1	89.1
6	4	4	5	2	4	1	5	1	4	4	5	2	3	2	5	4	73.4
7	4	2	5	1	5	1	5	1	4	1	4	1	5	1	5	1	93.8
8	5	2	5	1	5	1	5	1	5	1	5	1	3	2	4	2	90.6
9	4	3	4	1	3	2	4	1	4	1	5	1	3	3	5	2	78.1
10	3	2	4	1	4	1	5	3	4	1	4	1	3	3	5	3	76.6
11	4	2	4	2	4	1	5	1	4	2	4	2	3	3	4	3	75.0
12	4	3	4	1	5	1	5	2	4	1	4	2	4	3	4	2	79.7
13	4	2	4	1	5	1	1	1	4	2	4	2	4	2	5	3	76.6
14	4	2	4	1	4	3	5	1	4	1	4	1	3	3	5	1	81.3
15	4	2	4	1	4	2	4	2	4	2	4	2	3	3	4	2	73.4
16	5	3	5	1	5	1	5	1	5	1	5	1	3	3	5	1	90.6
17	3	2	4	1	2	4	4	2	3	4	2	4	3	3	4	2	54.7
18	4	2	4	2	2	3	1	1	3	1	4	2	3	3	5	2	65.6
19	4	2	4	1	4	1	5	1	4	1	4	1	3	3	5	1	84.4
20	4	2	5	1	4	1	5	2	4	2	4	1	3	3	4	4	76.6
21	5	1	4	1	4	1	5	1	5	1	4	1	3	2	5	1	90.6
22	2	4	4	1	2	4	4	2	4	1	3	2	3	3	5	1	64.1
23	4	2	4	2	2	4	5	1	4	2	3	2	4	2	5	2	71.9
24	4	2	5	1	5	1	3	1	5	1	5	1	2	3	4	3	81.3
25	4	3	4	1	4	2	4	3	4	2	4	2	4	2	3	2	71.9
26	4	2	4	1	4	3	4	2	4	1	4	2	3	1	5	2	78.1
27	4	3	4	1	4	2	4	2	4	2	4	2	3	3	5	3	71.9
28	4	4	4	1	2	4	4	2	2	4	2	4	2	4	4	2	48.4

P	Question																CUQ Score
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	
29	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	100.0
30	4	4	5	1	4	2	4	1	4	2	4	2	3	3	4	1	75.0
31	2	2	4	4	4	4	4	2	4	2	3	3	1	5	4	3	51.6
32	5	1	4	1	4	1	5	1	3	3	3	3	3	2	5	3	76.6
33	4	2	5	1	5	1	4	4	4	2	4	2	4	2	4	2	78.5
34	3	4	3	3	3	4	4	3	4	4	4	4	4	4	4	4	48.4
35	4	5	5	1	3	2	4	4	3	5	4	1	3	4	3	4	54.7
36	2	4	2	3	4	2	2	4	2	2	4	2	1	2	3	2	48.4
37	4	3	4	1	4	2	4	2	4	2	4	2	3	3	4	2	72.2
38	3	2	4	4	4	2	4	2	4	2	4	2	4	2	4	2	70.3
39	1	5	4	2	4	3	4	4	4	2	4	2	4	2	4	2	60.9
40	1	4	3	2	1	3	3	2	2	3	1	4	5	1	5	1	51.6
41	4	3	5	1	5	1	5	1	5	1	4	1	3	2	5	1	89.1
42	4	3	4	1	3	1	3	2	4	2	3	2	2	3	4	3	65.8
43	5	4	5	1	2	3	5	1	3	1	4	4	4	2	5	5	68.8
44	5	1	5	1	5	1	5	2	4	1	5	1	4	1	5	2	93.8
45	4	4	4	1	3	4	4	3	2	4	3	4	4	2	4	2	56.3
46	3	4	5	1	5	2	5	1	5	1	5	1	5	1	5	1	90.6
47	3	4	5	2	4	2	4	3	3	2	2	4	2	4	4	1	57.8
48	2	3	1	3	2	4	3	4	2	3	4	2	1	4	4	3	39.1
49	4	4	5	1	4	1	5	3	4	1	4	2	4	1	5	5	76.1
50	3	5	5	1	4	2	4	1	4	1	4	3	3	2	4	2	71.9

Appendix 26: Participant Recruitment Email, CUQ Validation Study

Are you interested in technology research?

Researchers from the School of Computing and School of Communication & Media are validating a new tool for measuring the usability of chatbots. Chatbots are intelligent systems that simulate conversation with a human being and are a relatively new technology.

If you consent to participate in this validation study, you will be required to use three different chatbots, spending a few minutes using each one, to see what it does and what sort of functions it has. You will then complete a Chatbot Usability Questionnaire (16 questions) for each one (3 questionnaires in total). The study will take no longer than 30-35 minutes to complete.

If you are interested in taking part, please visit the study website at http://scm.ulster.ac.uk/~B00625757/cuq/. There you can download a copy of the participant information sheet and sign the consent form.

If you have any questions, please contact the PhD Researcher (Samuel Holmes) at holmes-w@ulster.ac.uk

This research study has received ethical approval from the Communication Filter Committee.

Appendix 27: Participant Information Sheet, CUQ Validation Study

Validation of the Chatbot Usability Questionnaire Participant Information Sheet

Background

You are being invited to participate in a research study as part of a PhD. The aim of this study is to validate the Chatbot Usability Questionnaire (CUQ), a custom questionnaire for assessing the usability of chatbots. We are recruiting participants for this validation.

What is the CUQ and why do we need it?

Usability is a measure of how easy a computer system is to use. There are numerous validated questionnaires for measuring system usability, however while these are effective for measuring the usability of computer software and websites, they may not be as effective for measuring the usability of chatbots. Chatbots, intelligent systems that simulate human conversation, are a fairly new technology. The CUQ is designed to measure usability of chatbots by looking at aspects of usability that are unique to chatbots (such as intelligence, personality, etc.)

Who can participate in this study?

We are inviting adults (aged 18+) who are willing to use three different chatbots to validate the Chatbot Usability Questionnaire. You will be asked to provide online consent to participate in this study.

What is involved?

If you decide to participate in this study, you will be provided with weblinks and will access and complete this study online at a time and place of your convenience. You will use three different chatbots which have been rated as “good”, “average” and “poor” by a panel of academics from the School of Computing. You would spend approximately five to eight minutes interacting with each chatbot and exploring the different functions but will not know the rating of the chatbot you are using. After using each chatbot, you would complete three different usability questionnaires, one of which will be the CUQ.

Approximately two to three weeks later you will be contacted again and asked to repeat this procedure, spending approximately five to eight minutes interacting with the same three chatbots and completing the same three usability questionnaires as before. Repeating the procedure like this will determine how your questionnaire scores have changed.

What are the benefits of taking part?

You will be part of a project that will contribute to new knowledge on assessing the usability of conversational user interfaces. This will contribute to further research with the aim of improving procedures for testing this type of system.

This research project has received ethical approval from the Communication Filter Committee.

What happens if I decide that I want to withdraw?

You have the right to withdraw from the study at any time without giving a reason for doing so. If you withdraw from the study, any data that has been collected from you will not be used in the research or published as part of the findings. All data that will be collected about you will be managed, stored and destroyed in compliance with the General Data Protection Regulation 2018.

What will happen to the data collected?

Data collected will only be accessed by the research team. Please be assured that all information will be anonymous and treated as confidential.
Findings will be used to determine the validity of the CUQ and may also be published in research papers that will contribute to the existing body of literature by the addition of this new knowledge and will also be used to inform further research into this area.

Contact Details

If you would like further information, or to discuss participating in this project, please do not hesitate to contact either the Chief Investigator or the PhD researcher.

