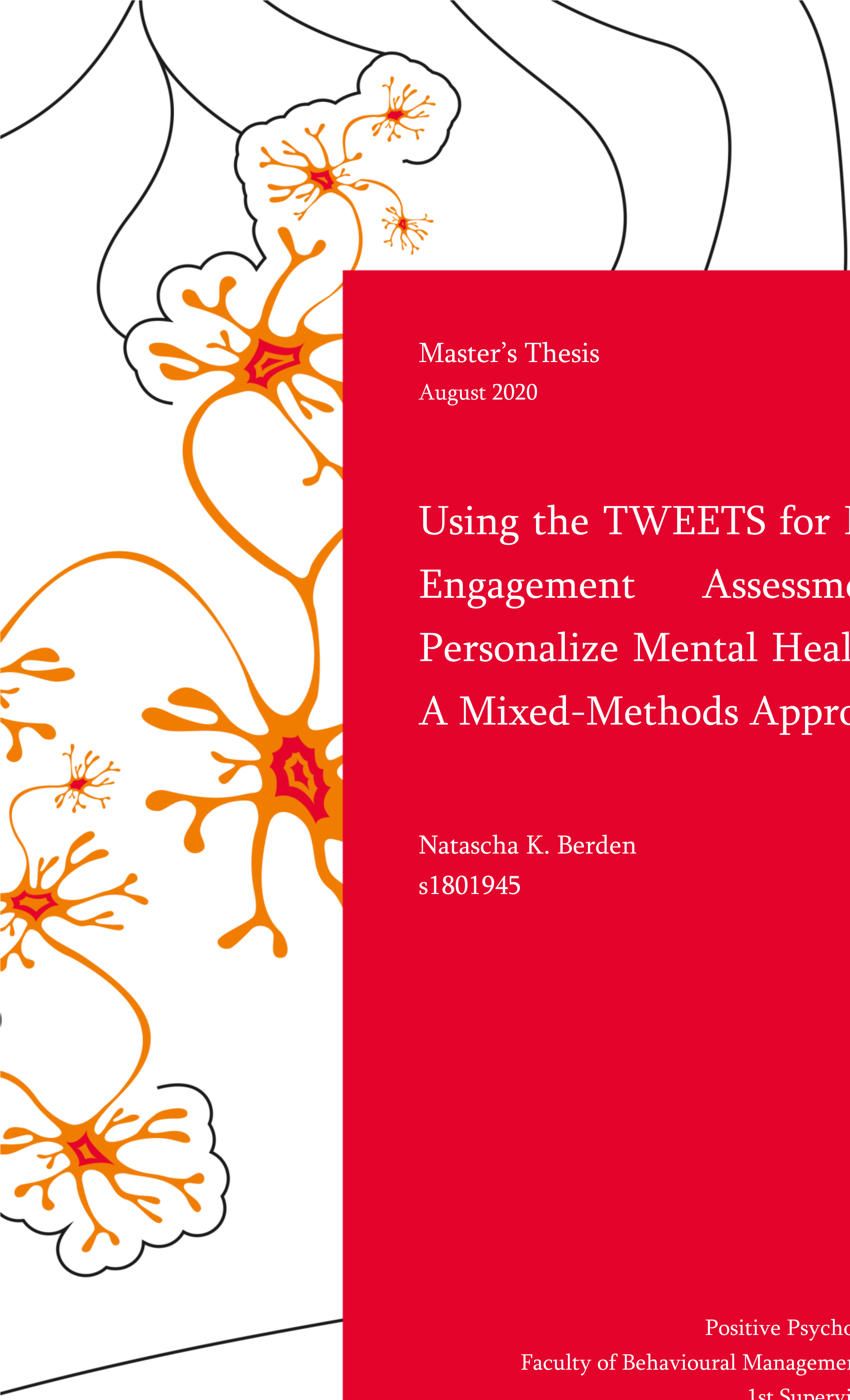


Master's Thesis
August 2020

Using the TWEETS for Expected Engagement Assessment to Personalize Mental Health Apps. A Mixed-Methods Approach.

Natascha K. Berden
s1801945

Positive Psychology and Technology
Faculty of Behavioural Management and Social Sciences
1st Supervisor: Dr. S. M. Kelders
2nd Supervisor: Dr. N. Köhle

UNIVERSITY OF TWENTE.

TABLE OF CONTENTS

Abstract

Background: Besides the public availability and continuously increasing uptake rates of mobile mental health (mMH) interventions, potential users rarely continue interacting with those apps after their download. One reason for this seems to be the failure of those apps to engage the users and motivate further interaction with the interventions. In this context, engagement is seen as a subjective experience with an app which is created by the expectation that a certain app could satisfy certain mental health needs as well as the individual's intention to direct one's thoughts, emotions and behaviour towards the interaction with that app. Though research in this area is still scarce at this point, findings of previous studies point at a connection between engagement and personalization of mMH apps. Here, the users' engagement with mMH apps was suggested to improve through personalization, which in turn related to an increased effectiveness of such apps. This leads to the current study's assumption that by assessing people's expected engagement with different features of mMH apps, individual feature preferences could be detected to compile a personalized app version. To do so, the present study tested the applicability of the TWente Engagement with Ehealth and Technologies Scale (TWEETS) as a personalization tool for mMH apps.

Methods: The mixed-methods design of this study included two consecutive online surveys and one voluntary follow-up interview (N= 11). During the first survey (N= 62), the TWEETS was used to assess participants' expected engagement regarding different mMH app features. In the second survey (N= 58), participants were confronted with four different app versions that entailed personalized app feature combinations based on the scores of the first survey. Regarding these different app versions, participants again had to indicate their expected engagement using the TWEETS. Additional voluntary interviews were conducted to obtain information about the participants' experiences with the TWEETS and feedback on the personalization procedure and possible improvement suggestions for both.

Results: As it was hypothesised; (1) the TWEETS scores of the single app features from the first survey predicted how the suggested app versions in the second survey would be ranked; and (2) the TWEETS helped to discriminate different degrees of expected engagement between different features and app versions. Thus, the individual combination of the single features that received the highest, medium or lowest expected engagement scores in the first round were also the individually highest, medium, or lowest scored feature combinations in the second round, respectively, with significant scoring differences between those ranks. During the interviews the participants emphasized that they appreciate the opportunity to design the mMH app according to their preferences. It was also pointed out that they supported their preferences in scoring of the features and app versions and, thus, that they adjusted their scoring patterns according to the preferences they had formed before they completed the TWEETS. Nevertheless, repetitively completing the TWEETS was perceived as too time consuming and participants would prefer a quicker technique to personalize their apps.

Conclusion: The current findings complement previous research about personalization and engagement by showing that the TWEETS was successful in detecting the participants' feature preferences based on their expected engagement. However, before using it as a personalization tool in real life scenarios, it is recommended to adjust its length and wording and to test its added value compared to a simpler personalization procedure. Due to the participants' use of the TWEETS to explain their preference choices, the TWEETS seems to be useful in evaluating app features regarding their engagement potential. Further results and implications for future studies are discussed.

1. Introduction

Over the last decade, mental health care has experienced momentous developments. Thanks to technological advances, people nowadays are not dependent on distinct people and locations anymore to receive mental health related information and support. Instead, they can easily access educational material and advice through the internet, or more specifically through their smartphones, which is enabling them to participate in versatile mobile interventions (Gliddon, Barnes, Murray & Michalak, 2017; Gowen, Deschaine, Gruttadara & Markey, 2012; Naslund, Marsch, McHugo & Bartels, 2015; Vockley, 2015). This form of mobile psychological health care service is also described as mMental Health (mMH). It has the potential to increase the scalability and quality of care, to enable anonymity of the receiver and reduce healthcare costs (Berger, Wagner & Baker, 2005; van Gemert-Pijnen, Kelders, Kip & Sanderman, 2018). Making use of mMH, it is not only possible to access the desired mental health information content on individually perceived demand, but also at any place or time (van Gemert-Pijnen, Kelders, Kip & Sanderman, 2018). Moreover, by transferring psychological interventions from therapeutic facilities to the user's everyday environment, the individual can be supported in applying behavioural and cognitive changes under real-life conditions. This turns smartphones into potentially efficient tools in mental health care (Bakker, Kazantzis, Rickwood & Rickard, 2016).

Making use of mMH to manage mental health becomes increasingly important, especially in younger populations. Not only do most psychological disorders start during adolescents and early adulthood (age 12-24), but also is the prevalence of mental health problems like depression, anxiety and mood disorders in younger populations continuously increasing (Bayram & Bilgel, 2008; Dennison, Morrison, Conway & Yardley, 2013; Patel, Flisher, Hetrick & McGorry, 2007; Punukollu & Marques, 2019). Here, mMH is a promising option to help young people manage their mental health (Gipson, Torous & Maneta, 2017; Tal & Torous, 2017). It can prevent the often feared confrontation with stigmatization by others and can solve the problem of low accessibility and lack of local and affordable mental health care services (Berger, Wagner & Baker, 2005; van Gemert-Pijnen, Kelders, Kip & Sanderman, 2018). Moreover, by using smartphones from an early age on and being familiar with operating them, adolescents and young adults are supposedly well receptive for mMH (Gipson, Torous & Maneta, 2017; Patel, Flisher, Hetrick & McGorry, 2007).

Reviews about web- and mobile-based mental health interventions in pupils, adolescents and young adults have indeed shown to be sufficient in decreasing psychiatric disorders like depression or anxiety and increasing mental health (Barak, Hen, Boniel-Nissim & Shapira, 2008; Bennett, Ruggero, Sever & Yanouri, 2020; Davies, Morriss & Glazebrook, 2014; Punukollu & Marques, 2019). Also, though both approaches had significantly positive effects on decreasing depressive symptoms, the meta-analysis of Firth and colleagues (2017) found evidence for larger effects achieved by mMH interventions which focused on mental health improvement than by those which addressed symptom reduction only. But not only do mMH interventions show effectiveness in increasing mental health when there is already demand for improvement, there is also evidence that mMH interventions can support illness and relapse preventions (Flett, Hayne, Riordan, Thompson & Conner, 2018; Naslund, Aschbrenner Araya, Marsch, Unützer, Patel & Bartels, 2017; Rathbone & Prescott, 2017). Thus, considering the advantages and therapeutic potential of mMH interventions, the increasing numbers of mental health problems in younger populations, but also their technological skills, mMH interventions seem like a promising tool to provide a remedy here. But besides the promising findings of recent research, those studies on mMH interventions repetitively emphasise the need for further research on guidelines for designing and conceptualizing efficient and effective mMH interventions.

On the one side, people's interest in mMH and, thus, individually managing their mental health can be concluded from the quick uptake of mMH apps, which is replicated in their high download rates (Liquid State, 2018). On the other side, however, the use as intended, also called adherence, and final completion of such mobile interventions only range from 1% to 29% (Torous, Wisniewski, Liu & Keshavan, 2018). This means that though people seem to have an initial motivation or interest to explore mMH apps, the probability of them to make use of the app and interact further with it after download is small. This phenomenon of high interest but low adherence and, thus, low expected effectiveness of mMH apps has recently received increasing attention in research (Punukollu & Marques, 2019). Regarding this, a study by McCurdie and colleagues (2012) points out that mMH apps most times lose their potential users shortly after the download of such apps, before users start to properly interact with the program. This is also supported by Torous, Nicholas, Larsen, Firth and Christensen's (2018) and Fanning, Mullen and McAuley's (2012) review and meta-analysis of mMH apps regarding their effectiveness to increase mental health and the interaction between users and the apps. These studies concluded that the origin behind low adherence and discontinuation might be connected to engagement problems.

1.1 Engagement

Engagement with health technology is described to be shown in the person's positive experiences with health technology and the presence of an intention and perceived need to change with the help of health technology (van Gemert-Pijnen, Kelders, Kip & Sanderman, 2018; Tippey & Weinger, 2017). Furthermore, it is also assumed to be expressed in the match between the operational elements of a health technology device and the user's experience, skills, and characteristics of technology use. Also important to mention here are the findings of Perski, Blandford, West, and Michie's (2016) systematic review on engagement and the conceptualized framework for people's engagement with digital behavioural change interventions. Here, it was pointed out that engagement is often defined to be a subjective experience with an intervention. This is supposed to include the individual's motivation to interact with the app and direct one's thoughts and emotions towards actions that need to be taken in such interventions. More precisely, this includes the interaction with the device via which the intervention is accessed as demanded and performing the tasks that are suggested.

Overall, engagement was frequently pointed out to play a central role in translating mental health care treatments to technological devices and that it is a crucial element in motivating potential users to interact with and adhere to an mobile apps (Torous, Staples, & Onnela, 2015; Torous, Wiśniewski, Liu & Keshavan, 2018). However, besides these supposedly highly influential aspects of engagement, there is still a lack of sufficient guidelines and concrete design choices to modify engagement and overcome initial interaction deficits with mMH apps (Torous, Nicholas, Larsen, Firth & Christensen, 2018). Moreover, looking at how engagement is defined, there seems to be a mismatch with the way it is assessed and taken care of in health technology. Looking at the depicted characteristics of engagement, a descriptive and predictive image of interacting with health technology is shown. More precisely, it is examined how the intervention users' individual stance towards and evaluation of such programs preferably need to be so users' initiate and continue the use of those apps.

But the way engagement is dominantly assessed is retrospectively in the course of product evaluation, more focusing on the frequency and patterns of users' interaction with and attitudes of a health technology like mMH (Fanning, Mullen & McAuley, 2012; Serrano, Coa, Yu, Wolff-Hughes & Atienza, 2017; Taki, Lymer, Russell, Campbell, Laws, Ong & Denney-Wilson, 2017; Torous, Nicholas, Larsen, Firth & Christensen, 2018). Hence, the findings of such assessments have only been used to adjust certain mMH's specific issues in hindsight, without the possibility to modify mMHs in advance (Ng, Firth, Minen & Torous, 2019). So, on the one side there is the awareness of the importance of the preconceptions in individually experienced necessity and overall perception of technological interventions. This includes the elementary motivation and willingness to adjust one's behaviour, emotions, and thoughts to engage with these technologies. But on the other side, the results of people's mMH attitudes and patterns of use are only incorporated in the modification of the mMH technology in retrospection. This is expressed in adjusting the overall intervention program to improve attitudes and regularity of interaction with the specific mMH. Hence, the aspects of mMH that are adjusted to increase the interaction frequency and improve people's opinions based on the retrospective assessments commonly done until now are no direct predictor for future engagement but mainly a trial and error procedure of modifying those technologies.

Due to this mismatch between the predictive definitions of engagement, including users' attitude, intention and perceived value of health technology, and the retrospective, evaluative measurement methods of analysing user patterns and frequency, the Twente Engagement with Ehealth and Technologies Scale (TWEETS) was developed (Kelders & Kip, 2019; Kelders, Kip & Greeff, 2019). Combining qualitative assessment of mMH users and findings from systematic review of research about user engagement like those from Perski, Blandford, West, and Michie (2016), this scale was conceptualised to assess engagement in users of technological healthcare services. This includes general electronic healthcare, but therein also mobile healthcare. Due to the key roles of the experience of and disposition to immerse in an activity emotionally and cognitively for predicting ongoing interaction, this scale assesses people's perceived level of affective, behavioural, and cognitive engagement. The resulting TWEETS is a self-report scale with an overall reasonable test-retest reliability, convergent and predictive validity, good divergent validity, and high internal consistency. The scale entails three parts that are created to determine the eHealth users' expected, current, and past engagement with a technological healthcare service. This makes it both a predictive as well as evaluative tool for engagement with electronic healthcare (Kelders, Kip & Greeff, 2019; Appendix A).

1.2 Personalization

The current study, however, goes one step further and tests, if the TWEETS can, besides assessing engagement at different points, also be deployed as a personalization tool for eHealth. Emphasized by multiple studies, personalization in eHealth is seen as a potent tool to increase the effectiveness of and adherence in healthcare interventions, and is suggested to enhance engagement (Ghane, Huynh, Andrews, Legg, Tabuenca & Sweeny, 2014; Handelzalts & Keinan, 2010; McCurdie, et al., 2012; Oinas-Kukkonen & Harjumaa, 2009; Rajanen & Rajanen, 2017; Tippey & Weinger, 2017). This is assumed to be done by giving the potential consumers the choice to modify the eHealth technology they are interacting with before or during their interaction process. Depending on their personal demands and preferences, they have the option to choose between several elements and features they would like to have integrated in their intervention or program. Supported by qualitative analysis of customer satisfaction and their feedback on mMH (Donkin & Glozier, 2012; Kim, Kim & Wachter, 2013), personalization is also described to enable users to feel directly involved and active in adjusting the intervention according to their own preferences and needs.