Chief Investigator	PhD Researcher
Dr Anne Moorhead	Samuel Holmes
School of Communication & Media	School of Communication & Media
(Contact Details Removed)	(Contact Details Removed)

Project Team

Samuel Holmes (School of Communication & Media)
Dr Anne Moorhead (School of Communication & Media/Institute of Nursing & Health Research)
Dr Raymond Bond (School of Computing)
Professor Huiru (Jane) Zheng (School of Computing)
Professor Vivien Coates (School of Nursing/Institute of Nursing and Health Research)

Appendix 28: Consent Form, CUQ Validation Study

Consent Form

(This is a text version of the form, which participants completed online via Qualtrics)

Study Title:	Validation of the Chatbot Usability Questionnaire

Chief Investigator:	Dr Anne Moorhead	PhD Student: Samuel Holmes

Please initial the boxes to confirm you have read and agree with the following statements:

1. I have been given and have read and understood the relevant information sheet for the above project and have asked and received answers to any questions raised. ☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without my rights being affected in any way. ☐
3. I understand that the researchers will hold all information and data collected during the project securely and in confidence and that all efforts will be made to ensure that I cannot be identified as a participant in the project (except as might be required by law) and I give permission for the researchers to hold relevant personal data. ☐
4. I agree to take part in this study and for data relating to me to be collected for analysis purposes only. ☐

Participant Name (Print)	Signature	Date
Name of person taking consent (If different from researcher)	Signature	Date
Researcher Name	Signature	Date

Appendix 29: Participant Demographic Questionnaire, CUQ Validation Study

(This is a text version of the questionnaire, which participants completed online via Qualtrics)

Please enter your Participant ID (This will have been given to you when you signed your consent form:

Please indicate **which of the following age groups you fit into:**

18 - 25 years	26 - 30 years	31 - 40 years	41 - 45 years	46 - 50 years	Over 50 years

1. Please indicate your **gender:**

Male ☐ Female ☐ Other ☐ Prefer not to say ☐

2. What is your **occupation?**

_____

3. What is your **first language?**

_____

4. Please rate your level of **technical ability:**

Low	1	2	3	4	5	High

Appendix 30: Mean CUQ Question scores per round, with p-values and correlations

WoeBot				
Question	**Round 1 Mean Score**	**Round 2 Mean Score**	**T-Test**	**Correlation**
1	3.8±1.0	3.8±0.9	0.83	0.58
2	2.5±1.0	2.5±0.9	1.0	0.68
3	4.2±0.8	4.2±0.6	0.81	0.52
4	1.6±0.7	1.6±0.6	1.0	0.59
5	4.2±0.7	4.1±0.7	0.56	0.55
6	2.0±1.0	1.8±0.5	0.33	0.48
7	4.1±1.1	4.2±0.8	0.51	0.55
8	1.9±0.9	1.9±0.8	0.76	0.71
9	3.6±1.1	3.9±0.7	0.08	0.67
10	2.1±1.0	2.0±0.6	0.54	0.40
11	3.8±1.0	4.0±0.8	0.28	0.75
12	2.0±1.0	1.9±0.6	0.46	0.55
13	3.2±0.8	3.4±0.7	0.49	0.32
14	2.5±0.9	2.3±0.8	0.31	0.35
15	4.2±1.0	4.2±1.0	0.70	0.77
16	2.1±1.1	1.8±0.9	0.06	0.75

Weight Loss Bot				
Question	**Round 1 Mean Score**	**Round 2 Mean Score**	**T-Test**	**Correlation**
1	3.7±0.9	3.6±1	0.29	0.53
2	2.7±1.1	2.6±0.9	0.65	0.62
3	4.0±0.9	4.0±0.7	1.00	0.54
4	2.2±1.1	1.8±0.6	0.15	0.27
5	4.2±0.9	4.0±0.7	0.08	0.49
6	1.7±0.8	2.0±0.8	0.13	0.34
7	3.9±1.0	3.7±1.2	0.22	0.51
8	2.2±0.9	2.6±1.1	0.02	0.56

Weight Loss Bot				
Question	**Round 1 Mean Score**	**Round 2 Mean Score**	**T-Test**	**Correlation**
9	3.5±0.9	3.4±0.9	0.42	0.66
10	2.0±1.1	2.4±1.0	0.12	0.41
11	3.9±0.8	3.6±1.0	0.18	0.53
12	2.0±0.8	2.0±0.8	0.49	0.79
13	3.0±1.0	3.1±0.8	0.33	0.53
14	2.9±1.1	2.8±1.1	0.86	0.35
15	3.9±1.0	3.7±1.1	0.54	0.60
16	2.2±1.1	2.3±1.2	0.38	0.73

FlyBot				
Question	**Round 1 Mean Score**	**Round 2 Mean Score**	**T-Test**	**Correlation**
1	2.8±1.1	2.6±1.3	0.47	0.72
2	3.7±1.1	3.5±1.1	0.32	0.60
3	3.0±1.2	3.1±1.0	0.63	0.64
4	2.4±1.0	2.6±1.0	0.34	0.63
5	2.7±1.1	2.8±1.2	0.37	0.63
6	2.6±1.2	2.4±1.0	0.18	0.84
7	3.9±0.8	3.9±0.9	0.82	0.50
8	2.4±1.0	2.2±1.0	0.50	0.45
9	3.3±1.1	3.4±0.8	0.54	0.44
10	2.4±1.0	2.4±1.0	0.68	0.33
11	2.9±1.1	3.2±1.0	0.07	0.67
12	2.2±0.8	2.3±0.8	0.60	0.40
13	2.6±1.1	2.8±1.1	0.18	0.55
14	3.2±1.2	2.9±1.2	0.27	0.26
15	4.1±0.6	4.0±0.9	0.63	0.51
16	1.7±0.7	1.6±0.7	0.50	0.48

Appendix 31: Conference & Journal Abstracts

Impact of digital technologies for communicating messages on weight loss maintenance: a systematic literature review

(Published in European Journal of Public Health, 18th September 2018 and April 2019)

Background

Research into the use of digital technology for weight loss maintenance (intentionally losing at least 10% of initial body weight and actively maintaining it) is limited. The aim of this article was to systematically review randomized controlled trials (RCTs) reporting on the use of digital technologies for communicating on weight loss maintenance to determine its effectiveness and identify gaps and areas for further research.

Methods

A systematic literature review was conducted by searching electronic databases to locate publications dated between 2006 and February 2018. Criteria were applied, and RCTs using digital technologies for weight loss maintenance were selected.

Results

Seven RCTs were selected from a total of 6541 hits after de-duplication and criteria applied. Three trials used text messaging, one used e-mail, one used a web-based system and two compared such a system with face-to-face contact. From the seven RCTs, one included children (n = 141) and reported no difference in BMI Standard Deviation between groups. From the seven trials, four reported that technology is effective for significantly aiding weight loss maintenance compared with control (no contact) or face-to face-contact in the short term (between 3 and 24 months).

Conclusions

It was concluded that digital technologies have the potential to be effective communication tools for significantly aiding weight loss maintenance, especially in the short term (from 3 to 24 months). Further research is required into the long-term effectiveness of contemporary technologies.

Digital communication tools in support of weight loss maintenance: A systematic literature review

(Presented at the 3rd Annual Public Health PhD Symposium, Liverpool John Moore's University, July 2017)

Background

Obesity and overweight present health risks[1]. Individuals who successfully lose weight experience the bigger challenge of maintaining this weight loss. Research into the effectiveness of technologies for weight loss maintenance is limited[2]. The aim of this research was to conduct a systematic literature review of previous research into the use of digital weight loss maintenance technologies.

Methods

A systematic literature review was conducted according to PRISMA guidelines[3]. The MEDLINE, EMBASE and PUBMED databases were searched, returning 3000 results. These were screened using inclusion and exclusion criteria. Publications included dated from 2006 to 2017 focusing on weight loss maintenance where the intervention primarily used digital health technologies. Twenty-eight papers were screened, six of which were randomized controlled trials (RCTs).

Results

From the six RCTs focusing on weight loss maintenance, four used text-messaging, one was e-mail based, and one used multiple technologies. Only one study did not show positive effects for body mass index and body weight, and another reported positive effects on depression but not anxiety. Younger people were found to engage with newer technologies. Age was not determined to be a barrier for communication technologies. It was reported that communication can be improved through personalising internet-based interventions[4].

Discussion

It was concluded that digital technology interventions were feasible for weight loss maintenance and showed positive short-term outcomes. Further research is needed into their effectiveness over longer periods.

Implications

Digital technologies provide opportunities for tailored messaging for weight loss maintenance and can effectively reach a diverse range of the population.

Digital technologies for communicating messages on weight loss maintenance: A systematic literature review

(Presented at the EACH (International Association for Communication in Healthcare) Summer Event, London, September 2017)

Problem Statement: Obesity and overweight pose significant risks to health1. Many individuals find it very difficult to lose weight. Even if weight loss is achieved, it is an even bigger challenge to maintain and therefore weight regain is common. This provides opportunities for effective communication of motivational messaging to retain weight loss. There is limited research into the effectiveness of such technologies for weight loss maintenance2. The aim was to conduct a systematic literature review of previous research on the use of digital technologies for weight loss maintenance.

Methods: A systematic literature review was conducted in November 2016 as per PRISMA guidelines3. Three databases were searched – MEDLINE, EMBASE and PUBMED. Inclusion and exclusion criteria were applied, which included publications dated from 2006 to 2017 focusing on obesity, specifically weight loss maintenance, where digital health technologies are the primary focus of the intervention. Initially, 3000 papers were retrieved and following screening and applying the criteria, 30 papers were identified, of which four were randomised controlled trials (RCTs). A further three RCTs were identified in June 2017

Results: From the seven RCTs, three used text messaging to send motivational messages, and one used e-mail. Two used a web-based portal, and one used a web-based portal plus Electronic Health Record tools. All except for one showed positive effects for outcome measures. The one study that assessed both anxiety and depression showed positive effects for depression but not anxiety. It was found that newer technologies are popular with younger people and so could be a feasible intervention for this demographic. Age is not a barrier for communication technologies. It was reported that personalising internet-based interventions will improve communication.

Discussion: It was concluded that while in general text message or email-based interventions were cost-effective and feasible for delivering messages on weight loss maintenance that showed favourable outcomes in the short term, further research is needed into their effectiveness over a longer period.

Implications: Digital technologies are useful for communicating on weight loss maintenance as they provide opportunities for personalised and tailored messaging and to reach a diverse range of the population.

WeightMentor: A new automated chatbot for weight loss maintenance

(Presented at the 32nd Human Computing Interaction (HCI) conference, Belfast, July 2018)

Obesity and Overweight are significant risk factors for many chronic conditions, such as type 2 diabetes. Weight loss is difficult and maintaining weight loss is a greater challenge. Effective communication is hindered by conflicting information and the sensitivity of the subject of obesity. Recent technological solutions for weight loss maintenance are limited. A chatbot would be appropriate for supporting weight loss as it does not require a human operator, is available 24 hours a day, and supports natural communication while maintaining anonymity. Such a system may also be integrated with popular social media platforms such as Facebook Messenger. This paper describes the design and development of the WeightMentor chatbot, a self-help motivational tool for weight loss maintenance. Chatbots may have the potential to contribute to obesity prevention and management.

Digital technologies for communicating on weight loss maintenance: A systematic literature review

(Presented at the EACH (International Association for Communication in Healthcare) International Conference on Communication in Healthcare (ICCH), Portugal, September 2018)

Background: Obesity and overweight are recognised globally as significant health risks. Weight loss maintenance is a challenge but may be improved by effective communication of motivational messages. Considerable research exists into the use of digital health technologies for weight management and healthy living, however research into their use for weight loss maintenance is limited. The aim was to systematically review previous randomised controlled trials (RCTs) reporting on the use of digital health technologies for weight loss maintenance, to determine the effectiveness of the technology and identify gaps and areas for further research.

Methods: A systematic literature review was conducted, which followed the PRISMA guidelines. A systematic search was conducted using electronic databases to locate publications dated between 2006 and 2017. The search was designed to focus on the key concepts of "Digital Technologies" and "Weight Loss Maintenance." Inclusion and exclusion criteria were applied, including publications focusing on weight loss maintenance where digital technology was used for intervention delivery.

Findings: From 52 papers identified, seven reported on RCTS. There were six trials with adult participants who had successfully lost weight. One other trial included children who engaged in a behaviour change intervention. Three trials used text messaging, one used e-mail, one used a web-based system and two compared web-based systems with face-to-face contact. The trial involving children reported no difference in outcome measures between groups. Overall, the trials using text messaging showed some significant benefits for maintenance of positive lifestyle changes. E-mail trial intervention participants maintained a greater average weight loss throughout than those in the control group. Face-to-face contact interventions yielded greater long-term benefits for weight loss maintenance and resulted in lower mean weight regain than web-based systems.

Discussion: Digital health technologies have the potential to feasibly deliver effective weight loss maintenance interventions, at least in the short term, but were observed to be slightly less effective when compared with face-to-face interventions. Further research is required into long-term effectiveness of contemporary technologies, efficiency of technology-based interventions, and more recent technological advances for weight loss maintenance.

Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?

(Presented at the 31st Annual European Conference on Cognitive Ergonomics (ECCE), Belfast, September 2019)

Chatbots are becoming increasingly popular as a human-computer interface. The traditional best practices normally applied to User Experience (UX) design cannot easily be applied to chatbots, nor can conventional usability testing techniques guarantee accuracy. WeightMentor is a bespoke self-help motivational tool for weight loss maintenance. This study addresses the following four research questions: How usable is the WeightMentor chatbot, according to conventional usability methods?; To what extend will different conventional usability questionnaires correlate when evaluating chatbot usability?; And how do they correlate to a tailored chatbot usability survey score?; What is the optimum number of users required to identify chatbot usability issues?; How many task repetitions are required for a first-time chatbot users to reach optimum task performance (i.e. efficiency based on task completion times)? This paper describes the procedure for testing the WeightMentor chatbot, assesses correlation between typical usability testing metrics, and suggests that conventional wisdom on participant numbers for identifying usability issues may not apply to chatbots. The study design was a usability study. WeightMentor was tested using a pre-determined usability testing protocol, evaluating ease of task completion, unique usability errors and participant opinions on the chatbot (collected using usability questionnaires). WeightMentor usability scores were generally high, and correlation between questionnaires was strong. The optimum number of users for identifying chatbot usability errors was 26, which challenges previous research. Chatbot users reached optimum proficiency in tasks after just one repetition. Usability test outcomes confirm what is already known about chatbots - that they are highly usable (due to their simple interface and conversation-driven functionality) but conventional methods for assessing usability and user experience may not be as accurate when applied to chatbots.

Communication tool for weight loss maintenance: Chatbot, WeightMentor - needs analysis & development

(Presented at the EACH (International Association for Communication in Healthcare) Forum on Healthcare Communication, Netherlands, September 2019)

Background

Obesity and overweight are significant health risks. Weight loss maintenance may be improved by effective communication. While text messaging is reportedly useful for self-reporting and motivation, users cannot expect replies immediately, and automation of replies makes messages less engaging. As conversation-driven interfaces, chatbots may lead to higher compliance with health interventions because they can react quickly to human emotions and create empathy. The aim of this study was to determine the needs of individuals who are maintaining weight loss in order to design and develop WeightMentor, a weight loss maintenance chatbot.

Methods

This was a needs assessment and technology (chatbot) development study. It included qualitative interviews, which were conducted with fifteen participants, recruited through the University. Participants were asked to identify weight loss and maintenance goals, challenges and solutions, comment on use of apps and types and tones of personal messages. Interviews were transcribed and analysed using thematic analysis. Based on the findings of the needs assessment, a chatbot, WeightMentor, was designed and developed. Ethical approval was obtained by the University Research Ethics Committee.

Findings

From the needs assessment, five key themes were identified: (1) Weight loss maintenance is challenging; (2) Social contact is beneficial but may also reinforce unhealthy habits; (3) Apps should be convenient and support progress tracking; (4) Personal messages should be specific and relevant; (5) Chatbots are potentially useful but interactions should be minimal.

Chatbot, WeightMentor, was designed and developed based on findings from the needs assessment and a review of the most popular nutrition apps. WeightMentor operates within Facebook messenger, which is currently one of the most popular social media platforms. WeightMentor's purpose is to provide a source of self-help and motivation for maintaining weight loss. Its functions include self-reporting of diet and physical activity, viewing previous self-reporting trends, and motivational messaging. Further research is required to test the usability and effectiveness of this WeightMentor.

Discussion

Chatbots are a valuable communication tool for supporting and motivating individuals. This newly developed chatbot, WeightMentor, has the potential to provide motivational messages to support individuals who are maintaining weight loss.

WeightMentor: A bespoke chatbot for Weight Loss Maintenance

(Presented at the 13th European Nutrition Conference, Federation of European Nutrition Societies (FENS), Dublin, October 2019)

Introduction: Recently, there has been significant research into the use of technology for weight loss, but research into technology for weight loss maintenance is limited. To date, social media and communication platforms are the most popular. Chatbots are intelligent systems, designed to simulate human conversation. Their user interface is often very simple, and may be integrated with existing communication tools such as Facebook, eliminating the need to download and learn a new app. Chatbots can mimic the variable nature of human conversation, so interacting with them seems natural and personal. The aim of this research study was to design, develop and test a chatbot to aid in weight loss maintenance.