Participants of personalized eHealth interventions were also shown to be more satisfied and achieve better results compared to participants who had no choice of making changes in their eHealth interventions (Ghane, Huynh, Andrews, Legg, Tabuenca & Sweeny, 2014; Handelzalts & Keinan, 2010). It, therefore, also seems as if people can have an individually better understanding of from which kinds of interventions their needs would benefit most and how often they would have to interact with the eHealth or mMH intervention to achieve their desired health state than the providers of such interventions (Bartley, Faasse, Horne & Petrie, 2016; Geers, Rose, Fowler, Rasinski, Brown & Helfer, 2013). In turn, this also leads to the assumption that the suggested treatment type and pattern of adherence of eHealth or mMH interventions may not always predict the maximum profit individuals can achieve with such interventions (Achilles, Anderson, Li, Subotic-Kerry, Parker & O'Dea, 2020; Donkin, Hickie, Christensen, Naismith, Neal, Cockayne & Glozier, 2013)

Thus, considering the characteristics and predictors of engagement, personalization of mMH seems like an efficient and effective way for letting people create interventions that satisfy their engagement needs. More specifically, by letting the potential users take control over the content and course of the interventions, personalization could accommodate the users' need to associate an intervention with positive emotions and feel motivated and capable to follow the exercises (Perski, Blandford, West & Michie, 2016; van Gemert-Pijnen, Kelders, Kip & Sanderman, 2018; Tippey & Weinger, 2017). In turn, this could increase the probability to adhere to and complete the upcoming treatment and increase its effectiveness, since users may perceive a stronger sense of identifying with the app, assuming it will help them to tackle their individual points of concern. Contributing to this are the findings of studies by Wallace, Bogard and Zbikowski (2018) and Achilles and colleagues (2020) on different goal setting and adherence behaviours of people participating in general health coaching as well as mHealth programs respectively. Both studies concluded that there are intrapersonal differences in the goal setting, the way participants prefer to interact with the program and the kind of programs they tend to pick. Here, personalization of future health interventions according to individual needs and demands was recommended to increase the participants' motivation to progress and maximize the probability of successful behavioural change.

For now, the common procedure for personalizing mMH apps is usually disease-centred. Depending on the users' complaints, symptom matching interventions are delivered. This way it can also be tested how effective distinct elements of mMH apps are in treating different mental health issues. These elements are focusing on suitable content, design, and feedback and have several effective options therein. However, though it is known that there are different treatments for certain health concerns and that people vary in their preferences in how to handle these (Whalley & Hyland, 2009), like with engagement, there are currently no common methods for personalizing electronic healthcare. Starting with the content element, evidence-based theoretical frameworks and exercises from cognitive behavioural therapy (CBT), positive psychological interventions (PPI) and mindfulness interventions, like acceptance and commitment therapy (ACT), have been found to be effective content choices for mMH apps (Bakker, Kazantzis, Rickwood & Rickard, 2016; Chida & Steptoe, 2008; Flett, Hayne, Riordan, Thompson & Conner, 2018; Hetrick, et al., 2017). Not only have mMH apps with these frameworks been associated with the users' increased mental well-being and decreased psychological problems like depression and anxiety disorders, but also shown preventative functions by increasing participants' resilience and self-regulating abilities. However, until now it is not known which mMH interventions are most effective for which individual needs and goals since more research is needed on this.

Suggested options for design choices incorporate the findings of qualitative and quantitative research on user preferences in mobile health applications. Here, customers report that visual demonstration of accomplishments, mastered and upcoming tasks are perceived as attractive characteristics of health apps (Dennison, Morrison, Conway & Yardley, 2013; Gowin, Cheney, Gwin, & Franklin Wann, 2015; van Gemert-Pijnen, Kelders, Kip & Sanderman, 2018). But it was also emphasized that a playful manner of teaching new skills motivated adherence and encouraged to perform the desired behaviour. Combining these findings, the illustrated characteristics can be found in gamification designs. Gamification is often used in mobile interventions that target behavioural change (Cotton & Patel, 2018) and integrates gaming principles into non-gaming environments. This includes the rewarding of desired behaviour in mMH apps, which encourages adherence and motivates adoption of such behaviour (Deterding, Dixon, Khaled & Nacke, 2011; Zichermann & Cunningham, 2011). The playful concept and reward elements aim to create enjoyment that results into a positive emotional experience of the intervention. This in turn increases the probability of a positive learning experience and supports the performance of the desired behaviour (Mullins & Sabherwal, 2020). Studies about the applicability of gamified educational material has shown that it is an effective tool to increase peoples' intrinsic motivation to engage with the program, which supports the acquisition of new practical skills and performance of behavioural change techniques (Bui, Veit & Webster, 2015; Dale, 2014; de-Marcos, Domínguez, Saenz-de-Navarrete & Pagés, 2014; Jones, Madden & Wengreen, 2014).

However, it is important to consider that the success of gamified interventions may be influenced by the users' individual preconditions. It was found that users vary in their levels of gaming attitudes, skills and experiences, which in turn are expressed in different ways of approaching those interventions (Hamari, Juho, & Tuunanen, 2014; Huotari & Hamari, 2016; Yee, 2006). This might have consequences on the way people process and interact with the mobile application and what they put their focus on, which can then lead to different outcomes. On the one hand, some focus more on mastering the game itself and others aim for mastering the skills that are taught. On the other hand, there are experienced and open people with a positive gaming attitude, but also people who prefer a more pragmatic manner in learning new skills. All these aspects can result in different outcomes. Therefore, different options of gamification and design need to be presented to the users to satisfy different tendencies. Up to personal preferences, this can take, for instance, a competitive form where users need to go through several levels and master certain tasks, a more narrative form where users are accompanied by avatars that represent individual achievements and receive rewards for accomplishing personal challenges, or a traditional and non-gamified form of pragmatically presenting and explaining different exercises (Hamari, Koivisto & Sarsa, 2014; Rajanen & Rajanen, 2017; Zichermann & Cunningham, 2011).

Looking at feedback options, motivational feedback has been shown to be most beneficial and motivational in terms of intervention effectiveness and adherence (Musiat, Hoffmann & Schmidt, 2012). Feedback via electronic devices can be delivered via text, video, image and/or sound formats, can replace the absence of real life human support and can function as a motivator and aid to improve one's skills (Dixon, 2015; Oinas-Kukkonen & Harjumaa, 2009). Studies on electronic health interventions have also demonstrated that support in the form of feedback and/or reminders increase the effectiveness of such programs and contribute to prolonged changes and performance of desired behaviour (Hurling, Fairley & Dias, 2006). However, though feedback has been shown to increase motivation and effectiveness of interventions (Bennett & Glasziou, 2003; Cunningham, Hodgins, Toneatto, Rai & Cordingley, 2009), there are mixed findings about the most effective style of feedback. Studies on different styles of feedback like audio, video, and/or text feedback have each been shown to be effective (Ice, Swan, Diaz, Kupczynski & Swan-Dagen, 2010). Though it is suggested to be most effective if it is customized towards personal levels of ability and needs (Berner, 2019; Copeland, Rooke, Rodriquez, Norberg & Gibson, 2017; Rassaei, 2019), and that most people prefer a combination of audio/video and text feedback compared to text only (Borup, West & Graham, 2012; Ice, Swan, Diaz, Kupczynski & Swan-Dagen, 2010; Lalley, 1998), there are also studies that support the notion that the effectiveness of feedback may vary due to individual preferences (Ice, Swan, Diaz, Kupczynski & Swan-Dagen, 2010; Olesova, Richardson, Weasenforth & Meloni, 2011). Thus, while video and text feedback may be the best option for one person, another person may achieve equal results with text feedback only.

Overall, combining the need for mental health care in younger populations, specifically undergraduate students (Beiter, Nash, McCrady, Rhoades, Linscomb, Clarahan & Sammut, 2015; Cvetkovski, Reavley & Jorm, 2012; Eisenberg, Gollust, Golberstein & Hefner, 2007), the promising potential of mMH interventions (Bakker, Kazantzis, Rickwood & Rickard, 2016; Barak, Hen, Boniel-Nissim & Shapira, 2008; Punukollu & Marques, 2019; Tal & Torous, 2017), and missing guidelines for personalization in electronic healthcare, the current study examined if the TWEETS can be used as a personalization tool for mMH interventions. This was done by assessing participants' expected engagement (TWEETS version) regarding the three fundamental building blocks, also called domains, of an mMH app, namely, the content, the format of received feedback, and the design of an mMH app. Thus, for the TWEETS to be an adequate personalization tool, it was hypothesized that the preferences in design, feedback and content of mMH apps, assessed in advance by the TWEETS, result into the detection of the best combination of mMH app elements for each individual participant. Firstly, this is hypothesized to be demonstrated by the participants' choice of their final intervention version, which is assumed to be similar to their feature preferences. Secondly, this is also hypothesized to be shown in the detection of significantly different TWEETS scores between the different domain features and between the different presented app versions. Additionally, to get a better impression of the users' experiences with their final choice of interventions and the personalization process, and to gather more information about the feasibility of the TWEETS as a personalization tool, interviews were conducted at the end of the study.

2. Methods

The current study is a pilot study about the feasibility of the TWEETS as a potential personalization tool for mMH apps and was conducted between April and June 2020. A mixed methods design with quantitative and qualitative data analysis was deployed, including convenience sampling. After completing online surveys at two consecutive times, interviews were conducted with participants who indicated to be interested in volunteering as interviewees for an evaluative assessment of the study. This study was approved by the ethical committee of the University of Twente (registration no.: 200213).

2.1 Participants

Due to convenience sampling, the sample consisted of students from the University of Twente who individually decided to take part in this study via the SONA systems platform of the University of Twente as well as students from the researcher's private network. From the initially 62 enrolled participants, 5 were excluded from further analysis of the data, since they did not take part in the second round of the quantitative online assessment. During the quantitative phase of this pilot study, 58 participants took part in the complete online assessment, out of whom 56.5% were female, the age range was 18 to 33 with a mean of 21.92, 90% were Bachelor's students and the majority of 66.2% was German. The subsequent qualitative phase included 11 voluntary students from the former participant sample with 45.45% females. People were eligible to participate in this study if they were at least 18 years old, currently a student, in possession of a smartphone and a laptop, in case they want to take part in the evaluative interview via a communication tool on their laptops, and able to read, speak and write in English fluently. Participants were recruited via the researcher's personal networks and the SONA system of the University of Twente through which students could take part in the study in exchange for research credits.

2.2 Material

2.2.1 TWEETS

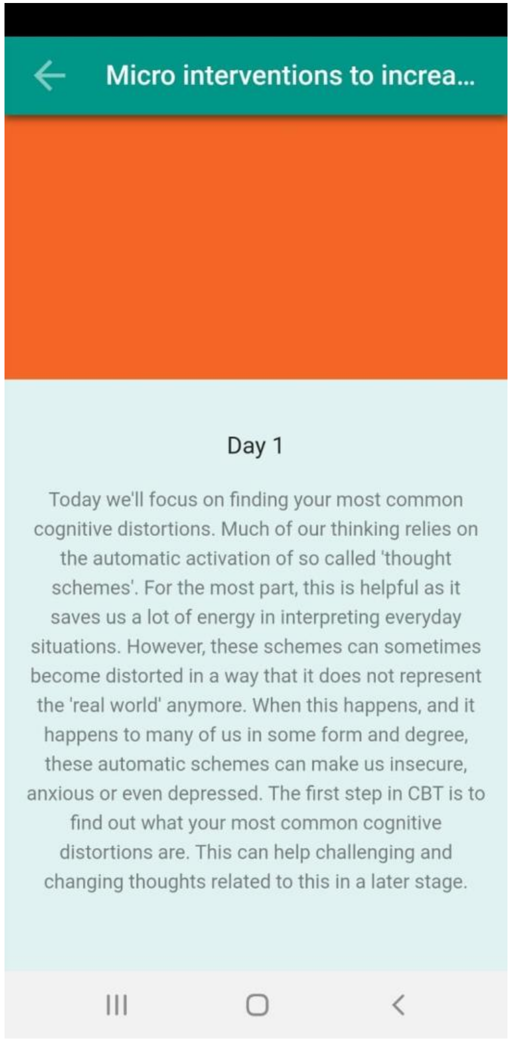
The TWente Engagement with Ehealth Technologies Scale (TWEETS) contains nine items that assess user-engagement with eHealth technologies across three different areas (Kelders, Kip & Greeff, 2019). It includes each three items to assess the areas of behavioural engagement (items 1-3), cognitive engagement (items 4-6) and affective engagement (items 7-9). Though there are three adaptations of the TWEETS for measuring user-engagement at different points (expectational, current and past engagement), for the purpose of the current study only the expectational engagement adaptation has been used and adjusted to the respective phases of the study. Thus, adjustments in the items' wording have been made in session one and two separately. For session one, the wording has been adapted with regards to the three different app domains of content, feedback and design, and for the second session, the item phrasings have been changed to specify the focus on the presented app versions (Table 1). Internal consistency of the adjusted scales was .93, .95, and .93 for content, feedback, and design respectively during the first session, and .96 for the app specification during the second session. The TWEETS has shown to have high internal consistency, good divergent validity, and reasonable convergent and predictive validity and test-retest reliability (Kelders, Kip & Greeff, 2019).

Table 1. *Adjusted TWEETS items for expectational engagement regarding feature specific and app specific assessment.*

Item	Content specific TWEETS	App specific TWEETS
1	Using an app with this *content* can become part of my daily routine.	Using this *app* can become part of my daily routine.
2	The *content* of this app is easy to use.	This *app* is easy to use.
3	I will be able to use an app with this *content* as often as needed to improve my well-being.	I will be able to use this *app* as often as needed to increase my well-being.
4	An app with this *content* will make it easier for me to work on increasing my well-being.	This *app* will make it easier for me to work on increasing my well-being.
5	This *content* motivates me to increase my well-being.	This *app* motivates me to increase my well-being.
6	This *content* will help me to get more insight into my well-being.	This *app* will help me to get more insight into my well-being.
7	I will enjoy using an app with this *content*.	I will enjoy using this *app*.
8	I will enjoy seeing the progress I make by using an app with this *content*.	I will enjoy seeing the progress I make in this *app*.
9	An app with this *content* will fit me as a person.	This *app* will fit me as a person.

2.2.2 App features

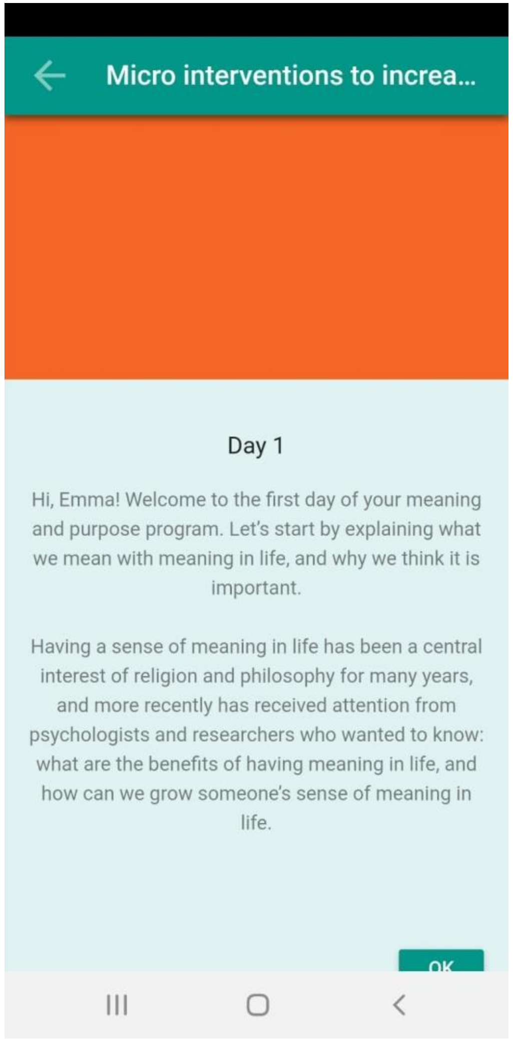
To resemble the best app version, participants had to show their feature preferences by indicating their expected engagement levels for each of the three different feature templates per domain. The templates showed how the features would look like when they are incorporated in the well-being app. For the content domain, Cognitive Behavioural Therapy (CBT), Acceptance and Commitment Therapy (ACT) and Positive Psychology (PP) templates were presented to them (Figure 1-3). Each content variation included an introduction to an example exercise with a short explanation of the psychological theory behind the respective exercise. The templates were deprived from other app elements like feedback and specific design variations, to only give an idea how the content would look like. The feedback templates showed how video feedback by a counsellor (VFC), the text feedback presented by an avatar (TFA), and the text only feedback (TOF) would be incorporated (Figure 4-6). The design templates illustrated the competitive gamification (CompG), non-competitive gamification (NoCompG) and no gamification features (noG) (Figure 7-9).



Day 1

Today we'll focus on finding your most common cognitive distortions. Much of our thinking relies on the automatic activation of so called 'thought schemes'. For the most part, this is helpful as it saves us a lot of energy in interpreting everyday situations. However, these schemes can sometimes become distorted in a way that it does not represent the 'real world' anymore. When this happens, and it happens to many of us in some form and degree, these automatic schemes can make us insecure, anxious or even depressed. The first step in CBT is to find out what your most common cognitive distortions are. This can help challenging and changing thoughts related to this in a later stage.