Materials & Methods: This study design involved development and testing of a chatbot. Based from our previous research from needs assessments using interviews with adults who have lost weight and the literature, a Chatbot was designed to be used on a smartphone or tablet. This intervention, the chatbot, WeightMentor, is a Facebook Messenger based self-help motivational chatbot for weight loss maintenance. Usability testing was conducted among 50 adults, who were video and audio recorded attempting four tasks using the chatbot (user profile creation, self-reporting physical activity (4 times), self-reporting food intake (4 times), and motivation). After the tasks, participants evaluated the chatbot using three different usability questionnaires.

Results: This bespoke chatbot for weight loss maintenance, WeightMentor, consisted of the following functions: user profile creation, self-reporting of food and physical activity, progress tracking and reviewing, and motivation. Varied greetings and responses, animated graphics, and storage and use of the user's name create an interesting user experience and improve user engagement. From the usability testing, WeightMentor was ranked well above average, suggesting that participants found the chatbot highly usable and acceptable. During repetitions of tasks, participants achieved optimum performance after just one attempt, suggesting that proficiency may be reached very quickly with the WeightMentor chatbot.

Discussion: WeightMentor has been shown to be easy to use and acceptable to users. It has potential as a self-help motivational tool, providing a friendly, lightweight interface conveniently accessible from a smartphone. Further testing would determine feasibility for delivering structured weight loss maintenance interventions. Nutritionists and Dietitians could make their clients and patients aware of such technologies, such as this WeightMentor, to assist them in weight loss maintenance, as part of obesity prevention and management.

References

Abraham A. A. Chow W. C. So H. K. Woo J. Chan S. M. and Nelson E. A. S. (2014) Feasibility of Using Cell Phone Reminders to Motivate Behaviour Change in Obese Adolescents in Hong Kong. In: Hong Kong Journal of Paediatrics. Conference: Joint Annual Scientific Meeting on Hong Kong Paediatric Society and Hong Kong Paediatric Nurses Association 2014. United Kingdom. Conference Publication: (pp. 192-206). 19 (3) (pp 197),

Aguilar-Martinez, A. Sole-Sedeno, J.M. Mancebo-Moreno, G. Medina, F.X. Carreras-Collado, R. and Saigi-Rubio, F. (2014) Use of mobile phones as a tool for weight loss: a systematic review. Journal of Telemedicine & Telecare, 20 (6), 339-349.

Arbitron & Edison Research (2013) Mums and Media. Arbitron & Edison Research. Available from: http:// www.slideshare.net/webby2001/moms-and-media-2012

Baki Kocaballi, A. Laranjo, L. and Coiera, E. (2018) Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires. In: Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI 2018). 4th - 6th July, Belfast.

Bandura, A. (1977) Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 84 (2), 191-215.

Bangor, A. Kortum, P.T. and Miller, J.T. (2009) Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. Journal of Usability Studies, 4 (3), 114-123.

Bangor, A. Kortum, P.T. and Miller, J.T. (2008) An Empirical Evaluation of the System Usability Scale. International Journal of Human Computer Interaction, 24 (6), 574-594.

Barak, A. Klein, B. and Proudfoot, J.G. (2009) Defining Internet-Supported Therapeutic Interventions. Annals of Behavioral Medicine, 38 (1), 4-17.

Baranowski, T. Buday, R. Thompson, D.I. and Baranowski, J. (2008) Playing for Real: Video Games and Stories for Health-Related Behavior Change. American Journal of Preventive Medicine, 34 (1), 74-82. e10.

Baranowski, T. and Frankel, L. (2012) Let's get technical! Gaming and technology for weight control and health promotion in children. Childhood Obesity, 8 (1), 34-37. Barnett, A. Cerin, E.F. and Baranowski, T. (1107) Active video games for youth: a systematic review.

Bauer, S. de Niet, J. Timman, R. and Kordy, H. (2010) Enhancement of care through self-monitoring and tailored feedback via text messaging and their use in the treatment of childhood overweight. Patient Education & Counseling, 79 (3), 315-319.

Bellamy, N. (2015) Principles of clinical outcome assessment. In: Hochberg, M.C. Silman, A.J. Smolen, J.S. Weinblatt, M.E. and Weisman, M.H. eds. Rheumatology. Sixth Edition ed. Philadelphia: Mosby, 9-19.

Bennett G.G. and Glasgow, R.E. (2009) The delivery of public health interventions via the internet: Actualizing their potential. Annual Review of Public Health, 30 273-292.

Bickmore, T.W. Trinh, H. Olafsson, S. O'Leary, T.K. Asadi, R. Rickles, N.M. and Cruz, R. (2018) Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant. Journal of medical Internet research, 20 (9), e11510.

Brindal, E. Hendrie, G. Freyne, J. Coombe, M. Berkovsky, S. and Noakes, M. (2013) Design and pilot results of a mobile phone weight-loss application for women starting a meal replacement programme. Journal of telemedicine and telecare, 19 (3), 166-174.

Brindal, E. Hendrie, G.A. and Freyne, J. (2016) Combining Persuasive Technology with Behavioral Theory to Support Weight Maintenance Through a Mobile Phone App: Protocol for the MotiMate App. JMIR Research Protocols, 5 (1), e5.

Brooke, J. (1996) SUS: A 'quick and dirty' usability scale. In: Jordan, P.W. Thomas, B. McClelland, I.L. and Weerdmeester, B. eds. Usability Evaluation in Industry. CRC Press, 189-194.

Budiu, R. (2018) The User Experience of Chatbots. California: Nielsen Norman Group. Available from: https://www.nngroup.com/articles/chatbots/ [Accessed March 2019].

Burke L.E. Styn M.A. Conroy M.B. Ye L. Glanz K. Sevick M.A. Ewing L.J. and Sereika, S.M. (2012) Adherence to weight loss treatment across three self-monitoring approaches and the association with weight change. In: Circulation. Conference: Epidemiology and Prevention/Physical Activity, Nutrition and Metabolism 2012. San Deigo, CA United States. Conference Publication: (var. pagings). 125 (10 SUPPL. 1) (no pagination); Lippincott Williams and Wilkins,

Cameron, G. Cameron, D. Megaw, G. Bond, R. Mulvenna, M. O'Neill, S. Armour, C. and McTear, M. (2018) Best practices for designing chatbots in mental healthcare – A case study on iHelpr. In: Raymond Bond, Maurice Mulvenna, Jonathan Wallace and Michaela Black, eds. Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI-2018),

Carter, M.C. FAU, B.V. Nykjaer, C.F. and Cade, J.E. (1113) Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial.

Castleberry, A. and Nolen, A. (2018) Thematic analysis of qualitative research data: Is it as easy as it sounds? Available from: http://www.sciencedirect.com/science/article/pii/S1877129717300606

Cherry, K. (2019) Gardner's Theory of Multiple Intelligences. New York: Verywell Mind. Available from: https://www.verywellmind.com/gardners-theory-of-multiple-intelligences-2795161 [Accessed April 2020].

Cherry, K. (2019) How General Intelligence Influences Performance on Cognitive Tasks. New York: Verywell Mind. Available from: https://www.verywellmind.com/what-is-general-intelligence-2795210 [Accessed April 2020].

Cherry, K. (2019) Theories of Intelligence in Psychology. New York: Verywell Mind. Available from: https://www.verywellmind.com/theories-of-intelligence-2795035#theories-of-intelligence [Accessed April 2020].

Cohen, A. Perozich, A. Rajan, R. Persky, S. Parisi, J. Bowie, J. Fahle, J. Cho, J. Krishnan, A. Cohen, Z. Ezike, A. Schulte, C. Taylor, J. Storey, D. Ahmed, R.S. and Cheskin, L.J. (2017) Framed, Interactive Theory-Driven Texting: Effects of Message Framing on Health Behavior Change for Weight Loss. Family & community health, 40 (1), 43-51.

Conroy, M.B. Bryce, C.L. McTigue, K.M. Tudorascu, D. Gibbs, B.B. Comer, D. Hess, R. Huber, K. Simkin-Silverman, L.R. and Fischer, G.S. (2017) Promoting weight maintenance with electronic health record tools in a primary care setting: Baseline results from the MAINTAIN-pc trial. Contemporary clinical trials, 54 60-67.

Cussler, E.C. Teixeira, P.J. Going, S.B. Houtkooper, L.B. Metcalfe, L.L. Blew, R.M. Ricketts, J.R. Lohman, J. Stanford, V.A. and Lohman, T.G. (2008) Maintenance of weight loss in overweight middle-aged women through the Internet. Obesity (Silver Spring, Md.), 16 (5), 1052-1060.

de Niet, J. Timman, R. Bauer, S. van den Akker, E. Buijks, H. de Klerk, C. Kordy, H. and Passchier, J. (2012) The effect of a short message service maintenance treatment on body mass index and psychological well-being in overweight and obese children: a randomized controlled trial. Pediatric Obesity, 7 (3), 205-219.

Department of Health Northern Ireland (2018) Health Survey Northern Ireland: First Results 2017/18. Northern Ireland: Department of Health Northern Ireland. Available from: https://www.health-ni.gov.uk/sites/default/files/publications/health/hsni-first-results-17-18.pdf [Accessed March 2019].

Department of Health Northern Ireland (2017) Health Survey Northern Ireland: First Results 2016/2017. Northern Ireland: Department of Health Northern Ireland. Available from: https://www.health-ni.gov.uk/sites/default/files/publications/health/hsni-first-results-16-17.pdf [Accessed August 2019].

DePoy, E. and Gitlin, L.N. (2016) Collecting Data Through Measurement in Experimental-Type Research. In: DePoy, E. and Gitlin, L.N. eds. Introduction to Research: Understanding and Applying Multiple Strategies. 5th Edition ed. St. Louis: Mosby, 227-247.

Derbyshire E. and Dancey, D. (2013) Smartphone medical applications for women's health: What is the evidence-base and feedback? International Journal of Telemedicine and Applications, 2013

DeVellis, R.F. (1991) Scale Development: Theory and Applications. California: SAGE Publications.

Donaldson, E.L. Fallows, S. and Morris, M. (2014) A text message-based weight management intervention for overweight adults. Journal of Human Nutrition & Dietetics, 27 (Suppl 2), 90-97.

Dudovskiy, J. (2018) Convenience Sampling. In: Anon. The Ultimate Guide to Writing a Dissertation in Business Studies: A Step-by-Step Assistance.

Dudovskiy, J. (2018) Purposive Sampling. In: Anon. The Ultimate Guide to Writing a Dissertation in Business Studies: A Step-By-Step Assistance.

Elbert, S.P. Dijkstra, A. and Rozema, A.D. (2017) Effects of tailoring ingredients in auditory persuasive health messages on fruit and vegetable intake. Psychology & Health, 32 (7), 781-797.

Elsevier (March 2019) *Embase Coverage & Content.* Elsevier. Available from: https://www.elsevier.com/solutions/embase-biomedical-research/embase-coverage-and-content [Accessed February 2020].

Eurostat (2019) Overweight and obesity - BMI statistics. European Union: Eurostat. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php/Overweight_and_obesity_-_BMI_statistics [Accessed September 2019].

Evans, E.H. Araujo-Soares, V. Adamson, A. Batterham, A.M. Brown, H. Campbell, M. Dombrowski, S.U. Guest, A. Jackson, D. Kwasnicka, D. Ladha, K. McColl, E. Olivier, P. Rothman, A.J. Sainsbury, K. Steel, A.J. Steen, I.N. Vale, L. White, M. Wright, P. and Sniehotta, F.F. (2015) The NULevel trial of a scalable, technology-assisted weight loss maintenance intervention for obese adults after clinically significant weight loss: study protocol for a randomised controlled trial. Trials [Electronic Resource], 16 421.

Ezendam, N.P. Brug, J. and Oenema, A. (2012) Evaluation of the Web-based computer-tailored FATaintPHAT intervention to promote energy balance among adolescents: results from a school cluster randomized trial. Archives of Pediatrics & Adolescent Medicine, 166 (3), 248-255.

Fadil, A. and Gabrielli, S. (2017) Addressing Challenges in Promoting Healthy Lifestyles: The AI-Chatbot Approach. Information ed. Researchgate. Available from: https://www.researchgate.net/profile/Ahmed_Fadhil4/publication/316141692_Addressing_Challenges_in_Promoting_Healthy_Lifestyles_The_AI-Chatbot_Approach/links/596286a9458515a3574aad1d/Addressing-Challenges-in-Promoting-Healthy-Lifestyles-The-AI-Chatbot-Approach.pdf [Accessed August 2017].
Festinger, L. (1957) A theory of cognitive dissonance. Stanford: Stanford University Press.

Fitzpatrick, K.K. Darcy, A. and Vierhile, M. (2017) Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health, 4 (2), e19.

Fjeldsoe, B.S. Goode, A.D. Phongsavan, P. Bauman, A. Maher, G. Winkler, E. and Eakin, E.G. (2016) Evaluating the Maintenance of Lifestyle Changes in a Randomized

Controlled Trial of the 'Get Healthy, Stay Healthy' Program. JMIR MHealth and UHealth, 4 (2), e42.

Fjeldsoe, B.S. Marshall, A.L. and Miller, Y.D. (2009) Behavior change interventions delivered by mobile telephone short-message service. American Journal of Preventive Medicine, 36 (2), 165-173.

Foggia, L. (2018) Usability Testing: What is it and how to do it? UX Collective. Available from: https://uxdesign.cc/usability-testing-what-is-it-how-to-do-it-51356e5de5d [Accessed August 2019].

Franco, R.Z. Fallaize, R. Lovegrove, J.A. and Hwang, F. (2016) Popular Nutrition-Related Mobile Apps: A Feature Assessment. JMIR MHealth and UHealth, 4 (3), e85.
Fry, J.P. and Neff, R.A. (2009) Periodic prompts and reminders in health promotion and health behavior interventions: systematic review. Journal of Medical Internet Research, 11 (2), e16.

Funk, K.L. Stevens, V.J. Appel, L.J. Bauck, A. Brantley, P.J. Champagne, C.M. Coughlin, J. Dalcin, A.T. Harvey-Berino, J. Hollis, J.F. Jerome, G.J. Kennedy, B.M. Lien, L.F. Myers, V.H. Samuel-Hodge, C. Svetkey, L.P. and Vollmer, W.M. (2010) Associations of internet website use with weight change in a long-term weight loss maintenance program. Journal of Medical Internet Research, 12 (3), e29.

Funk, K.L. Stevens, V.J. Bauck, A. Brantley, P.J. Hornbrook, M. Jerome, G.J. Myers, V.H. and Appel, L. (2011) Development and Implementation of a Tailored Self-assessment Tool in an Internet-based Weight Loss Maintenance Program. Clinical Practice & Epidemiology in Mental Health [Electronic Resource]: CP & EMH, 7 67-73.

Gasch M. Dunleavy G.J. Kyaw B.M. Lean M.E.J. and Nikolaou, C.K. (2016) Personalized Health, eLearning, and mHealth Interventions to Improve Nutritional Status. Current Nutrition Reports, 5 (4), 295-306.

Gerber, B.S. Stolley, M.R. Thompson, A.L. Sharp, L.K. and Fitzgibbon, M.L. (2009) Mobile phone text messaging to promote healthy behaviors and weight loss maintenance: a feasibility study. Health Informatics Journal, 15 (1), 17-25.

Gnewuch, U. Morana, S. and Maedche, A. (2017) Towards Designing Cooperative and Social Conversational Agents for Customer Service. In: International Conference on Information Systems (ICIS). 10th - 13th December 2017, Seoul, South Korea. South Korea: Association for Information Systems, Available from: https://www.researchgate.net/publication/320015931_Towards_Designing_Cooperative_and_Social_Conversational_Agents_for_Customer_Service [Accessed April 2020].

Gonzalez C. Herrero P. Cubero J.M. Iniesta J.M. Hernando M.E. GarciaSaez G. Serrano A.J. MartinezSarriegui I. PerezGandia C. Gomez E.J. Rubinat E. Alcantara V. Brugues E. Chico A. Mato E. Bell O. Corcoy R. and De Leiva, A. (2013) PREDIRCAM eHealth platform for individualized telemedical assistance for lifestyle modification in the treatment of obesity, diabetes, and cardiometabolic risk prevention: A pilot study (PREDIRCAM 1). Journal of Diabetes Science and Technology, 7 (4), 888-897.