Figure 1. CBT content template

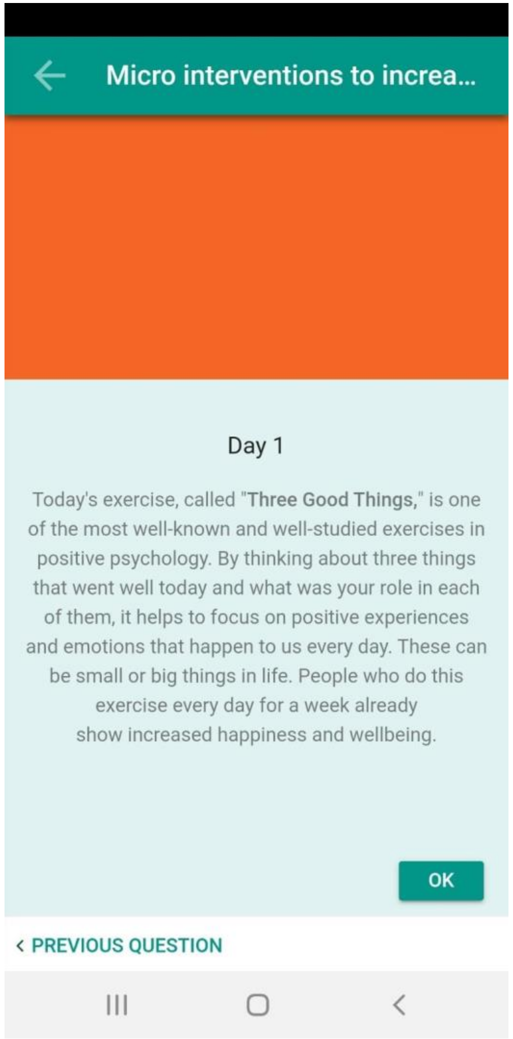


Day 1

Hi, Emma! Welcome to the first day of your meaning and purpose program. Let's start by explaining what we mean with meaning in life, and why we think it is important.

Having a sense of meaning in life has been a central interest of religion and philosophy for many years, and more recently has received attention from psychologists and researchers who wanted to know: what are the benefits of having meaning in life, and how can we grow someone's sense of meaning in life.

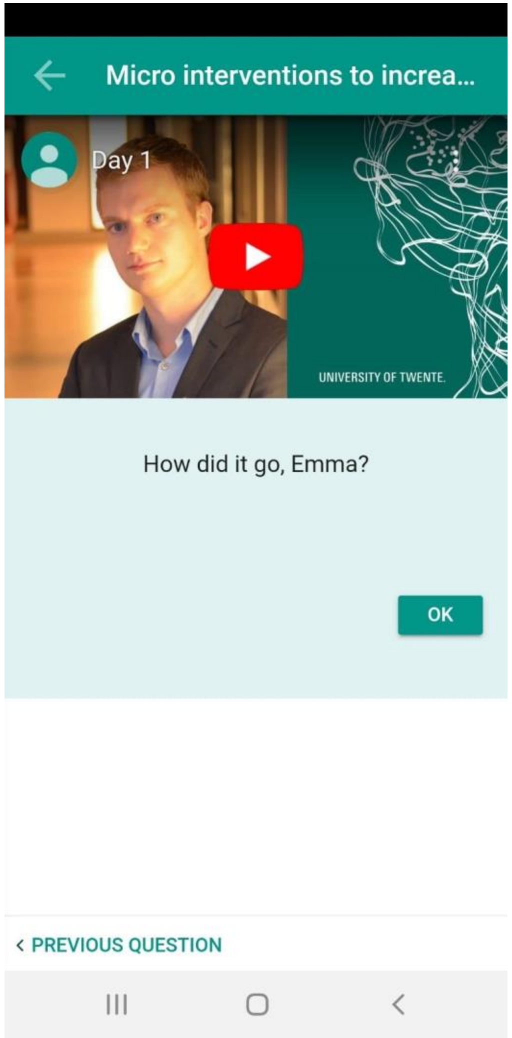
Figure 2. ACT content template



Day 1

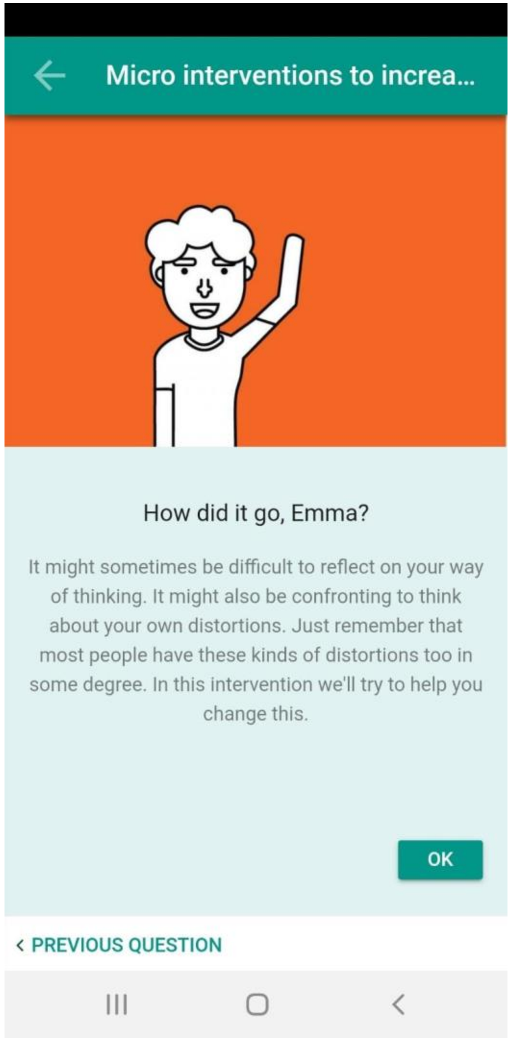
Today's exercise, called "**Three Good Things,**" is one of the most well-known and well-studied exercises in positive psychology. By thinking about three things that went well today and what was your role in each of them, it helps to focus on positive experiences and emotions that happen to us every day. These can be small or big things in life. People who do this exercise every day for a week already show increased happiness and wellbeing.

Figure 3. PPI content template



How did it go, Emma?

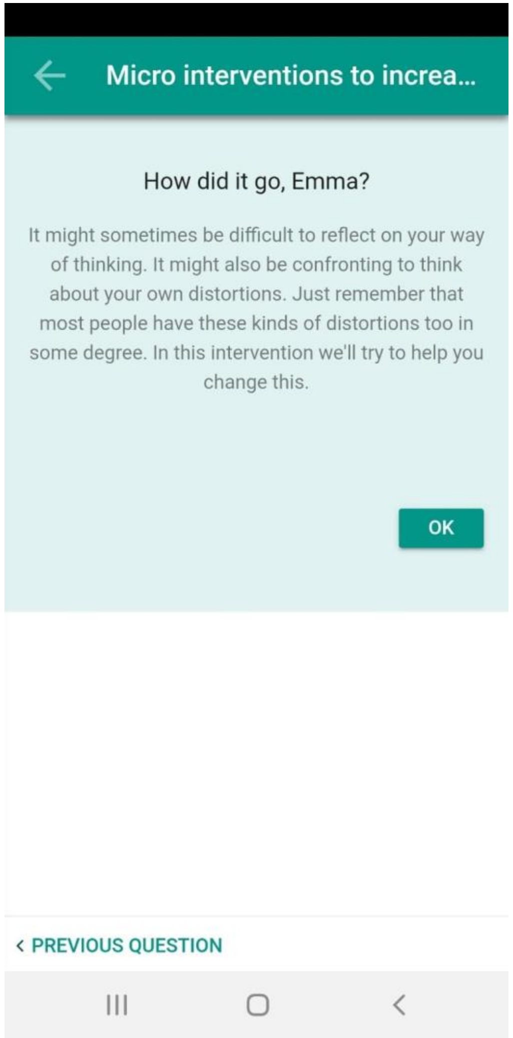
Figure 4. VFC feedback template



How did it go, Emma?

It might sometimes be difficult to reflect on your way of thinking. It might also be confronting to think about your own distortions. Just remember that most people have these kinds of distortions too in some degree. In this intervention we'll try to help you change this.

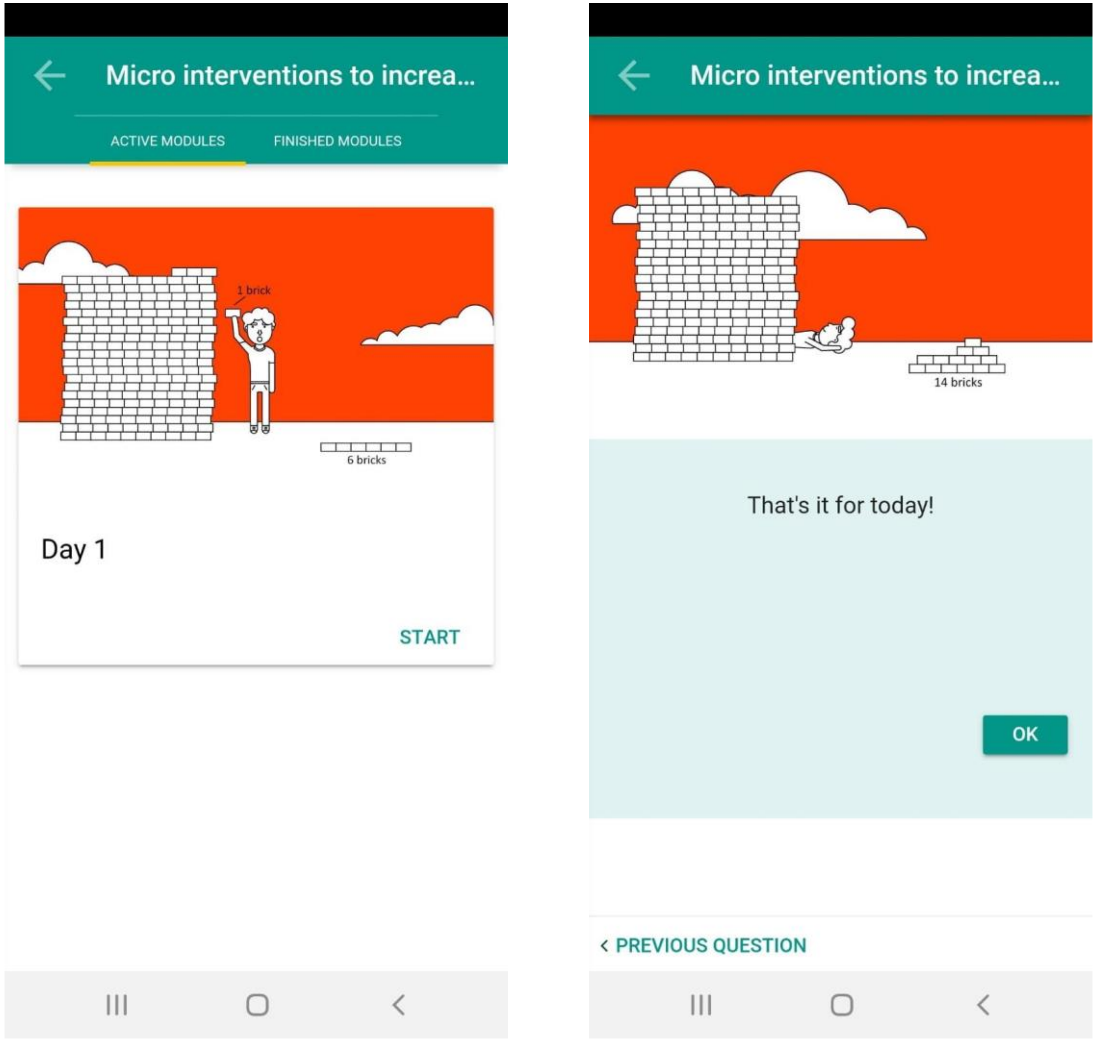
Figure 5. TFA feedback template



How did it go, Emma?

It might sometimes be difficult to reflect on your way of thinking. It might also be confronting to think about your own distortions. Just remember that most people have these kinds of distortions too in some degree. In this intervention we'll try to help you change this.

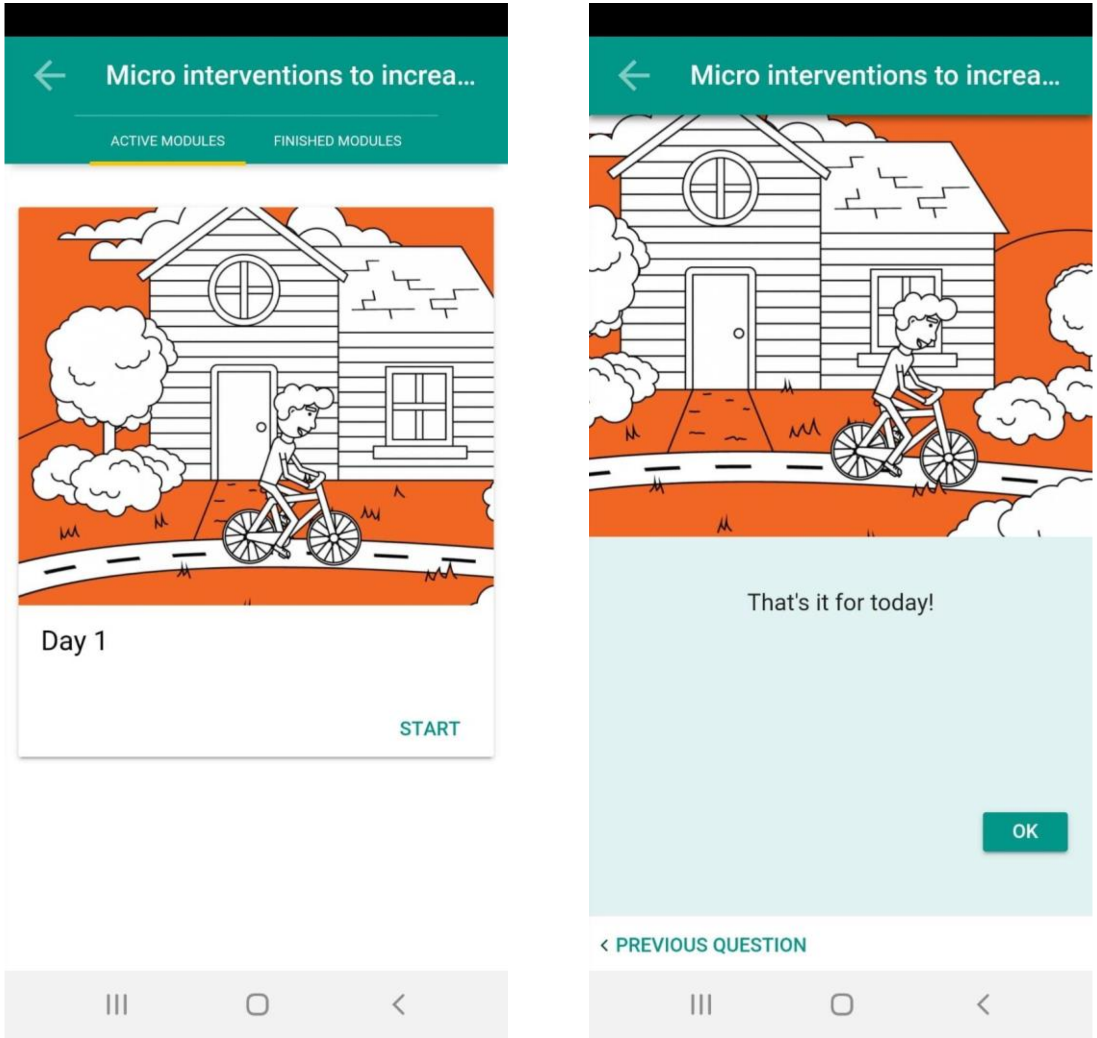
Figure 6. TFO feedback template



Day 1

That's it for today!

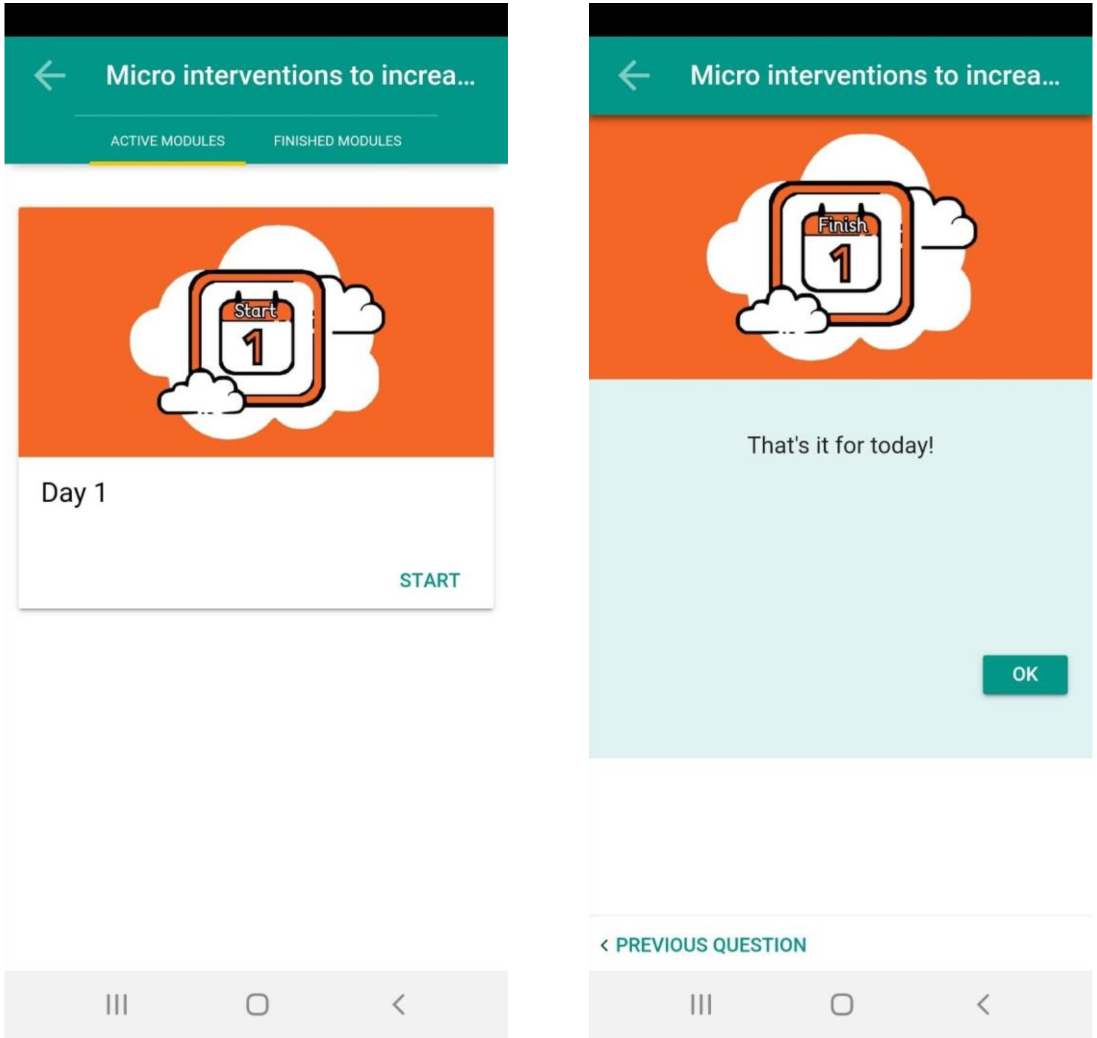
Figure 7. CompG design templates


Micro interventions to increa...
ACTIVE MODULES
FINISHED MODULES
START
Micro interventions to increa...
OK
‹ PREVIOUS QUESTION

Day 1

That's it for today!