Greene, G.W. Schembre, S.M. White, A.A. Hoerr, S.L. Lohse, B. Shoff, S. Horacek, T. Riebe, D. Patterson, J. Phillips, B.W. Kattelmann, K.K. and Blissmer, B. (2011)

Identifying Clusters of College Students at Elevated Health Risk Based on Eating and Exercise Behaviors and Psychosocial Determinants of Body Weight. Available from: http://www.sciencedirect.com/science/article/pii/S000282231001970X. [Accessed August 2017]

Greene, G.W. White, A.A. Hoerr, S.L. Lohse, B. Schembre, S.M. Riebe, D. Patterson, J. Kattelmann, K.K. Shoff, S. Horacek, T. Blissmer, B. and Phillips, B.W. (2012) Impact of an Online Healthful Eating and Physical Activity Program for College Students. Am J Health Promot, 27 (2), e47-e58.

Haapala, I. Barengo, N. Biggs, S. Surakka, L. and Manninen, P. (2009) Weight loss by mobile phone: a 1-year effectiveness study. Public Health Nutr. 12 2382-2391.

Half, R. (2017) 6 Basic SDLC Methodologies: Which One Is Best? Robert Half International Inc. Available from: https://www.roberthalf.com/blog/salaries-and-skills/6-basic-sdlc-methodologies-which-one-is-best [Accessed April 2019].

Harford, T. (2019) Can a computer fool you into thinking it is human? London: BBC. Available from: https://www.bbc.co.uk/news/business-49344596 [Accessed April 2020].

Hassenzahl, M. Burmester, M. and Koller, F. (2003) AttrakDiff: A questionnaire to measure perceived hedonic and pragmatic quality. Mensch & Computer, 187-196.

Haugen, H.A. Tran, Z.V. Wyatt, H.R. Barry, M.J. and Hill, J.O. (2007) Using telehealth to increase participation in weight maintenance programs. Obesity (Silver Spring, Md.), 15 (12), 3067-3077.

Hebden, L. Cook, A. van der Ploeg, P. and Allman-Farinelli, M. (2012) Development of Smartphone applications for nutrition and physical activity behavior change. JMIR Research Protocols, 1 (e9),

Hebden, L. Balestracci, K. McGeechan, K. Denney-Wilson, E. Harris, M. Bauman, A. and Allman-Farinelli, M. (2013) 'TXT2BFiT' a mobile phone-based healthy lifestyle program for preventing unhealthy weight gain in young adults: study protocol for a randomized controlled trial. Trials [Electronic Resource], 14 75.

Hibbard, J. and Gilburt, H. (2014) Supporting people to manage their health: An introduction to patient activation. Available from: https://www.kingsfund.org.uk/sites/default/files/field/field_publication_file/supporting-people-manage-health-patient-activation-may14.pdf [Accessed October 2019].

Higgins, E.T. Friedman, R.S. Harlow, R.E. Idson, L.C. Ayduk, O.N. and Taylor, A. (2001) Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. European Journal of Social Psychology, 31 (1), 3-23.

Holden, R.R. (2010) Face Validity. In: Weiner, I.B. and Craighead, W.E. eds. The Corsini Encyclopedia of Psychology, Volume 2. New Jersey: John Wiley & Sons, 637-638.

Holmes, W.S. Moorhead, S.A. Coates, V.E. Zheng, H. and Bond, R.R. (2019) Impact of digital technologies for communicating messages on weight loss maintenance: a systematic literature review. European Journal of Public Health, 29 (2), 320-328.

Hone, K.S. and Graham, R. (2000) Towards a tool for the subjective assessment of speech system interfaces (SASSI). Natural Language Engineering, 6 (3-4), 287-303.

Jones, M. FAU, L.K. FAU, O.M. Taylor, K.F. Cunning, D.F. FAU, D.A. FAU, W.D. and Taylor, C.B. (0320) Randomized, controlled trial of an internet-facilitated intervention for reducing binge eating and overweight in adolescents.

Joo, N. and Kim, B. (2007) Mobile phone short message service messaging for behaviour modification in a community-based weight control programme in Korea. J Telemed Telecare, 13 (8), 416-420.

Kattelmann K. White A. Bredbenner C. Greene G. Horacek T. Kidd T. Phillips B. Colby S. Brown O. Hoerr S. Shelnutt K. Olfert M. Koenings M. and Stabile Morrell, J. (2013) Project yeah: Development of a web-delivered theory-based intervention for preventing excess weight gain in young adults. In: Annals of Nutrition and Metabolism. Conference: 20th International Congress of Nutrition. Granada Spain. Conference Publication: (var. pagings). 63 (pp 487-488); S. Karger AG,

Kattelmann, K.K. Bredbenner, C.B. White, A.A. Greene, G.W. Hoerr, S.L. Kidd, T. Colby, S. Horacek, T.M. Phillips, B.W. Koenings, M.M. Brown, O.N. Olfert, M.D. Shelnutt, K.P. and Morrell, J.S. (2014) The Effects of Young Adults Eating and Active for Health (YEAH): A Theory-Based Web-Delivered Intervention. Available from: http://www.sciencedirect.com/science/article/pii/S149940461400637X. [Accessed August 2018].

Kelders S.M. van GemertPijnen J.E.W.C. Werkman A. and Seydel, E.R. (2009) Evaluation of a web-based lifestyle coach designed to maintain a healthy bodyweight. Journal of telemedicine and telecare, 16 (1), 3-7.

Kessels L.T.E. Ruiter R.A.C. Brug J. and Jansma, B.M. (2011) The effects of tailored and threatening nutrition information on message attention. Evidence from an event-related potential study. Appetite, 56 (1), 32-38.

Kings Fund (2017) What is social prescribing? London: King's Fund. Available from: https://www.kingsfund.org.uk/publications/social-prescribing [Accessed April 2020].

Kivits, J., Ricci, L. and Minari, L. (2019) Interdisciplinary Research in Public Health, the 'why' and the 'how'. Journal of Epidemiology and Community Health, 73 (12), 1061-1062.

Knowledge@Wharton (2016) The Rise of the Chatbots: Is It Time to Embrace Them? Wharton University of Pennsylvania. Available from: http://knowledge.wharton.upenn.edu/article/rise-chatbots-time-embrace/. [Accessed July 2017].

Korotitsch, W.J. and Nelson-Gray, R. (1999) An overview of self-monitoring research in assessment and treatment. Psychological assessment, 11 (4), 415-425.

Kreuter, M. Farrell, D. Olevitch, L. and Brennan, L. (2000) Tailoring health messages: Customizing communication with computer technology. Lawrence Erlbaum Associates Publishers.

Kvale, S. (2007) Chapter 5: Conducting an Interview. In: Fick, U. ed. Doing Interviews. First Edition ed. London: Sage Publications, 57.

Kyselova, V. (2018) Why Brands Abandon Bot Technology and How to Make Chatbots Work. Florida: Jasoren. Available from: https://jasoren.com/why-brands-abandon-bot-technology-and-how-to-make-chatbots-work/ [Accessed March 2020].

Laugwitz, B. Held, T. and Schrepp, M. (2008) Construction and Evaluation of a User Experience Questionnaire. In: HCI and Usability for Education and Work, 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 11/20, Graz, Austria. 63-76.

Lee, J. Song, S. Ahn, J.S. Kim, Y. and Lee, J.E. (2017) Use of a Mobile Application for Self-Monitoring Dietary Intake: Feasibility Test and an Intervention Study. Nutrients, 9 (7),

Lee, W. Chae, Y.M. Kim, S. Ho, S.H. and Choi, I. (2010) Evaluation of a mobile phone-based diet game for weight control. J Telemed Telecare, 16 (5), 270-275.

Leventhal, H. (1971) Fear appeals and persuasion: the differentiation of a motivational construct. American Journal of Public Health, 61 1208-1224.

Liberati, A. Altman, D.G. Tetzlaff, J. Mulrow, C. Gøtzsche, P.C. Ioannidis, J.P.A. Clarke, M. Devereaux, P.J. Kleijnen, J. and Moher, D. (2009) The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. PLOS Medicine, 6 (7), e1000097.

Liberman, A. and Chaiken, S. (1992) Defensive processing of personally relevant health messages. Personality and Social Psychology Bulletin, 18 669-679.

Lin, W. and Yao, G. (2014) Concurrent Validity. In: Michalos, A.C. ed. Encyclopedia of Quality of Life and Well-Being Research. Dordrecht: Springer Netherlands, 1184-1185.

Lin, W. and Yao, G. (2014) Predictive Validity. In: Michalos, A.C. ed. Encyclopedia of Quality of Life and Well-Being Research. Dordrecht: Springer Netherlands, 5020-5021.

Linney, S. (2019) The Flesch Reading Ease and Flesch-Kincaid Grade Level. Brighton: Readable. Available from: https://readable.com/blog/the-flesch-reading-ease-and-flesch-kincaid-grade-level/ [Accessed September 2019].

MacLean, P.S. Wing, R.R. Davidson, T. Epstein, L. Goodpaster, B. Hall, K.D. Levin, B.E. Perri, M.G. Rolls, B.J. Rosenbaum, M. Rothman, A.J. and Ryan, D. (2015) NIH working group report: Innovative research to improve maintenance of weight loss. Obesity, 23 (1), 7-15.

Martín, J. Muñoz-Romero, C. and Ábalos, N. (2017) chatbottest - Improve your chatbot's design. chatbottest.com. Available from: https://chatbottest.com/ [Accessed March 2019].

McCaffrey, J. (2017) Revealing Secrets with R and Factor Analysis. Irvine, California: Converge 360. Available from: https://visualstudiomagazine.com/articles/2017/03/01/revealing-secrets-r-factor-statistics.aspx [Accessed September 2019].

McLeod, S. (2016) Bandura - Social Learning Theory. Simply Psychology. Available from: www.simplypsychology.org/bandura.html [Accessed August 2017].

McLeod, S. (2014) Bobo Doll Experiment. Simply Psychology. Available from: https://www.simplypsychology.org/bobo-doll.html [Accessed June 2019].

Medical Research Council (2019) Developing and evaluating complex interventions: Updated 2019. London: Medical Research Council. Available from: https://mrc.ukri.org/documents/pdf/complex-interventions-guidance/ [Accessed July 2019].

Moher, D. Liberati, A. Tetzlaff, J. Altman, D.G. and The PRISMA Group (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine, 6 (7), e1000097.

Moore, R.J. Arar, R. Ren, G. and Szymanski, M.H. (2017) Conversational UX Design. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. Denver, Colorado, USA. New York, NY, USA: ACM, 492-497.

Moorhead, S. Coates, V. Gallagher, A. Nolan, G. Murphy, K. and Hazlett, D. (2013) Obesity communication among patients by health professionals: Findings from the Weight Care Project. Health, 5 (8C), 100-109.

Moran, G. (2015) The Science Behind Why Inspirational Quotes Motivate Us. USA: Fast Company. Available from: https://www.fastcompany.com/3051432/why-inspirational-quotes-motivate-us [Accessed April 2020].

Mosa, A. Yoo, I. and Sheets, L. (2012) A Systematic Review of Healthcare Applications for Smartphones. BMC Medical Informatics & Decision Making, 12 67.

Mulvenna, M. Boger, J. and Bond, R. (2017) Ethical by Design - A Manifesto. In: European Conference on Cognitive Ergonomics 2017 (ECCE 2017). September, ACM Digital Library, 51-54.

National Health Service (2019) Obesity. National Health Service. Available from: http://www.nhs.uk/Conditions/Obesity/Pages/Introduction.aspx [Accessed March 2019].

National Health Service (2016) Should you lose weight fast? United Kingdom: National Health Service. Available from: https://www.nhs.uk/live-well/healthy-weight/should-you-lose-weight-fast/ [Accessed May 2019].

National Library of Medicine (2019) MEDLINE: Description of the Database. National Library of Medicine. Available from: https://www.nlm.nih.gov/bsd/medline.html [Accessed February 2020].

National Library of Medicine (2019) PubMed Overview. National Library of Medicine. Available from: https://www.nlm.nih.gov/bsd/pubmed.html [Accessed February 2020].

Needleman, R. (2017) Is it OK for a chatbot to impersonate a human? Chatbots Magazine. Available from: https://chatbotsmagazine.com/is-it-ok-for-a-chatbot-to-impersonate-a-human-75632dea6e33 [Accessed May 2019].

Newton, K.H. Wiltshire, E.J. and Elley, C.R. (2009) Pedometers and Text Messaging to Increase Physical Activity. Diabetes care, 32 (5), 813-815.

Nielsen, J. (1994) 10 Heuristics for User Interface design. Nielsen Norman Group. Available from: https://www.nngroup.com/articles/ten-usability-heuristics/ [Accessed March 2019].

Nielsen, J. (2012) Usability 101: Introduction to Usability. California: Nielsen Norman Group. Available from: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ [Accessed April 2019].

Nielsen, J. and Landauer, T.K. (1993) A Mathematical Model of the Finding of Usability Problems. In: Proceedings of the INTERCHI '93 Conference on Human Factors in Computing Systems. Amsterdam, The Netherlands. Amsterdam, The Netherlands, The Netherlands: IOS Press, 206-213.

Ofcom (2018) Communications Market Report. London: Ofcom. Available from: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/117256/CMR-2018-narrative-report.pdf

Parahoo, K. (2014) Questionnaires. In: Anon.Nursing Research: Principles, Process and Issues. Third Edition ed. London: Macmillan International Higher Education, 294-303.

Park, M. and Kim, H. (2012) Evaluation of mobile phone and Internet intervention on waist circumference and blood pressure in post-menopausal women with abdominal obesity. International Journal of Medical Informatics, 81 388-394.

Park, M. Kim, H. and Kim, K. (2009) Cellular phone and Internet based individual intervention on blood pressure and obesity in obese patients with hypertension. International Journal of Medical Informatics, 78 704-710.

Partridge, S.R. Allman-Farinelli, M. McGeechan, K. Balestracci, K. Wong, A.T.Y. Hebden, L. Harris, M.F. Bauman, A. and Phongsavan, P. (2016) Process evaluation of TXT2BFiT: a multi-component mHealth randomised controlled trial to prevent weight gain in young adults. International Journal of Behavioral Nutrition and Physical Activity, 13 (1), 7.

Patrick, K. Raab, F. Adams, A.M. Dillon, L. Zabinski, M. Rock, L.C. Griswold, G.W. and Norman, J.G. (2009) A Text Message Based Intervention for Weight Loss: Randomized Controlled Trial. J Med Internet Res, 11 (1), e1.

Personetics (2016) How Chatbots Fit into Your Omnichannel Strategy. Personetics. Available from: http://personetics.com/wp-content/uploads/2016/10/Personetics-Omnichannel-Bot-Brief.pdf [Accessed June 2017].

Petrie, P. (2013) Pregnancy and the Changing World of Technology. Available from: http://corporate.mumsviews.co.uk/about.php.

Pollard, C.M. Howat, P.A. Pratt, I.S. Boushey, C.J. Delp, E.J. and Kerr, D.A. (2016) Preferred Tone of Nutrition Text Messages for Young Adults: Focus Group Testing. JMIR MHealth and UHealth, 4 (1), e1.

Preedy, V. and Watson, R. (2009) Handbook of Disease Burdens and Quality of Life Measures. New York: Springer.

Prochaska, J. DiClemente, C. and Norcross, J. (1992) In search of how people change: applications to addictive behaviors. American Journal of Psychology, 47 1102-1114.

Prochaska, J. and Velicer, W. (1997) The Transtheoretical Model of Health Behavior Change. American Journal of Health Promotion, 12 (1), 38-48.

Reichheld, F.F. (2003) The One Number You Need to Grow. Massachusetts: Harvard Business Review. Available from: https://hbr.org/2003/12/the-one-number-you-need-to-grow [Accessed March 2019].

Saenz, J. Burgess, W. Gustitis, E. Mena, A. and Sasangohar, F. (2017) The Usability Analysis of Chatbot Technologies for Internal Personnel Communications. IIE Annual Conference. Proceedings, 1357-1362.

Saunders, B. Sim, J. Kingstone, T. Baker, S. Waterfield, J. Bartlam, B. Burroughs, H. and Jinks, C. (2018) Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & Quantity, 52 (4), 1893-1907.

Sauro, J. (2011) Measuring Usability with the System Usability Scale (SUS). Colorado: MeasuringU. Available from: https://measuringu.com/sus/ [Accessed March 2020].