Figure 8. NoCompG design templates



Micro interventions to increa...
ACTIVE MODULES
FINISHED MODULES
Start
1
Day 1
START
Micro interventions to increa...
Finish
1

That's it for today!

Figure 9. NoG design templates

2.2.3 Interviews

Interviews were taken 1:1 and recorded via the online video communication tool Skype. While covering the 3 main topics of interest, the interview included open questions and followed a flexible structure. This enabled to go deeper into individual arguments and comments to obtain a representative picture of the participants as potential users of mMH apps. The essential topics of the interview occupied (1) the users' perception of the presented apps, (2) the experiences with the TWEETS questions regarding their helpfulness in constructing an engaging app version, and (3) if there were any suggestions in further improving the personalization process. The order and precise wording of the questions listed below in table 2 may have varied between the interviews, depending on the natural flow of the conversation.

Table 2. *List of essential questions of the semi-structured interview.*

Topics	Questions
Perception of apps.	Which app version did you like the most? What was most important for you in this choice? What is your personally most important feature of an app?
Perceived helpfulness of the TWEETS to design an attractive app.	Did the questions help you to make a choice?/Did you find the questions helpful in making a choice? Which questions were most/least helpful? How would you prefer the best choice of an app to be presented to you?
Further suggestions to improve the personalization process.	Is there anything you missed along the way of designing your own app? Did you feel guided along the way? Do you have any further suggestions on how this could be made easier or more efficient?

2.3 Procedure

At the beginning of the study, participants were informed about the purpose of the study and upcoming session(-s), duration and number of sessions and the possibility to withdraw at any time. The study could be accessed via the SONA system platform of the University of Twente or via a link which was delivered by the researcher in case potential participants were not enrolled at the University of Twente. Students from the University of Twente received SONA credits as incentive after completing the second online survey and additional credits when taking part in the follow-up interview. The main phase of the study included two quantitative assessments via the online survey platform Qualtrics.

Starting with the first assessment, after giving informed consent, various template versions of three essential app domains (content, design, feedback) of a well-being app were presented to the participants. Participants completed the TWEETS for every domain separately by indicating their level of agreeableness with each item of the scale regarding the different templates. Thus, taking the domain of content as an example, three different content versions were presented to them, namely an example version of a CBT, PPI, and Acceptance and Commitment (ACT) content (Figure 1-3). Followed by the 9 items of the TWEETS (e.g. “This content will help me to get more insight on how to increase my well-being.”) with the answering options being the three content versions of CBT, PPI and ACT and adjacent slides ranging from 0 to 10 (0= strongly disagree; 10= strongly agree) to indicate the perceived applicability of the item’s statement on the content versions. Afterwards, the same procedure was conducted with the templates of the feedback domain (text only feedback, video feedback by counsellor, text presented by virtual agent) (Figure 4-6) and design domain (competitive gamification, non-competitive gamification, no gamification) (Figure 7-9).

After the first session, participants received an email, including an introduction to the following session and its length. The time that passed between activating the link for the first survey and the link for the second survey round was on average 148.45 hours (6.19 days) and participants received 1 to 3 reminder emails to complete the second survey. Attached to the emails regarding the second survey was a unique Qualtrics link which included a personalized survey, based on their individual results from the first survey. Thus, after activating the link, participants were confronted with four different blueprint versions of a well-being app. These blueprints resembled each participant’s preference indications from the first session. To compensate for any effects of the order of presentation on the participants’ app preference rankings, the presentation order of the blueprints was randomized, so the positions at which what kind of app versions were presented varied between each participant (Appendix B). One blueprint included the features that were ranked highest in each app domain (content, feedback, design), called suggested best-fit app. Another one included the highest ranked features from content and feedback and medium ranked in design, called suggested second-best-fit app. A third option included the medium ranked features of all three domains, called suggested medium-fit app, while a fourth included the lowest ranked features, called suggested least-fit app. Important to mention here is that since the randomization was done manually, it accidentally turned out that the versions that were presented at the third position were dominantly the suggested least-fit app versions and the ones presented at the fourth position were most often the suggested best-fit or second-best-fit app versions (Appendix B).

Participants were asked to answer the TWEETS for each version of the app separately, by indicating their agreeableness with each item regarding the app versions with adjacent slides from 0 to 10 (0= strongly disagree; 10= strongly agree). For transparency reasons, participants received the information on which of the presented four versions of the app was the one including features which were indicated to be most preferred during the first session, which version included the second-most preferred features, which the medium preferred and which the least preferred, after completion of the second session.

At the beginning of the main phase of the study, participants were asked if they were interested in taking part in a follow-up interview. This follow-up interview was conducted after the second session via the telecommunication application Skype. Participants were asked for their approval to record the interview before and at the beginning of the recording to obtain a recorded form of consent.

2.4 Statistical Analysis

For quantitative analysis, the data collected via the online survey platform Qualtrics were exported to IBM SPSS Statistics 26. After screening for completion, complete datasets of the first and second quantitative online session were available for 62 and 58 (91.94%) of the participants, respectively. All of them were included in the final analyses since there were neither missing data nor noticeable answering patterns. To check if parametric or non-parametric tests needed to be conducted, the Shapiro-Wilk test for small sample sizes was deployed. Using a significance level of α= 0.05, the Shapiro-Wilk test showed a normal distribution in the scoring of the TWEETS in the first as well as the second survey round, with $p>.18$. Hence parametric tests were used for data analysis.

2.4.1 Personalization procedure

To filter the individual preferences in the app feature domains of content, feedback and design, sub score means for each of the nine features were created. These three sub scores per feature domain were manually compared for each individual to establish every participant's expected engagement rankings. By using these individual rankings, four different personalized mMH app versions were compiled for each participant. Thus, the app that was suggested to be the participant’s best-fit resembled the features with the highest TWEETS scores per domain in the first round. This procedure was repeated for the suggested second-best-fit version with the highest ranked content and feedback templates and medium design template, the suggested medium-fit app version with the medium ranked features and the suggested least-fit version with the lowest ranked features per domain.

2.4.2 Personalization and discrimination checks

To test the hypothesis if the TWEETS is effective in detecting the best combination of mMH app elements for each participant, several steps were taken. To test if there were any features that scored significantly different from other features, analyses of variance were performed. For every domain, the categorical app domain variable was taken as the factor and the three feature mean variables as dependent variables. Hence, to look for significant differences in scorings of content, the categorical variable for content was taken as the factor in the equation and the continuous mean score variables for content (meanCBT, meanACT, meanPPI) as the dependent variables. This was repeated respectively for every feature domain, regardless of the individuals' feature ranks. Frequency distributions were examined on how many participants scored which feature highest, medium, or lowest to see if there are any features that are preferred by more people than others.

Next, since each participant received a personalized version of the second survey, including their individual versions of the four different app versions, presented in randomly varying orders, each dataset was made uniform and rearranged before merging. Therefore, mean engagement score variables were created for each app version and manually compared with each other for each participant to detect the best-fit, second-best-fit, medium-fit and least-fit app versions of the second round. Additional to each app version's mean engagement score variable, categorical variables expressing their new ranking were generated as well as categorical variables that indicated which of the 27 possible feature combinations were expressed with each app ranking. After merging, the frequencies and descriptives of the new app version engagement means and rankings were calculated to create a general overview of the participants' choices, thus, which app versions scored highest, second highest, medium, or lowest. This was also done to see if the suggested app versions based on the features' mean expected engagement scores from the first survey matched the rankings of the app versions of the second round. Furthermore, for each app rank of the second round, analyses of variance were conducted to see if there were feature combinations that scored significantly different from the other combinations within the rank category.

Moreover, to test the discriminative value of the TWEETS and if there is a significant difference between the feature preferences or between the four different app versions of the first and second round, or if there were features that scored significantly higher or lower than others, paired-sample t-tests were conducted. Therefore, regarding the scores of the first round, the individually highest, medium and lowest ranked feature scores per domain were compared with each other, as well as the means of the different features per domain, independent from the individuals' feature rankings. Regarding the second round, the means of the individuals' four new app ranks were compared with each other.

2.4.3 Interview coding

The interview recordings were transcribed using the transcription software AmberScript and were coded and analysed with the qualitative data analysis and research software Atlas.ti version 8. The interview analysis and coding were conducted by one coder and all information that could indicate the identity of the interviewees was anonymized. Both deductive and inductive coding was performed during an iterative process. All interviews were read three times to first get a general impression of the content, then secondly to identify the main topics and further individual expressions and comments regarding these, and a third time to look for overlapping statements in those expressions. Thus, after filtering the information on the three main topics (1) perception of the apps, (2) perceived helpfulness of the TWEETS to design an attractive app and (3) further suggestions to improve the personalization process, sub codes were identified in case there was consensus in content between multiple interviews (Tables 9-11). Depending on the precision and compactness of the statement, the codes encompassed single or multiple sentences.

3. Results

3.1 Preference distribution of features and app versions

Comparing the first survey round's TWEETS scores of the different features within the three domains of content, feedback and design, the results of the paired-sample t-tests show that some features scored significantly higher or lower than others in their domain (Table 3). Looking at the content domain, PPI (*M*= 6.27; *SD*= 1.77) scored significantly higher than CBT (*M*= 4.96; *SD*= 2.02) or ACT (*M*= 4.79; *SD*= 1.88), with 62.9% of participants favouring PPI, 27.4% favouring CBT and 8.1% scoring ACT the highest (Table 3). In the feedback domain, a rank order of first Avatar (*M*= 6.49; *SD*= 1.85), followed by Video (*M*= 5.53; *SD*= 2.20) and finishing with Text (*M*= 4.73;*SD*= 1.91) with each significant differences between them was exposed. Here, the preference frequency distribution per feedback feature was 64.5%, 29% and 4%, respectively. And regarding the design domain, 58.1% scored Bricks (*M*= 6.57;*SD*= 2.17) and 32.3% scored Bike (*M*= 6.37;*SD*= 1.83) both significantly higher than 9.7% scored Calendar (*M*= 4.70;*SD*= 2.08) . These feature preference indications of the first round resulted in the compilation of 17 suggested best-fit app versions (Figure 10). Moreover, representing the feature preference distributions, the app versions that were most frequently suggested as best-fit versions were, firstly, the PPI, Avatar and Bricks version by nearly 25 % of the participants, followed by, secondly, 12.9% for the PPI, Avatar and Bike version and, thirdly, the CBT, Video, Bricks version favoured by 11.3%.

Table 3. *Paired-sample t-test results of TWEETS scores of features based on the first survey round.*

Domain	Features	*mean* (*SD*)	*meanDif (SD)*	*t*	*p*
Content	CBT - ACT	4.96 (*2.02*) 4.79 (*1.88*)	.17 (*1.82*)	.70	.49
	CBT - PPI	4.96 (*2.02*) 6.27 (*1.77*)	-1.31 (*2.55*)	-3.95	.00**
	ACT - PPI	4.79 (*1.88*) 6.27 (*1.77*)	-1.48 (*2.14*)	-5.31	.00**
Feedback	Video - Avatar	5.53 (*2.20*) 6.49 (*1.85*)	-.96 (*2.34*)	-3.16	.00*
	Video - Text	5.53 (*2.20*) 4.73 (*1.91*)	.80 (*2.37*)	2.59	.01*
	Avatar - Text	6.49 (*1.85*) 4.73 (*1.91*)	1.76 (*1.65*)	8.21	.00**
Design	Bricks - Bike	6.57 (*2.17*) 6.37 (*1.83*)	.21 (*2.34*)	.67	.50
	Bricks - Calendar	6.57 (*2.17*) 4.70 (*2.08*)	1.87 (*2.84*)	5.07	.00**
	Bike - Calendar	6.37 (*1.83*) 4.70 (*2.08*)	1.67 (*2.34*)	5.48	.00**

Note: *SD* = Standard Deviation; *meanDif* = mean Difference; * *p*< .05; ***p*< .001

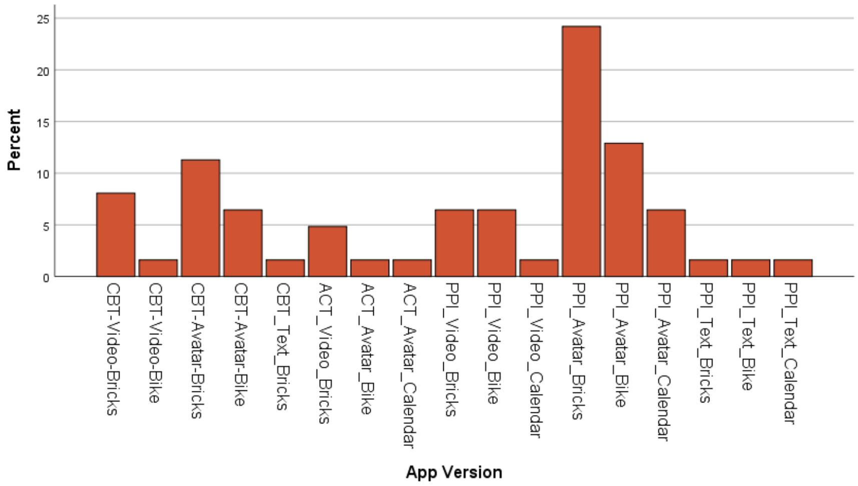


Figure 10. Frequencies of suggested best-fit app versions based on expected engagement TWEETS scores in the 1st survey.

Regarding the second round of the survey, best-fit app version preference distributions look similar to the best-fit feature preferences in the first survey round, with shifts from app versions including the Bricks design being favoured by most people towards versions with the same content and feedback but Bike design (Figure 11). For instance, looking at the three most frequently favoured versions from the first round, now the PPI, Avatar, Bike version was scored highest by 12.9% and PPI, Avatar, Bricks version by 11.3% of the participants. While the CBT, Avatar, Bricks app was on the third position, it is now on the fourth with 8.1% compared to 9.7% of participants who now scored the CBT, Avatar, Bike highest. However, while the highest scored app versions with CBT and PPI content also included the Avatar feedback in both rounds, the most appealing app version with ACT as content feature was not only combined with the Video instead of Avatar feedback but also the Bricks design in both rounds (both 4.8%).

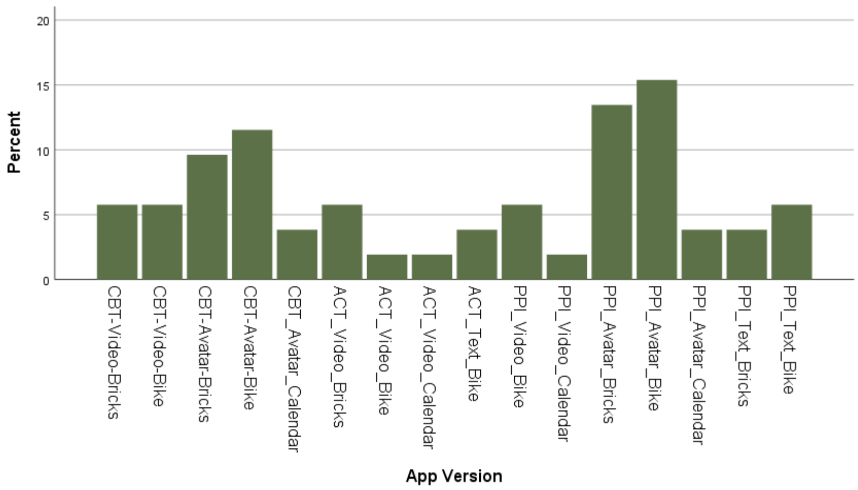


Figure 11. Frequencies of best-fit app versions based on expected engagement TWEETS scores in the 2nd survey.

Moreover, the results of the second round show that most people ranked their suggested best-fit and second-best-fit app versions based on feature preferences of the first survey round highest and thus as best-fit app version again, as it can be seen in figure 12. Here, the frequency of which type of app version was ranked highest in the second round is illustrated, with the types of app versions being defined by the ranks the app versions received in the first round. Hence, the “Best” bar shows the percentage of people who scored their suggested best-fit app version, containing the highest scored features from the first round, also highest in the second round. However, there were also suggested app versions that scored equally high in the second round, which is, for instance. demonstrated with the “Best/Second-Best”, saying that the suggested best-fit as well as the second-best-fit, based on the feature scores of the first round, scored both highest in the second round (Figure 12).