Sauro, J. (2012) 10 Things to know about the Single Ease Question (SEQ). Colorado: MeasuringU. Available from: https://measuringu.com/seq10/ [Accessed August 2019].

Sauro, J. (2012) Predicting Net Promoter scores from System Usability Scale Scores. Colorado: MeasuringU. Available from: https://measuringu.com/nps-sus/ [Accessed March 2019].

Sauro, J. (2018) 5 ways to interpret a SUS score. Colorado: MeasuringU. Available from: https://measuringu.com/interpret-sus-score/ [Accessed March 2019].

Sauro, J. (2016) Are the terms formative and summative helpful? Colorado: MeasuringU. Available from: https://measuringu.com/formative-summative/ [Accessed March 2019].

Schmidt-Nielsen, A. (1995) Intelligibility and acceptability testing for speech technology. In: Syrdal, A. Bennett, R. and Greenspan, S. eds. Applied Speech Technology. Boca Raton: CRC Press

Schrepp, M. (2019) User Experience Questionnaire Data Analysis Tool. SAP. Available from: https://www.ueq-online.org/Material/Data_Analysis_Tools.zip [Accessed March 2019].

Schrepp, M. (2019) User Experience Questionnaire Handbook. 4th ed. [Accessed March 2020].

Schwartz, O. (2019) In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation. New York: IEEE Spectrum. Available from: https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation [Accessed April 2020].

Searle, J. (1980) Minds, Brains and Programs. The Behavioral and Brain Sciences, 3 (3), 417-424.

Sedgwick, P. and Greenwood, N. (2015) Understanding the Hawthorne effect. BMJ, 351

Shapiro, J.R. Koro, T.F. Doran, N.F. Thompson S FAU - Sallis, James, F. FAU, S.J. Calfas, K.F. and Patrick, K. (0423) Text4Diet: a randomized controlled study using text messaging for weight loss behaviors.

Sheth, B. (2016) The Bot Lifecycle: What to know before you make your bot. Chatbots Magazine. Available from: https://chatbotsmagazine.com/the-bot-lifecycle-1ff357430db7 [Accessed April 2019].

Shneiderman, B. Plaisant, C. Cohen, M. Jacobs, S. and Elmqvist, N. (2016) Designing the User Interface: Strategies for Effective Human-Computer Interaction. Sixth edition ed. Pearson.

Sky News (2019) Obesity crisis: The UK's weight problem in seven charts. United Kingdom: Sky News. Available from: https://news.sky.com/story/seven-charts-on-the-uks-obesity-problem-11583981 [Accessed 2019].

Smith, K.L. Kerr, D.A. Fenner, A.A. and Straker, L.M. (2014) Adolescents just do not know what they want: a qualitative study to describe obese adolescents' experiences of text messaging to support behavior change maintenance post intervention. Journal of Medical Internet Research, 16 (4), e103.

Sniehotta, F.F. Evans, E.H. Sainsbury, K. Adamson, A. Batterham, A. Becker, F. Brown, H. Dombrowski, S.U. Jackson, D. Howell, D. Ladha, K. McColl, E. Olivier, P. Rothman, A.J. Steel, A. Vale, L. Vieira, R. White, M. Wright, P. and Araújo-Soares, V. (2019) Behavioural intervention for weight loss maintenance versus standard weight advice in adults with obesity: A randomised controlled trial in the UK (NULevel Trial). PLOS Medicine, 16 (5), e1002793.

Spector, P.E. (1992) Summated Rating Scale Construction. Iowa: Sage Publications, Inc.

Stabb, S. (1995) Needs assessment methodologies. In: Stabb, S. Harris, S. and Talley, J. eds. Multicultural Needs Assessment for College and University Student Populations. C.C. Thomas, 51-115.

Stephens J. and Allen, J. (2015) Young adults, technology, and weight loss: A focus group study. In: Circulation. Conference: American Heart Association's Epidemiology and Prevention/Lifestyle and Cardiometabolic Health 2015 Scientific Sessions. Baltimore, MD United States. Conference Publication: (var. pagings). 131 (no pagination); Lippincott Williams and Wilkins,

Stevens, V.J. Funk, K.L. Brantley, P.J. Erlinger, T.P. Myers, V.H. Champagne, C.M. Bauck, A. Samuel-Hodge, C.D. and Hollis, J.F. (2008) Design and implementation of an interactive website to support long-term maintenance of weight loss. Journal of Medical Internet Research, 10 (1), e1.

Svetkey LP, Stevens VJ, Brantley PJ, et al (2008) Comparison of strategies for sustaining weight loss: The weight loss maintenance randomized controlled trial. JAMA, 299 (10), 1139-1148.

Teixeira, P.J., Silva, M.N., Mata, J., Palmeira, A.L. and Markland, D. (2012) Motivation, self-determination and long-term weight control. International Journal of Behavioural Nutrition and Physical Activity, 9 (22)

Thomas, J. and Wing, R. (2013) Health-E-Call, a Smartphone-Assisted Behavioral Obesity Treatment: Pilot Study. JMIR MHealth and UHealth, 1 (1),

Thomas, D. Vydelingum, V. and Lawrence, J. (2011) E-mail contact as an effective strategy in the maintenance of weight loss in adults. Journal of Human Nutrition & Dietetics, 24 (1), 32-38.

Thorndyke, E.L. (1898) Animal intelligence: an experimental study of the associative processes in animals. New York: Macmillan.Available from: https://archive.org/details/animalintelligen00thoruoft/page/n3 [Accessed May 2019].

Travis, D. (2006) Standards Update: Usability Test Reporting. London: UserFocus. Available from: https://www.userfocus.co.uk/articles/cif.html [Accessed March 2019].

Truby, H. Baic, S. deLooy, A. Fox, K.R. Livingstone, M.B. Logan, C.M. Macdonald, I.A. Morgan, L.M. Taylor, M.A. and Millward, D.J. (2006) Randomised controlled trial of four commercial weight loss programmes in the UK: initial findings from the BBC 'diet trials'. BMJ, 332 (7553), 1309.

Tullis, T. and Albert, B. (2018) Measuring the User Experience: SUS Calculation Tool. MeasuringUX.com. Available from: http://www.measuringux.com/sus_calculation.xls [Accessed March 2019].
Turing, A. (1950) Computing Machinery and Intelligence. Mind, LIX (236), 433-460.

Tutty, L.M. and Rothery, M.A. (2010) Needs Assessments. In: Thyer, B. ed. The Handbook of Social Work Research Methods. Second Edition ed. California: Sage Publications, 149-162.

UKOM (2018) UK Digital Market Overview. UKOM. Available from: https://www.comscore.com/layout/set/popup/content/download/44264/2545450/version/2/file/UKOM-Digital-Marketing-Overview-December-2018-Final2.pdf [Accessed March 2019].

Usability 24/7 (2017) Formative and Summative Usability Testing. Usability 24/7. Available from: https://www.usability247.com/blog/formative-summative-usability-testing/ [Accessed August 2018].

Usability.gov (2017) System Usability Scale (SUS). Usability.gov. Available from: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html [Accessed March 2019].

UX24/7 (2017) Formative and Summative Usability Testing. London: UX24/7. Available from: https://ux247.com/formative-summative-usability-testing/ [Accessed March 2019].

White, G. (2018) Child advice chatbots fail to spot sexual abuse. London: BBC. Available from: https://www.bbc.co.uk/news/technology-46507900 [Accessed March 2019].

Wing, R. and Hill, J. (2001) Successful Weight Loss Maintenance. Annual Review of Nutrition, 21 323-341.

Wing, R.R. Tate, D.F. Gorin, A.A. Raynor, H.A. and Fava, J.L. (2006) A self-regulation program for maintenance of weight loss. The New England journal of medicine, 355 (15), 1563-1571.

Woolford, S.J. Barr, K.L.C. Derry, H.A. Jepson, C.M. Clark, S.J. Strecher, V.J. and Resnicow, K. (2011) OMG do not say LOL: obese adolescents' perspectives on the content of text messages to enhance weight loss efforts. Obesity, 19 (12), 2382-2387.

World Health Organisation (February 2018) Obesity & Overweight Fact Sheet. Geneva: World Health Organisation. Available from: http://www.who.int/mediacentre/factsheets/fs311/en/ [Accessed March 2019].

Yap, T.L. and Davis, L.S. (2008) Physical activity: the science of health promotion through tailored messages. Rehabilitation Nursing Journal, 33 (2), 55-62.