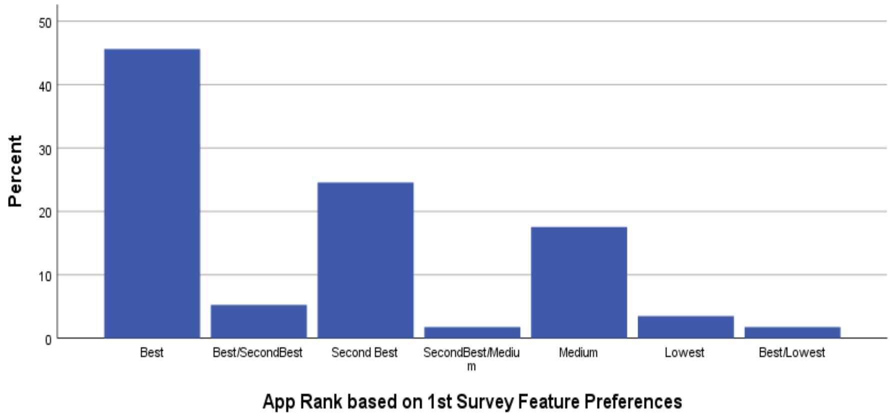


Figure 12. Frequencies of which app versions were scored highest in the second survey.

For exploratory purpose, to see if there are any app versions or features that elicit a significantly higher score than others within the highest scored features and app versions, additional analyses of variance have been conducted. However, there were neither any significant differences found in the scores of the TWEETS between the suggested best-fit app versions, *F*(16,45)= .50, *p*= .94, or within the feature domains regarding the outcomes of the first survey round, nor regarding the outcomes of the second survey round, *F*(15,36)= .78, *p*= .69 (Table 4). Hence, though there were in both rounds feature combinations that were scored highest more frequently, there were no single combinations of features that were scored significantly higher than others within the best-fit app versions and features. More precisely, no matter which feature combinations the suggested best-fit version of the first round (*M*= 7.17; *SD*= 1.41) or the highest scored app version of the second round (*M*= 7.40; *SD*= 1.35) contained, their scores did not significantly differ from each other.

Table 4. *Frequency distribution of and Differences between Best-Fit Features per App Domain based on TWEETS results of 1st and 2nd Survey.*

Domain	Feature	*n* (*%*)	mean (*SD*)	95% CI	*F*	*p*
1st Survey	CBT	17 (*27.4*)	7.36 (*1.73*)	6.47 - 8.25	.67	*.57*
Content	ACT	5 (*8.1*)	7.83 (*1.45*)	6.03 - 9.63		
	PPI	39 (*62.9*)	7.02 (*1.27*)	6.61 - 7.43		
	ALL	1 (*1.6*)	6.55	6.55		
Feedback	Video	18 (*29*)	7.29 (*1.52*)	6.53 - 8.05	.69	*.51*
	Avatar	40 (*64.5*)	7.20 (*1.41*)	6.75 - 7.65		
	Text	4 (*6.5*)	6.38 (*.79*)	5.13 - 7.62		
Design	Bricks	36 (*58.1*)	7.19 (*1.51*)	6.68 - 7.70	.21	*.81*
	Bike	20 (*32.3*)	7.24 (*1.38*)	6.59 - 7.89		
	Calendar	6 (*9.7*)	6.82 (*1.01*)	5.76 - 7.88		
Total		62	7.17 (*1.41*)	6.81 - 7.53	.50	*.94*
2nd Survey	CBT	19 (*30.6*)	7.35 (*1.62*)	6.57 - 8.14	.03	*.97*
Content	ACT	7 (*11.3*)	7.51 (*.74*)	6.82 - 8.20		
	PPI	26 (*41.9*)	7.40 (*1.30*)	6.88 - 7.93		
Feedback	Video	15 (*24.2*)	7.28 (*1.10*)	6.67 - 7.89	1.73	*.19*
	Avatar	30 (*48.4*)	7.64 (*1.47*)	7.09 - 8.19		
	Text	7 (*11.3*)	6.62 (*1.11*)	5.59 - 7.65		
Design	Bricks	20 (*32.3*)	7.29 (*1.04*)	6.81 - 7.78	.35	*.71*
	Bike	26 (*41.9*)	7.38 (*1.65*)	6.71 - 8.05		
	Calendar	6 (*9.7*)	7.83 (*.81*)	6.98 - 8.68		
Total	ALL	52 (*83.9*)	7.40 (*1.35*)	7.02 - 7.78	.78	*.69*

Note: *n*= Sample Size; *SD* = Standard Deviation

Results of further analyses of variance and paired-sample t-tests regarding the second-best, medium- and least-fit app versions of the first and second survey rounds can be found in Appendices C and D. These were conducted to also test if there were any feature combinations within the other app rankings with scores that are significantly different from other combinations of these specific ranks. Indeed, feature combinations that were ranked second highest in the second round scored significantly lower in case they contained the CBT content (*M*= 5.43; *SD*= 1.60) than apps with the PPI (*M*= 6.50; *SD*= 1.16) or ACT (*M*= 7.32; *SD*= 1.54) content, *F*= 5.08, *p*= .01. And the app versions that ultimately scored lowest, namely the lowest scored ones of the least-fit app versions, in the first, *F*= 3.52, *p*= .02, as well as in the second round, *F*= 9.04, *p*= .00, contained the Calendar design (1st round *M*= 3.77; *SD*= 1.40; 2nd round *M*= 3.76; *SD*= 1.52).

3.2 Discriminative value of TWEETS

To test the discriminative value of the TWEETS, paired-sample t-tests were performed for the results of the first and second survey. Regarding the TWEETS scores of the different features within the three domains, significant differences were found between every feature ranking per domain. Thus, within every domain, the scores of the highest, medium, and lowest scored features were all significantly different from each other (Table 5). Looking at the second survey, there were significant differences found between the mean expected engagement TWEETS scores of the different app ranks with the best-fit (*M*= 7.31; *SD*= 1.37) rank containing the highest scores, the second-best-fit (*M*= 6.41; *SD*= 1.42) the second-highest sores, the medium-fit (*M*= 5.71; *SD*= 1.54) the medium-high scores and the least-fit (*M*= 4.69; *SD*= 1.69) the lowest scores (Table 6). Regarding the order of presentation of the different app versions in the second round, there were significant differences found between the first and third presented app version (*t*(56)= 2.62, *p*= .01) and the second and third app version (*t*(56)= 2.36, *p*= .02) (Table 7).

Table 5. Paired-Sample t-test results on comparing TWEETS mean scores of different feature ranks within each feature domain from the first survey round.

	Mean (*SD*)	95% CI	*t*	*p*
Highest ranked content - Medium ranked content	1.69 (*1.44*)	1.32 - 2.05	9.19	.00*
Highest ranked content - Lowest ranked content	2.73 (*1.92*)	2.25 - 3.22	11.24	.00*
Medium ranked content - Lowest ranked content	1.05 (*1.17*)	.75 – 1.35	7.04	.00*
Highest ranked feedback - Medium ranked feedback	1.33 (*1.05*)	1.06 – 1.60	9.86	.00*
Highest ranked feedback - Lowest ranked feedback	3.05 (*1.66*)	2.62 – 3.47	14.35	.00*
Medium ranked feedback - Lowest ranked feedback	1.69 (*1.46*)	1.32 – 2.06	9.10	.00*
Highest ranked design - Medium ranked design	1.44 (*1.40*)	1.08 – 1.80	8.08	.00*
Highest ranked design - Lowest ranked design	3.14 (*1.99*)	2.64 - 3.65	12.46	.00*
Medium ranked design - Lowest ranked design	1.70 (*1.61*)	1.29 – 2.11	8.31	.00*

Notes: TWEETS= TWente Engagement with Ehealth Technologies Scale; *SD*= standard deviation.

* $p < .001$

Table 6. Paired-Sample t-test results on comparing TWEETS mean scores of different app versions from the second survey round.

	Mean (*SD*)	95% CI	*t*	*p*
Best App Version - Second Best App Version	.89 (*.82*)	.68 - 1.11	8.12	.00*
Best App Version - Medium App Version	1.59 (*1.08*)	1.30 - 1.88	10.97	.00*
Best App Version - Lowest App Version	2.61 (*1.54*)	2.20 - 3.02	12.70	.00*
Second Best App Version - Medium App Version	.70 (.71)	.51 - .89	7.40	.00*
Second Best App Version - Lowest App Version	1.72 (*1.13*)	1.42 - 2.02	11.38	.00*
Medium App Version - Lowest App Version	1.02 (*1.04*)	.74 - 1.30	7.33	.00*

Notes: TWEETS= TWente Engagement with Ehealth Technologies Scale; *SD*= standard deviation.

* *p* < .001

Table 7. Paired-Sample t-test Results on comparing TWEETS mean scores of different app versions from the second survey round based on the order of presentation.

	Mean (*SD*)	95% CI	*t*	*p*
First App Version - Second App Version	.00 (*1.83*)	-.48 - .49	.01	.99
First App Version - Third App Version	.56 (*1.63*)	.13 - 1.01	2.62	.01*
First App Version - Fourth App Version	.34 (*1.89*)	-.17 - .84	1.34	.19
Second App Version - Third App Version	.56 (*1.80*)	.09 - 1.04	2.36	.02*
Second App Version - Fourth App Version	.33 (*2.07*)	-.22 - .88	1.22	.23
Third App Version - Fourth App Version	-.23 (*2.08*)	-.78 - .32	-.83	.41

Notes: TWEETS= TWente Engagement with Ehealth Technologies Scale;
SD= standard deviation.
* *p* < .05

3.3 Interview Results

The interviews covered the three main topics of (1) perception of the presented apps, (2) perceived helpfulness of the TWEETS to design personalized apps and (3) further suggestions to improve personalization.

3.3.1 Perception of presented apps.

Belonging to the first topic, three main codes have been identified, namely *Personalization, Validation* and *Convincing Design* (Table 8). The first one called *Personalization* includes 5 subcodes that categorize all the quotes in which the interviewees describe their impressions of the presented app versions regarding their personalization background. Here, the most frequently mentioned attribute is combined under the subcode *choice satisfaction,* which was expressed as followed by one participant;

> When apps are very limited, I usually find something that annoys me. And that's sometimes really the point where I look for other solutions when I'm feeling too limited by the software or the experience. (Participant 7)

Personalization also includes the participants' positive perception of control and *recognizing their own influence* through interacting with the app and being able to pick from a repertoire of multiple features and examining a minimalistic preview of the app. This was described like, for instance;

> And especially also when it comes to the second round that we had to complete where we had like a personalized survey. This I actually really liked a lot because I could see that the answers I gave are based on the first round. I could see that you take this into account, what I preferred and what I didn't prefer so much. This is something that I really liked. (Participant 1)

But *Personalization* also refers to emotional components of a feeling of being cared for and focusing on the *joy and entertainment* factor when picking features.

> I would say maybe because in the previous rounds, when I had to do you know, when I had to choose which ones I most preferred, actually, these are the ones that I mostly preferred. So I just, just in my opinion, these are the ones that are more appealing and more kind of fun, if I may say so for myself. Two, yeah, that's why I have stated it's the first one (best-fit). (Participant 6)

Moreover, a mental-health app was thought of as a kind of *omniscient supporter* that can help you reach your personal goals by activating customized services, as depicted by one participant in the following statement;

> And I look at my phone and say, wow, I have to do this. Then I remember and I'm doing that task. And, yeah, I did like to see that an app is connected with me even if I don't use it in that moment. (Participant 2)

Therefore, tasks and content that created a *minimalistic and simple* impression were perceived most positive due to the expected easy use in daily life and low amount of effort needed to fulfil.

> (...) then I'm just like, okay (...), nice information, but it goes in one ear and goes out the other. And that's about it for the day. And I might not even read everything just because I'm like yeah, information overflow a little bit like I don't really care. (Participant 5)

Another aspect of priority and appreciation was described regarding the app's ability for *progress demonstration* and tracking of the user's performances and action. This is captured in the next main code named *Validation*, which was representatively summarized by participant 6 by saying; *"*I think this will motivate me even further to continue using the app and to continue, you know. Working with it and having fun and actually increasing my mental health.". This was especially perceived to be done by the feedback features, which were interpreted as *feedback on performance* and reactions on task accomplishments and, thus, seen as reassuring of the optimal app use, since it is perceived as "(...) very important to receive (...) good feedback just to know what to improve, what to keep." (Participant 2). Moreover, there was an underlying need to use the technology as intended and master the given tasks to receive reward by progress indications in design and positive feedback, wherein video feedback played a controversial role. While some perceived a video from a human protagonist as soothing and motivating, others indicated discomfort when imagining a stranger would react to one's mental health work.

Besides this, participants also repetitively expressed the focus on *Convincing Design*, which is the third main code. Important attributes participants were looking out for here were, for example, supporting a *professional impression*, which was created by an expected reliable theoretical background of the app, which participant 2 described in one statement like;

> And I also like the fact that I can see the logo of the university of twente. (...) That makes me feel that it's something approved. Educational, approve. Yeah. And I really, really like it.

By associating the suggested features and apps with academic origins, supported by reliable research results, the participants expressed a sense of trust in the structures and tasks of the application. This appreciation further included the assumed app designers' effort in creating the features as well as the study design, which initiated an increased willingness to progress and establishing reciprocity, in the sense of enjoying a well-structured and design product and fulfilling the demanded tasks in return. This is not only connected to the expected use of the app, but also with the completion of the questionnaires in the study.

In addition to this, the *visualization of progress* was also described to be one of the participants' major concerns when examining the different app versions; “I think I liked the third one more (...) where you see the bricks. The brick at the beginning, where you start and then afterwards where he sleeps. I enjoy this the most (Participant 7).” This acknowledgement of *visual aesthetics* in app design expanded from preferring designs that were perceived as attractive to attributing more importance and interest in this visual enjoyment than in the content that is delivered. Regarding this, comments that are summarized underneath the subcode *design influencing content perception* were made, especially connected to the gamified design features;

> So I really focused also on (...) the pictures and the images because mostly I'm also like a person, like when the image suits me, when I like it, then I also get more attached to it, also automatically to the content it provides. (Participant 1)

Table 8. Frequency distribution of main and subcodes regarding the topic Perception of Presented Apps.

Main Code (q/n)	Subcode (q/n)
Personalization (78/11)	Choice satisfaction (49/9)
	Recognizing their own influence (17/7)
	Identification (16/5)
	Simplification of tasks/minimalistic design (13/5)
	Omniscient supporter (8/5)
Validation (7/5)	Progress demonstration (7/4)
	Feedback on performance (5/3)
Convincing Design (36/10)	Visual aesthetics (20/10)
	Visualized progress (7/4)
	Professional impression (7/4)
	Design influencing content perception (6/4)

Notes: q= number of quotations within this code; n= number of interviews containing this code

3.3.2 Perceived Helpfulness of the TWEETS.

The second topic covered the two categories of *Helpful* and *Not Helpful* (Table 9). Starting with *Not Helpful,* the theme that stood out the most here by being stated most often was centred around the theme of *preference formation* before answering the TWEETS. This was depicted by participant 10’s statement which condensed the general impression of other interviewees as well;

> And I think I could have answered it shorter. (...) Yeah, after I saw all these versions, I already had a favourite. Then answering all these questions, I was like, OK, yeah, why? Like, OK, now I have to put in the questions the way I feel

> about all of these three versions. And I just like one version the most and that's it.

Another issue of concern was the *number of items* presented to the participants. Though in general, participants were not complaining about the amount of items included in the survey within this research situation, the majority of interviewees mentioned that they would be disinclined to complete a questionnaire of this length when downloading an app in real-life;

> No, that would be too much. I always judge apps on what I can get from (them). Like, if I don't get something from it and it takes too much time, then I drop it. It has to serve me. And in a way, when I have to always answer eight or nine questions after something I got from it, then I wouldn't like, you know, I wouldn't take the time. It would take too long. (Participant 7)

Though an expressed openness for mental health work as well as previous experiences in programming and interface design seemed to increase the willingness, there was mostly strong aversion regarding answering such an amount of questions. The answering process was described as being *demotivating* and people would not feel sure if the app was worth the effort, as in the sense of uncertainty if their work would be rewarded afterwards. This is compatible with the previously pointed out strong motivator and focus on features that express progress and reward.

Another topic belonging to the *Not Helpful* code spectrum is expressed in the code about the *wording* of the items. Abstract concepts like motivation, ease of use and fitting one’s personality were hard to imagine before having ever interacted with the app;

> Also, the question, like, “as often as needed”, if I will be able to use this app as often as I needed. I don't know how often I will need it. Do I need it every day or once a week? (Participant 6)

Furthermore, though not belonging to the TWEETS itself, but still seeming to be very influential on the perception of it was the *design* of the questionnaires, which was speculated to have a negative influence on the answers. It was mentioned, that if the screenshots of the features and app versions were continuously, respectively placed next to the items, the probability of annoyance and recall decline of details by scrolling up and down to examine the features/apps again to think about a representative preference indication would be lower.

However, shifting now to more positive and *Helpful* aspects of the TWEETS, the abstract wording seemed to have resulted in *deeper reflection* of single features and app versions. This was also expressed to lead participants to ruminate about the applicability and meaning of the items; “Yes, exactly. But it also helped me to think more deeply about the features, because I was forced to examine them more often to answer the questions (Participant 8).” Furthermore, this reflection also seemed to inspire them to create different views on apps; “I think, because what I said earlier, that it opened my eyes to some dimensions of the designs that I have not yet thought of.” (Participant 4).

Moreover, though participants mostly created their preferences before examining the items, the items helped them to *support their choices* in more detail, which was expressed by participant 9 by saying;

> I think I made up my mind about the applications that I wanted to have for an app like this. I'd say pretty quick because I like the idea. I mean, I took my time read through it, but I, I kind of favourited one over the other pretty quickly. And yeah. I mean, I used the questions to express that and I think that went pretty well.

Furthermore, the items were seen as helpful especially in *discriminating* similarly liked features. In case participants liked two items equally much, the items helped them to reflect on their preferences and express them in more detail, which resulted in more detailed differentiations. And lastly, when asked if there was anything about the items that may have evoked a feeling of being guided towards preferring certain features or app versions, participants pointed out the *non-suggestive* impression of the items; “No, actually no. I think it felt like you were interested in what I think is the best version of those apps and you were interested in what I as a user liked the most (Participant 7).”

Table 9. Frequency distribution of main and subcodes regarding the topic Perceived Helpfulness of the TWEETS.

Main Code (q/n)	Subcode (q/n)
Helpful	Non-suggestive (24/10)
	Support choices (9/7)
	Deeper reflection (8/4)
	Discriminating (2/2)
Not Helpful	Preference formation (16/9)
	Demotivation (16/7)
	Design (7/3)
	Number of items (6/2)
	Wording (4/3)

Notes: q= number of quotations within this code; n= number of interviews containing this code

3.3.3 Improvement Suggestions.

Based on the interviewees gathered experiences, multiple improvement suggestions for the personalisation process as well as the app itself have been proposed, which can be split into suggestions regarding the *Personalisation Process* and the *App* itself (Table 11). Starting with the *Personalisation Process*, due to the perceived need for more *background information* about the mechanisms and handling of the app, a short *Preview* version of the app was suggested to be useful in helping to make the most customized choices in the app creation. This was suggested to take the form of a small interactive pilot version or introduction video;

> Maybe I would like a little video like a screen recording video or like a recording of like maybe the first (...) page. And then how you look in there and you know, like a (...) short tutorial, how (...) you can go through it. And like, when this

> video was done that there's a button or something under it, saying “You can download a version of the app”. (Participant 10)

Moreover, the *Personalization Procedure* at the beginning of the interaction with the mental-health app was preferred to be more *flexible and simplifiable.* In the sense of personalization, the user should have the choice to choose an extended personalization procedure like, for instance, in the study, or pick a minimalistic option, as described by participant 4’s tool kit idea;

> An individual tool kit. That would be preferable so that the person can select (...) from a variety of options. Which content they want to have in their app. So, imagine that they've created an app (...) and then it's, do you want (...) this cartoon character presenting the feedback or do you just want to have it stated there as it is. Do you want these bricks or a calendar or something like that? (...) I think this would (...) make the app much more suitable for every individual.

As also possible with this described *tool kit,* the user should have the opportunity to pick and match the preferred features so it can immediately be seen how they look together as a complete app version, which was described to make a different impression than viewing the features separately, disconnected. Then, after a certain time of interacting with the created app version, the user should be offered the option to adjust the app and switch features.

Looking now at additional attributes about the app itself, the interviewees described the influence of *visual attractiveness* and pointed out their desire to choose the colours and letter fonts, since it was expressed to cause visual discomfort and distract the participants while interacting with the app. Furthermore, participants were also interested in receiving more graphically *visualized content* like diagrams or symbolic pictures to further ease the processing of content information the app is presenting;

> I was more drawn at the images than the content. Like I also read, of course, the content that was presented, but (...) well, (...) most focus lies on the images, whether it would fit me, and then I read the text. (Participant 1)

Reminders like notification functions and the ability to connect the app with other programs and services was also mentioned to motivate interaction with the app and perform the suggested exercises. Connected to this, some of the interviewees also described the longing to expand the app’s abilities by *integrating it into daily tasks,* making it not only to a steady attribute of their lives but also connecting it with other applications like calendars, organization tools and physical activity programs.

Regarding the video feedback option there was confusion about if the video is recorded based on the users' individual experiences, or if it generally reflects on the suggested exercises. On the one side, it was perceived as discomforting when an unknown person reacts on one's mental-health accomplishments, since it is a very intimate and sensitive topic. But on the other side, there were also participants like participants like number 6, who would have preferred the therapist- or tutor-like *personalized video feedback*;

> And the thing that I like the most here is the feedback, because I can see that it's a person that gives me the feedback, which I really like. Because for me, it's easier and it makes me feel I don't know better, more motivated or ambitioned, to see that someone like a person is talking with me. And I really like that.

And finally, a last point that should be treated with care is the professional impression of an app. Independently from each other, interviewees regularly pointed out their *appreciation of effort* that was imagined to be put in the creation of the app features and the study. Participants expressed a positive mood and feeling of trust when talking about this in the interviews; "And I like it. I think that it's professional and it's consistent. (...) I like the professionalism behind it." (Participant 2).

Table 11. Frequency distribution of main and subcodes regarding the topic Improvement Suggestions.

Main Code (q/n)	Subcode (q/n)
Personalization Procedure	Flexible and simplifiable (35/9)
	More background information (28/8)
	Preview (9/6)
App	Integration in daily tasks (25/9)
	Visual attractiveness (20/10)
	Appreciation of effort (11/6)
	Visualized content (7/4)
	Reminders (4/3)
	Personalized video feedback (4/2)

Notes: q= number of quotations within this code; n= number of interviews containing this code

4. Discussion

The aim of this study was to test the applicability of the TWEETS as a personalization tool for mMH apps in a population of university students. Applying a mixed-methods approach, the results of the quantitative assessments support the hypothesized predictive as well as discriminative quality of the TWEETS. More precisely, the predictive quality is supported by the rankings of the different app versions during the second survey round, which mostly conform with the suggested app rankings based on the feature scores of the first survey round. The discriminative quality is shown in the TWEETS's ability to detect significant differences in the feature scores within the three app feature domains in the first survey round and between the different app rankings of the second survey round. Including the findings of the interviews here, it is important to mention that the participants described to use the TWEETS' items to support their feature and app preferences. Thus, the users' decision on their most favourite features or app versions was mostly made before completing the TWEETS and was less likely to be a result of the TWEETS. This leads to the conclusion that an engagement assessment tool like the TWEETS can indeed be used to detect mMH feature and app version preferences, but also as a tool to assess the underlying reasoning for those preferences.

These findings are important complements to existing research on the relationship between engagement and personalization. The fact that the participants seemed to already have chosen their most preferred features before answering the TWEETS leads to the assumption that people may adjust their expectations about future engagement accordingly. This would serve as an explanation for why treatments over which people were in control and able to adjust according to their own needs showed higher engagement, adherence, satisfaction and effectiveness compared to non-personalized treatments (Bartley, Faasse, Horne & Petrie, 2016; Cooper, Messow, McConnachie, Freire, Elliott, Heard, Williams & Morrison, 2017; Geers, Rose, Fowler, Rasinski, Brown & Helfer, 2013; Ghane, Huynh, Andrews, Legg, Tabuenca & Sweeny, 2014;). Subsequently, these findings turn the TWEETS into a useful tool to analyse and evaluate user preferences. By examining the precise scoring of the different TWEETS categories, improvement suggestions for different features could be made to increase the users expected behavioural, cognitive, or affective engagement. This way it could also be explored if and how different populations and age groups differ in their feature preferences and expected engagement scores. Subsequently, this could inspire adjustment suggestions to create more appealing feature options for different populations to increase the probability that people will make use of mMH apps and experience their benefits.

However, considering the large Cronbach's alpha of >.90 and the qualitative data from the interviews, the feasibility of the TWEETS in its original format as a personalization or evaluation tool for mMH apps outside research settings might be debatable. The interviewees predominantly emphasised that the TWEETS would include too many items. Therefore, the repetitive completing of the TWEETS for different feature domains and app versions was described to create feelings of fatigue and demotivation. Additionally, interviewees mentioned that they were less likely to complete a questionnaire like the TWEETS when privately downloading an app. This was further supported by their expressed priority of simplicity and quick reward accessibility when operating with smartphone apps, which was also found out in previous research on people’s interaction patterns and interests regarding eHealth and mMH apps (Dennison, Morrison, Conway & Yardley, 2013; Gliddon, Barnes, Murray & Michalak, 2017; McCurdie, Taneva, Casselman, Yeung, McDaniel, Ho & Cafazzo, 2012). Therefore, decreasing the TWEETS’ number of items and simplifying their wording is suggested before applying it outside the research setting.

This may require decreasing the item number to, for instance, 1 or 2 items per category (behaviour, cognitions, affect). Furthermore, the remaining items should only contain content that the users are able to apply on the presented feature options. More precisely, in case the user has not yet received information about the precise course of the intervention, items like "I will be able to use this app as often as needed." should be avoided. However, to overcome this issue and make it easier to complete the TWEETS, short informative texts or tutorials about the personalization procedure and the course of the intervention could be presented beforehand. This was also suggested during the interviews and would be in line with the findings of Torous, Nicholas, Larsen, Firth and Christensen's (2018) clinical review about user engagement regarding mental health apps which says that users of mMH apps value a good overview on how the app works and which tasks need to be done.

Furthermore, the added value of the TWEETS as a personalization tool for mMH apps should be tested. To do so it would be useful to compare the adjusted TWEETS version to a simple personalization and a no-personalization condition, to which the participants should be randomly assigned. The simple personalization condition may take the form of letting the users simply click on their most preferred feature or ranking them from most to least preferred, which was both suggested during the interviews. In the no-personalization condition, participants would receive a randomly selected app version. This is ideally suggested to be done in real life scenarios by, for instance, recruiting people who have privately downloaded a mMH app. This way it can more likely be assumed that the consumers have an internal interest in the app and are not motivated by external incentives as it was the case in this study. Additionally, by applying pre- and post-tests regarding the interventions' desired outcomes and providing options to give feedback on the personalization procedures it could be assessed if the different personalization procedures have different effects on the intervention outcomes and perceptions of these mMH apps.

Besides, it could also be examined if mMH users benefit from different personalization procedures at different time points. This way it can be tested, if the TWEETS would be of more benefit and people would be more willing to answer multiple questions after they have already interacted with the mMH app and experienced its positive effects. Thus, while mMH app users may desire a quick selection procedure at their first encounter with the app, they may be willing to invest more time for later personalization techniques. This would extend the findings of Rajanen and Rajanen's (2017) and Simon and Perlis' (2010) studies on the changing needs of app consumers with progressing time to not only affect the changing content and exercise demands but also personalization needs.

Looking at the participants' concrete choices, there were strong preference trends demonstrated in the rankings of the features and apps. The most often preferred app versions included the CBT and PPI content, Avatar feedback and the gamified Bike or Bricks design. Whereas the most frequently lowest scored features and app versions contained the ACT content, Text Only feedback and Calendar design. Explanations for this can be found in the interviews of this study as well as previous research. Advocating for the high ranking of the CBT and PPI content, students pointed out that an app that is not expected to demand a lot of effort to use and integrate into their daily lives is perceived as most attractive. Connected to this, it was also described that the desired app version should present the content in a minimalistic and illustrative way with the opportunity to interact with the program. Compared to those two, the ACT content included a lot of informative text, abstract and meaning charged concepts like meaning in life, religion and philosophy, plus little opportunity to interact with the app in the form of writing down one's experiences. This could have made it more probable for people to positively score for CBT and PPI when comparing them to the rather complicated ACT content.

Taking this into consideration, one recommendation for future mMH apps content is, firstly, to make use of simple wording when describing theories and exercises. Reading the instructions should not take much time due to the text length or complexity, which is also in line with the findings of Garrido and colleagues' (2019a; 2019b) qualitative study and review about young adults' eHealth app preferences. Delivering information in small doses may, thus, also help to keep up the user's motivation to continue working with the app and enable a quick feeling of reward and satisfaction. Secondly, the user should receive multiple opportunities to interact with the app, for instance, in the form of writing down their experiences or choosing between multiple options to express their opinions and attitudes. And thirdly, to further support the easiness of processing the content, increased use should be made of graphical illustrations and images to further simplify the information.

Supporting the gamified design choices, the students emphasized that the perceived fun and entertainment factor of apps was an important choice criterion. In addition to this it was also pointed out that the visual aesthetics of graphical progress and achievement representation as well as the easy accessibility of reward were also important. Aesthetic design was also mentioned to positively influence the perception of the app's content. Hence, in case the app had an appealing design, participants expressed a higher probability to accept the content and experience it in a positive way. This focus on entertainment, visual attractiveness and appealing and quick reward representation is also confirmed by Gowin, Cheney, Gwin and Franklin

Wann’s (2015) and Dennison, Morrison, Conway and Yardley’s (2013) qualitative studies on what kind of health apps are preferred by adolescents and Garrido and colleagues’ (2019b) review about digital mental health intervention preferences in young adults.

Hence, combining these findings, further support is found for making use of gamified design choices in mMH apps due to their motivating and appealing effects. Design choices should be made with caution since they seem to have significant influence on the way the content is perceived. Thus, a qualitative and helpful content option may be rejected when it is combined with an unappealing design. However, though the non-gamified calendar design version was less frequently preferred, there were no significant differences found between the scores of the highest ranked app versions that included a gamified design and those which included a non-gamified design. Thus, there were people expressing similar preference intensities for both types of design, which makes it advisable to keep offering the non-gamified option as well to serve everyone’s design preferences.

Additional to the focus on a rewarding progress demonstration, Middelweerd et al.’s (2015) qualitative analysis on physical activity app feature preferences in university students pointed out the desire of eHealth users’ to receive personalised feedback, which was also mentioned in the present interviews. Regarding this, interviewees of the current study emphasised that the topic of mental health was a sensitive issue, which is why interviewees were not sure if they would feel comfortable to receive feedback on their mental health achievements from a stranger via the video feedback option. Nevertheless, participants in general valued the motivational feedback for reassuring and validating one’s experiences. Hence, combining these reassurance and validation needs of the interviewees with the preference for graphic representations, the predominant favouring of the Avatar feedback seems to be a logical consequence. This is supported by the generally perceived comforting and therapeutic effect of anthropomorphic figures in electronic interventions (Amini, Lisetti, Yasavur & Rishe, 2013; Marsch, Lord, Dallery, 2014). However, it should not be concluded to only integrate Avatar feedback in mMH apps. There were also participants who favoured the Text Only version, as well as the Video version. Again, to satisfy every user’s preferences, it is recommended to continue offering all three versions of feedback. Still, it is advisable to give the users more information about the video feedback beforehand. Users should be told if the video feedback is personal and addresses the users’ individual experiences or is generally regarding possible experiences with and the meaning of a presented exercise.

Another important message that can be drawn from the interviews is that personalization itself was highly valued by the participants, though the repetitive completion of the TWEETS was described to be exhausting and demotivating. The perceived importance of and appreciation for treatment personalization is also demonstrated in previous qualitative studies on consumers' eHealth and mMH user preferences (Perski, Blandford, Ubhi, West & Michie, 2017; Torous, Nicholas, Larsen, Firth & Christensen, 2018; Cunningham, Hodgins, Toneatto, Rai & Cordingley, 2009). Furthermore, in line with findings of past research about the perceived control over treatment choices and app expressions (Brown, Oikawa, Rose, Haught, Oikawa & Geers, 2015; Geers, Rose, Fowler, Rasinski, Brown & Helfer, 2013; Ghane, Huynh, Andrews, Legg, Tabuenca & Sweeny, 2014), interviewees in the present study described that personalization created a feeling of comfort and satisfaction. Combining this with the findings of Handelzalts and Keinan's (2010), showing that people who were able to choose their upcoming treatment were not only more satisfied but also performed better and achieved significantly better results than people who were not able to choose, the current study gives further evidence for the importance of including personalization options in mMH apps.

4.1 Limitations

To the current knowledge of the researcher, the present study is the first to investigate the appropriateness of using an eHealth engagement scale to personalize a mMH app for increasing well-being in a university student population. Combining qualitative and quantitative research methods, this led to unique and authentic insights into the relationship between expected engagement and personalization, students' perception of mental health apps in general and what potential users' app personalization demands look like. However, besides the benefits of this study, there are some limitations that need to be taken into consideration.

Firstly, the randomization of the presentation order of the four app versions was done manually and resulted in an increased occurrence of suggested least-fit app versions at third position and suggested best- and second-best-fit app versions at fourth position during the second survey. Therefore, the significantly lower scoring of the third presented app compared to the apps presented at the other positions can then be taken as further support for the predictive and discriminative abilities of the TWEETS. However, due to this, there should then also be a significant difference between the third and the fourth presented app version, especially when considering the fact that the apps that were presented at fourth position were mostly the suggested best-fit or second-best-fit versions. But though the scores of the fourth app are slightly higher on average than those of the third presented app, this difference is not significant and a tentative assumption could, thus, be that participants might tend to score preferred apps that are presented in the first two positions higher than as if they are presented at the last two positions. Thus, to control for a potential effect of presentation order in future studies, the randomization should be done using appropriate tools (e.g. https://www.randomizer.org/).

Secondly, since the research was conducted by a single researcher who was of similar age and also a student of the same university as most of the participants, there is no inter-rater reliability regarding the qualitative analysis and supposedly decreased objectivity when conducting and analysing the interviews. Hence, it is well possible that additional researchers, especially those with a different background or age could have identified other additional topics and opened different perspectives on the reported content. Thirdly, since every interviewee emphasised the absence of any problems or need for help regarding mental-health issues, a tentative assumption can be made. Considering that mental well-being is an intimate and still stigmatized topic, the participants may have tried to distance themselves from any negative associations with it. Combining this with the earlier mentioned discomfort in receiving personalized feedback from a therapist-like stranger as it was expected in the video feedback feature, it could well be that the current video interview results led to different statements than a more anonymous interview procedure would have produced. Since the interviewer's video functions were activated during every interview, participants were able to observe the interviewer's mimic and gesture reactions on their comments. This may have influenced the statements of the interviewees. Hence, combining the second and third limitation, a deactivated camera function or telephone interviews, conducted by researchers who are not potentially from the same social environment may be advisable for future studies.

And fourthly, a final aspect that needs to be considered when interpreting the results is that participants in this study should not be expected to ultimately represent the target group that would get in touch with the TWEETS as a personalization tool for mMH. It is not known how motivated the current participants are to download an app to improve their mental well-being and complete a personalization procedure like the one that was presented in this study. Therefore, it is possible that people who would get in touch with mMH apps and the TWEETS out of their own motivation, not driven to receive any external incentives like research credits, would show differences in feature and app preferences and express different reactions to this procedure. Thus, to further test the feasibility of personalization procedures for mMH apps, a next step could be to test it with members of the actual target group, as it was done in the study by Bakker, Kazantzis, Rickwood and Richard (2018) on the effectiveness of mMH apps to increase public mental health. Here, participants were people who privately downloaded mMH apps and were then contacted by the researchers and asked if they would be interested to participate in their study. Doing this, individuals from a broader spectrum could be included and opinions and reactions of a more versatile population would be obtained. This in turn would also enable the examination if people from different backgrounds would have different perceptions of the TWEETS or mMH personalization needs.

4.2 Conclusion

Due to its mixed-methods approach, the findings of this study are important complements to the research about engagement and personalization in mMH interventions. While the quantitative assessment demonstrated that the TWEETS can be used to reliably detect individual differences in expected engagement regarding mMH apps and features, the qualitative data support that the indicated that these differences represent the participants' feature and app version preferences. This shows that consumers' mMH feature and app preferences can indeed be explained by their expected behavioural, cognitive, and affective engagement levels. Nevertheless, it was predominantly stated that feature and app choices were made before the scales were completed and that the participants used the TWEETS to express and support their preferences, which leads to the assumption that the TWEETS may have been less helpful in assisting the participants during their choice processes but was more supportive in reasoning their decisions. Therefore, the TWEETS may also be applicable as an evaluation tool for mMH app features. However, it is advisable to further adjust the length and wording of the TWEETS according to smartphone user demands and to test its added value compared to simpler personalization techniques before applying it in everyday scenarios. And finally, regarding the fact that the qualitative findings of the current study are in line with previous studies on mMH consumer preferences in general, it can be assumed that the suggestions based on the findings of the present study are also applicable on non-student populations as well.

References

Achilles, M. R., Anderson, M., Li, S. H., Subotic-Kerry, M., Parker, B., & O'Dea, B. (2020). Adherence to e-mental health among youth: Considerations for intervention development and research design. *DIGITAL HEALTH, 6*, 205520762092606. doi:10.1177/2055207620926064

Bakker, D., Kazantzis, N., Rickwood, D., & Rickard, N. (2016). Mental Health Smartphone Apps: Review and Evidence-Based Recommendations for Future Developments. *JMIR Mental Health, 3*(1), e7. https://doi.org/10.2196/mental.4984

Bakker, D., Kazantzis, N., Rickwood, D., & Rickard, N. (2018). A randomized controlled trial of three smartphone apps for enhancing public mental health. *Behaviour Research and Therapy, 109*, 75–83. https://doi.org/10.1016/j.brat.2018.08.003

Bartley, H., Faasse, K., Horne, R., & Petrie, K. J. (2016). You Can't Always Get What You Want: The Influence of Choice on Nocebo and Placebo Responding. *Annals of Behavioral Medicine, 50*(3), 445–451. doi:10.1007/s12160-016-9772-1

Bayram, N., & Bilgel, N. (2008). The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students. *Social Psychiatry and Psychiatric Epidemiology, 43*(8), 667–672. https://doi.org/10.1007/s00127-008-0345-x

Beiter, R., Nash, R., McCrady, M., Rhoades, D., Linscomb, M., Clarahan, M., & Sammut, S. (2015). The prevalence and correlates of depression, anxiety, and stress in a sample of college students. *Journal of Affective Disorders, 173*, 90–96. https://doi.org/10.1016/j.jad.2014.10.054

Bennett, J. W., & Glasziou, P. P. (2003). Computerised reminders and feedback in medication management: a systematic review of randomised controlled trials. *Medical Journal of Australia, 178*(5), 217–222. https://doi.org/10.5694/j.1326-5377.2003.tb05166.x

Bennett, C. B., Ruggero, C. J., Sever, A. C., & Yanouri, L. (2020). eHealth to redress psychotherapy access barriers both new and old: A review of reviews and meta-analyses. *Journal of Psychotherapy Integration, 30*(2), 188–207. https://doi.org/10.1037/int0000217

Berger, M., Wagner, T. H., & Baker, L. C. (2005). Internet use and stigmatized illness. *Social Science & Medicine, 61*(8), 1821–1827. doi:10.1016/j.socscimed.2005.03.025

Berner, E. S. (2019). Capsule Commentary on Bond et al., Real-time Feedback in Pay-for-Performance: Does More Information Lead to Improvement*? Journal of General Internal Medicine, 34*(9), 1852. https://doi.org/10.1007/s11606-019-05146-9

Borup, J., West, R. E., & Graham, C. R. (2012). Improving online social presence through asynchronous video. *The Internet and Higher Education, 15*, 195–203. http://dx.doi.org/10.1016/j.iheduc.2011.11.001

Bui, A., Veit, D., & Webster, J. (2015). Gamification – A Novel Phenomenon or a New Wrapping for Existing Concepts? *Presented at the Thirty Sixth International Conference on Information Systems*, Fort Worth. Retrieved from https://pdfs.semanticscholar.org/84a0/89cb606d3e868bc35238af2a6af3372be270.pdf

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin, 140*(4), 980–1008. https://doi.org/10.1037/a0035661

Chida, Y., & Steptoe, A. (2008). Positive Psychological Well-Being and Mortality: A Quantitative Review of Prospective Observational Studies. *Psychosomatic Medicine, 70*(7), 741–756. doi:10.1097/psy.0b013e31818105ba

Copeland, J., Rooke, S., Rodriquez, D., Norberg, M. M., & Gibson, L. (2017). Comparison of brief versus extended personalised feedback in an online intervention for cannabis users: Short-term findings of a randomised trial. *Journal of Substance Abuse Treatment, 76,* 43–48. https://doi.org/10.1016/j.jsat.2017.01.009

Cotton, V., & Patel, M. S. (2018). Gamification Use and Design in Popular Health and Fitness Mobile Applications. *American Journal of Health Promotion, 33*(3), 448–451. https://doi.org/10.1177/0890117118790394

Cunningham, J. A., Hodgins, D. C., Toneatto, T., Rai, A., & Cordingley, J. (2009). Pilot Study of a Personalized Feedback Intervention for Problem Gamblers. *Behavior Therapy, 40*(3), 219–224. https://doi.org/10.1016/j.beth.2008.06.005

Cvetkovski, S., Reavley, N. J., & Jorm, A. F. (2012). The prevalence and correlates of psychological distress in Australian tertiary students compared to their community peers. Australian & New Zealand Journal of Psychiatry, 46(5), 457–467. https://doi.org/10.1177/0004867411435290

Dale, S. (2014). Gamification. *Business Information Review, 31*(2), 82–90. doi:10.1177/0266382114538350

Davies, E. B., Morriss, R., & Glazebrook, C. (2014). Computer-Delivered and Web-Based Interventions to Improve Depression, Anxiety, and Psychological Well-Being of University Students: A Systematic Review and Meta-Analysis. *Journal of Medical Internet Research, 16*(5), e130. https://doi.org/10.2196/jmir.3142 de-Marcos, L., Domínguez, A., Saenz-de-Navarrete, J., & Pagés, C. (2014). An empirical study comparing gamification and social networking on e-learning. *Computers & Education, 75,* 82–91. doi:10.1016/j.compedu.2014.01.012

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 122. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dennison, L., Morrison, L., Conway, G., & Yardley, L. (2013). Opportunities and Challenges for Smartphone Applications in Supporting Health Behavior change: Qualitative Study. *Journal of Medical Internet Research, 15*(4), e86.https://doi.org/10.2196/jmir.2583

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness. *Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '11*. https://doi.org/10.1145/2181037.2181040

Dixon, S. (2015). The pastoral potential of audio feedback: a review of the literature. *Pastoral Care in Education, 33*(2), 96–104. https://doi.org/10.1080/02643944.2015.1035317

Donkin, L., & Glozier, N. (2012). Motivators and Motivations to Persist With Online Psychological Interventions: A Qualitative Study of Treatment Completers. *Journal of Medical Internet Research, 14*(3), e91. https://doi.org/10.2196/jmir.2100

Donkin, L., Hickie, I. B., Christensen, H., Naismith, S. L., Neal, B., Cockayne, N. L., & Glozier, N. (2013). Rethinking the Dose-Response Relationship Between Usage and Outcome in an Online Intervention for Depression: Randomized Controlled Trial. *Journal of Medical Internet Research, 15*(10), e231. https://doi.org/10.2196/jmir.2771

Eisenberg, D., Gollust, S. E., Golberstein, E., & Hefner, J. L. (2007). Prevalence and correlates of depression, anxiety, and suicidality among university students. *American Journal of Orthopsychiatry, 77*(4), 534–542. https://doi.org/10.1037/0002-9432.77.4.534

Fanning, J., Mullen, S. P., & McAuley, E. (2012). Increasing Physical Activity With Mobile Devices: A Meta-Analysis. *Journal of Medical Internet Research, 14*(6), e161. https://doi.org/10.2196/jmir.2171

Firth, J., Torous, J., Nicholas, J., Carney, R., Pratap, A., Rosenbaum, S., & Sarris, J. (2017). The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry, 16*(3), 287–298. https://doi.org/10.1002/wps.20472

Flett, J. A. M., Hayne, H., Riordan, B. C., Thompson, L. M., & Conner, T. S. (2018). Mobile Mindfulness Meditation: a Randomised Controlled Trial of the Effect of Two Popular Apps on Mental Health. *Mindfulness.10*(5), 863-876. doi:10.1007/s12671-018-1050-9

Garber, J. (2006). Depression in Children and Adolescents. *American Journal of Preventive Medicine, 31*(6), 104–125. doi: 10.1016/j.amepre.2006.07.007

Garrido, S., Cheers, D., Boydell, K., Nguyen, Q. V., Schubert, E., Dunne, L., & Meade, T. (2019). Young People's Response to Six Smartphone Apps for Anxiety and Depression: Focus Group Study. *JMIR Mental Health, 6*(10), e14385. https://doi.org/10.2196/14385

Garrido, S., Millington, C., Cheers, D., Boydell, K., Schubert, E., Meade, T., & Nguyen, Q. V. (2019). What Works and What Doesn't Work? A Systematic Review of Digital Mental Health Interventions for Depression and Anxiety in Young People. *Frontiers in Psychiatry, 10*, 759. https://doi.org/10.3389/fpsyt.2019.00759

Geers, A. L., Rose, J. P., Fowler, S. L., Rasinski, H. M., Brown, J. A., & Helfer, S. G. (2013). Why does choice enhance treatment effectiveness? Using placebo treatments to demonstrate the role of personal control. *Journal of Personality and Social Psychology, 105*(4), 549–566. doi:10.1037/a0034005

Ghane, A., Huynh, H. P., Andrews, S. E., Legg, A. M., Tabuenca, A., & Sweeny, K. (2014). The relative importance of patients' decisional control preferences and experiences. *Psychology & Health, 29*(10), 1105–1118. doi:10.1080/08870446.2014.911873

Gipson, S. Y.-M. T., Torous, J., & Maneta, E. (2017). Mobile Technologies in Child and Adolescent Psychiatry. *Harvard Review of Psychiatry, 1*. doi:10.1097/hrp.0000000000000144

Gliddon, E., Barnes, S. J., Murray, G., & Michalak, E. E. (2017). Online and mobile technologies for self-management in bipolar disorder: A systematic review. *Psychiatric Rehabilitation Journal, 40*(3), 309–319. https://doi.org/10.1037/prj0000270

Gowen, K., Deschaine, M., Gruttadara, D., & Markey, D. (2012). Young adults with mental health conditions and social networking websites: Seeking tools to build community. *Psychiatric Rehabilitation Journal, 35*(3), 245–250. https://doi.org/10.2975/35.3.2012.245.250

Gowin, M., Cheney, M., Gwin, S. & Franklin Wann, T. (2015). Health and Fitness App Use in College Students: A Qualitative Study. *American Journal of Health Education, 46*(4), 223–230. doi:10.1080/19325037.2015.1044140

Handelzalts, J. E., & Keinan, G. (2010). The effect of choice between test anxiety treatment options on treatment outcomes. *Psychotherapy Research, 20*(1), 100–112. doi:10.1080/10503300903121106

Hamari, J., Koivisto, J., & Sarsa, H. (2014, January). Does Gamification Work? -- A Literature Review of Empirical Studies on Gamification. *2014 47th Hawaii International Conference on System Sciences*. https://doi.org/10.1109/hicss.2014.377

Hamari, Juho, & Tuunanen, J. (2014). Player Types: A Meta-synthesis. *Transactions of the Digital Games Research Association, 1*(2). https://doi.org/10.26503/todigra.v1i2.13

Hayes, A. F. (2017). Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition: A Regression-Based Approach (2. Ed.). Guilford Publications.

Hetrick, S. E., Yuen, H. P., Bailey, E., Cox, G. R., Templer, K., Rice, S. M., … Robinson, J. (2017). Internet-based cognitive behavioural therapy for young people with suicide-related behaviour (Reframe-IT): a randomised controlled trial. *Evidence Based Mental Health, 20*(3), 76–82. doi:10.1136/eb-2017-102719

Huotari, K., & Hamari, J. (2016). A definition for gamification: anchoring gamification in the service marketing literature. *Electronic Markets, 27*(1), 21–31. https://doi.org/10.1007/s12525-015-0212-z

Hurling, R., Fairley, B. W., & Dias, M. B. (2006). Internet-based exercise intervention systems: Are more interactive designs better? *Psychology & Health, 21*(6), 757–772. doi:10.1080/14768320600603257

Ice, P., Swan, K., Diaz, S., Kupczynski, L., & Swan-Dagen, A. (2010). An Analysis of Students' Perceptions of the Value and Efficacy of Instructors' Auditory and Text-Based Feedback Modalities across Multiple Conceptual Levels. *Journal of Educational Computing Research, 43*(1), 113–134. https://doi.org/10.2190/ec.43.1.g

Jones, B. A., Madden, G. J., & Wengreen, H. J. (2014). The FIT Game: preliminary evaluation of a gamification approach to increasing fruit and vegetable consumption in school. *Preventive Medicine, 68*, 76–79. https://doi.org/10.1016/j.ypmed.2014.04.015

Kelders, S. M., & Kip, H. (2019). Development and initial validation of a scale to measure engagement with eHealth technologies. In *CHI EA 2019 - Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* [3312917] Association for Computing Machinery (ACM). https://doi.org/10.1145/3290607.3312917

Kelders, S. M., Kip, H. & Greeff, J. (2019). *Psychometric evaluation of the Twente Engagement with Ehealth Technologies Scale (TWEETS): Evaluation Study*. Article submitted for publication. Psychology, Health & Technology. University of Twente.

Kim, Y. H., Kim, D. J., & Wachter, K. (2013). A study of mobile user engagement (MoEN): Engagement motivations, perceived value, satisfaction, and continued engagement intention. *Decision Support Systems, 56*, 361-370.doi: 10.1016/j.dss.2013.07.002

Lalley, J. P. (1998). Comparison of Text and Video as Forms of Feedback during Computer Assisted Learning. *Journal of Educational Computing Research, 18*(4), 323–338. https://doi.org/10.2190/lxnp-wapb-vh9a-hfrw

Liquid State. (2018). *The Rise of mHealth Apps: A Market Snapshot*. Retrieved from https://liquid-state.com/mhealth-apps-market-snapshot/

McCurdie, T., Taneva, S., Casselman, M., Yeung, M., McDaniel, C., Ho, W., & Cafazzo, J. (2012). MHealth Consumer Apps: The Case for User-Centered Design. *Biomedical Instrumentation & Technology, 46*(s2), 49–56. https://doi.org/10.2345/0899-8205-46.s2.49

Middelweerd, A., van der Laan, D. M., van Stralen, M. M., Mollee, J. S., Stuij, M., te Velde, S. J., & Brug, J. (2015). What features do Dutch university students prefer in a smartphone application for promotion of physical activity? A qualitative approach. *International Journal of Behavioral Nutrition and Physical Activity, 12*(1), 31. https://doi.org/10.1186/s12966-015-0189-1

Mullins, J. K., & Sabherwal, R. (2020). Gamification: A cognitive-emotional view. *Journal of Business Research, 106*, 304–314. https://doi.org/10.1016/j.jbusres.2018.09.023

Musiat, P., Hoffmann, L., & Schmidt, U. (2012). Personalised computerised feedback in E-mental health*. Journal of Mental Health, 21*(4), 346–354. doi:10.3109/09638237.2011.648347

Naslund, J. A., Aschbrenner, K. A., Araya, R., Marsch, L. A., Unützer, J., Patel, V., & Bartels, S. J. (2017). Digital technology for treating and preventing mental disorders in low income and middle-income countries: a narrative review of the literature. *The Lancet Psychiatry, 4*(6), 486–500. https://doi.org/10.1016/s2215-0366(17)30096-2

Naslund, J. A., Marsch, L. A., McHugo, G. J. & Bartels, S. J. (2015) Emerging mHealth and eHealth interventions for serious mental illness: a review of the literature. *Journal of Mental Health, 24*(5), 321-332, doi 10.3109/09638237.2015.1019054

Ng, M. M., Firth, J., Minen, M., & Torous, J. (2019). User Engagement in Mental Health Apps: A Review of Measurement, Reporting, and Validity. *Psychiatric Services, 70*(7), 538–544. https://doi.org/10.1176/appi.ps.201800519

Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive Systems Design: Key Issues, Process Model, and System Features. *Communications of the Association for Information Systems, 24*. https://doi.org/10.17705/1cais.02428

Olesova, L. A., Richardson, J. C., Weasenforth, D. & Meloni, C. (2011). Using Asynchronous Instructional Audio Feedback in Online Environments: A Mixed Methods Study. *MERLOT Journal of Online Learning and Teaching, 7*(1). Retrieved from https://jolt.merlot.org/vol7no1/olesova_0311.pdf

Patel, V., Flisher, A. J., Hetrick, S., & McGorry, P. (2007). Mental health of young people: a global public-health challenge. *The Lancet, 369*(9569), 1302-1313.

Perski, O., Blandford, A., Ubhi, H. K., West, R., & Michie, S. (2017). Smokers' and drinkers' choice of smartphone applications and expectations of engagement: a think aloud and interview study. *BMC Medical Informatics and Decision Making, 17*(1), 25. https://doi.org/10.1186/s12911-017-0422-8

Pelletier, J.-F., Rowe, M., François, N., Bordeleau, J., & Lupien, S. (2013). No personalization without participation: on the active contribution of psychiatric patients to the development of a mobile application for mental health. *BMC Medical Informatics and Decision Making*, *13*(1). doi:10.1186/1472-6947-13-78

Punukollu, M., & Marques, M. (2019). Use of mobile apps and technologies in child and adolescent mental health: a systematic review. *Evidence Based Mental Health, 20* (4), 161-166. doi:10.1136/ebmental-2019-300093

Rajanen, D., & Rajanen, M. (2017). Personalized Gamification: A Model for Play Data Profiling. Presented at the AcademicMindtrek 2017, Tampere, Finland.Retrieved from https://www.researchgate.net/publication/320895382_Personalized_Gamification_A_Model_for_Play_Data_Profiling

Rassaei, E. (2019). Computer-mediated text-based and audio-based corrective feedback, perceptual style and L2 development. *System, 82*, 97–110. https://doi.org/10.1016/j.system.2019.03.004

Rathbone, A. L., & Prescott, J. (2017). The Use of Mobile Apps and SMS Messaging as Physical and Mental Health Interventions: Systematic Review. *Journal of Medical Internet Research, 19*(8), e295. https://doi.org/10.2196/jmir.7740

Serrano, K. J., Coa, K. I., Yu, M., Wolff-Hughes, D. L., & Atienza, A. A. (2017). Characterizing user engagement with health app data: a data mining approach. *Translational behavioral medicine, 7*(2), 277-285.

Simon, G. E., & Perlis, R. H. (2010). Personalized Medicine for Depression: Can We Match Patients With Treatments? *American Journal of Psychiatry, 167*(12), 1445–1455. https://doi.org/10.1176/appi.ajp.2010.09111680

Tal, A., & Torous, J. (2017). The digital mental health revolution: Opportunities and risks. *Psychiatric Rehabilitation Journal, 40*(3), 263–265. https://doi.org/10.1037/prj0000285

Tippey, K. G., & Weinger, M. B. (2017). User-Centered Design Means Better Patient Care. *Biomedical Instrumentation & Technology, 51*(3), 220–222. doi:10.2345/0899-8205-51.3.220

Torous, J., Staples, P., & Onnela, J.-P. (2015). Realizing the Potential of Mobile Mental Health: New Methods for New Data in Psychiatry. *Current Psychiatry Reports, 17*(8). doi:10.1007/s11920-015-0602-0

Torous, J., Wisniewski, H., Liu, G., & Keshavan, M. (2018). Mental Health Mobile Phone App Usage, Concerns, and Benefits Among Psychiatric Outpatients: Comparative Survey Study. *JMIR Mental Health, 5*(4), e11715. https://doi.org/10.2196/11715

Torous, J., Nicholas, J., Larsen, M. E., Firth, J., & Christensen, H. (2018). Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence Based Mental Health, 21*(3), 116–119. doi:10.1136/eb-2018-102891

van Gemert-Pijnen, L., Kelders, S. M., Kip, H., & Sanderman, R. (2018). *EHealth Research, Theory and Development: A Multi-Disciplinary Approach* (1st ed.). Abingdon, United Kingdom: Routledge.

Vockley, M. (2015). The Rise of Telehealth: ‘Triple Aim,’ Innovative Technology, and Popular Demand Are Spearheading New Models of Health and Wellness Care. *Biomedical Instrumentation & Technology, 49*(5), 306-320.doi:10.2345/0899-8205-49.5.306

Wallace, A. M., Bogard, M. T., & Zbikowski, S. M. (2018). Intrapersonal Variation in Goal Setting and Achievement in Health Coaching: Cross-Sectional Retrospective Analysis. *Journal of Medical Internet Research, 20*(1), e32. https://doi.org/10.2196/jmir.8892

Yardley, L., Spring, B. J., Riper, H., Morrison, L. G., Crane, D. H., Curtis, K., … Blandford, A. (2016). Understanding and Promoting Effective Engagement With Digital Behavior Change Interventions. *American Journal of Preventive Medicine, 51*(5), 833–842.

Yee, N. (2006). Motivations for Play in Online Games. *CyberPsychology & Behavior, 9*(6), 772–775. doi:10.1089/cpb.2006.9.772

Zichermann G, Cunningham C. *Gamification By Design: Implementing Game Mechanics In Web And Mobile Apps*. Sebastopol, California: O'Reilly Media; 201

Appendix A

TWente Engagement with Ehealth Technologies Scale (TWEETS)

The TWEETS can be used to measure engagement at different moments in time.

- After first (day of) use: **expectation** of engagement
- During usage: **current** engagement
- After finishing usage/when intervention is finished: **past** engagement

Please indicate to what extent you agree with the following statement:

5-point Likert scale: strongly disagree, disagree, neutral, agree, strongly agree

Expectations of engagement

I think

1. Using this [technology] can become part of my daily routine
2. This [technology] is easy to use
3. I will be able to use this [technology] as often as needed (to achieve my goals)
4. This [technology] will make it easier for me to work on [goal of the technology, e.g. increasing my well-being]
5. This [technology] will motivate me to [goal of the technology]
6. This [technology] will help me to get more insight into my [goal of the technology, e.g. well-being]
7. I will enjoy this [technology]
8. I will enjoy seeing the progress I make in this [technology]
9. This [technology] will fit me as a person

Current engagement

Thinking about using this [technology] **recently, I feel**

1. Using this [technology] is part of my daily routine
2. The [technology] is easy to use
3. I’m able to use the [technology] as often as needed (to achieve my goals)
4. This [technology] makes it easier for me to work on [goal of the technology, e.g. increasing my well-being]
5. This [technology] motivates me to [goal of the technology]

6. This [technology] helps me to get more insight into my [goal of the technology, e.g. well-being]
7. I enjoy using this [technology]
8. I enjoy seeing the progress i make in this [technology]
9. This [technology] gits me as a person

Past engagement

Looking back at using the [technology]**, I feel that**

1. Using this [technology] did become part of ma daily routine
2. The [technology] was easy to use
3. I was able to use the [technology] as often as needed (to achieve my goals)
4. This [technology] mad it easier for me to work on [goal of the technology, e.g. well-being]
5. This [technology] motivated me to [goal of the technology]
6. This [technology] helped me to get more insight into my [goal of the technology, e.g. well-being]
7. I enjoyed using this [technology]
8. I enjoyed seeing the progress I made in this [technology]
9. This [technology] fits me as a person

Notes:

- Items 1,2,3 cover behavioral engagement, items 4,5,6 cover cognitive engagement, items 7,8,9 cover affective engagement
- Each item can be scored on a 5-point Likert-scale: strongly disagree (0), disagree (1), neutral (2), agree (3), strongly agree (4). Scores can be totaled per component (range 0-12) and overall (range 0-36)

Appendix B

Bar charts on the frequencies of app ranks that were presented at different presentation position in the second survey round

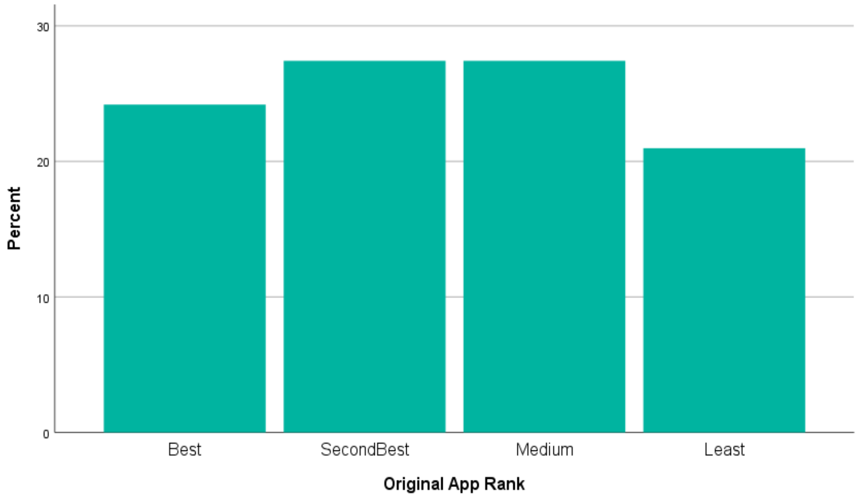


Figure 13. Frequencies of app versions that were presented at first position in the second survey round, based on expected engagement feature scores of the first survey round.

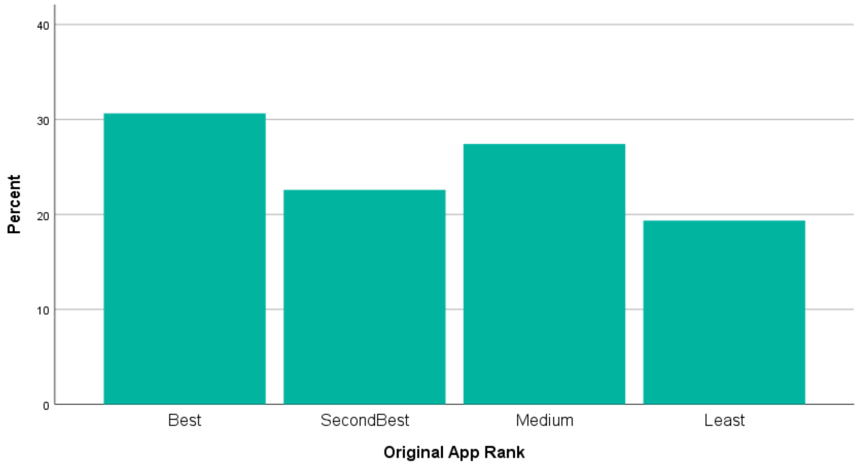


Figure 14. Frequencies of app versions that were presented at second position in the second survey round, based on expected engagement feature scores of the first survey round.

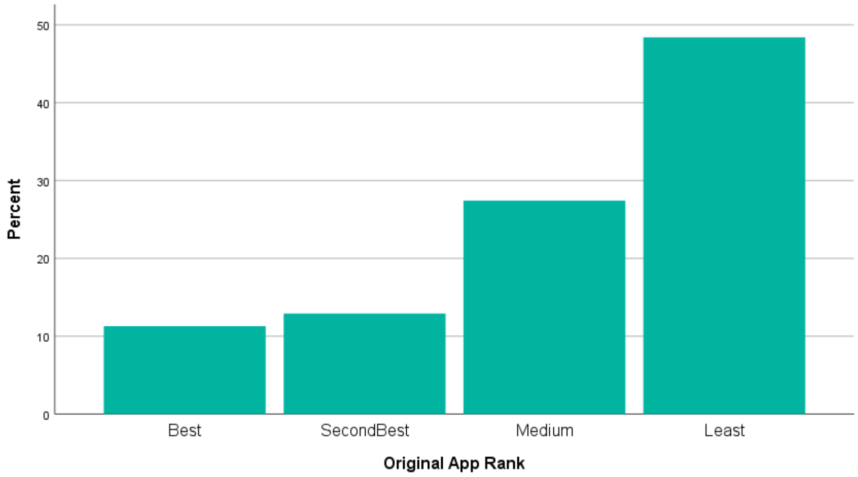


Figure 15. Frequencies of app versions that were presented at third position in the second survey round, based on expected engagement feature scores of the first survey round.

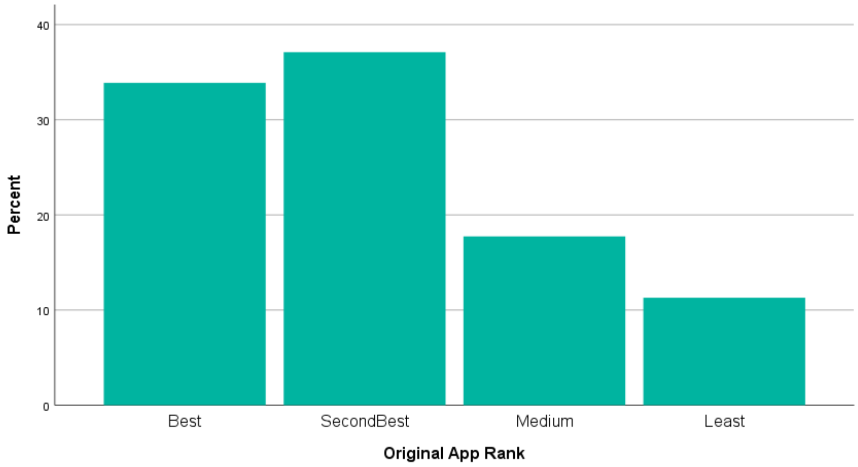


Figure 16. Frequencies of app versions that were presented at fourth position in the second survey round, based on expected engagement feature scores of the first survey round.

Appendix C

Tables of ANOVA results regarding app rank mean scores of the first and second survey

Table 12. *Frequency Distributions of and Differences between Second-Best-Fit Features per App Domain based on TWEETS results of 1st and 2nd Survey.*

Domain	Feature	*n* (*%*)	mean (*SD*)	95% CI	*F*	*p*
1st Survey	CBT	17 (*27.4*)	6.82 (*1.74*)	5.92 - 7.71	.29	(*0.83*)
Content	ACT	5 (*8.1*)	7.07 (*.77*)	6.11 - 8.02		
	PPI	39 (*62.9*)	6.55 (*1.34*)	6.11 - 6.98		
Feedback	Video	18 (*29*)	6.76 (*1.40*)	6.06 - 7.45	.84	(*0.44*)
	Avatar	40 (*64.5*)	6.71 (*1.42*)	6.25 - 7.16		
	Text	4 (*6.5*)	5.78 (*.65*)	3.72 - 7.85		
Design	Bricks	14 (*22.6*)	6.60 (*1.46*)	5.76 - 7.44	.17	(*0.85*)
	Bike	34 (*54.8*)	6.61 (*1.43*)	6.11 - 7.11		
	Calendar	14 (*22.6*)	6.85(*1.40*)	6.05 - 7.66		
Total	ALL	62	6.66 (*1.41*)	6.30 - 7.02	.78	(*0.70*)
2nd Survey	CBT	12	5.43 (*1.60*)	4.41 - 6.44	5.08	(*.01*)*
Content	ACT	8	7.32 (*1.54*)	6.03 - 8.61		
	PPI	27	6.50 (*1.16*)	6.04 - 6.96		
Feedback	Video	15	6.00 (1.75)	5.03 - 6.97	1.42	(*.25*)
	Avatar	23	6.35 (*1.19*)	5.83 - 6.86		
	Text	9	7.02 (*1.51*)	5.86 - 8.19		
Design	Bricks	16	6.35 (*1.46*)	5.58 - 7.13	2.72	(*.08*)
	Bike	17	5.84 (*1.47*)	5.08 - 6.59		
	Calendar	14	7.02 (*1.26*)	6.29 - 7.75		
Total	ALL	47	6.37 (*1.46*)	5.94 - 6.79	3.74	(*.003*)**

Note: n= Sample Size; SD = Standard Deviation; *p<.05 ; **p< .01

Table 13. *Frequency Distributions of and Differences between Medium Fit Features per App Domain based on TWEETS results of 1st and 2nd Survey.*

Domain	Feature	*n* (*%*)	mean (*SD*)	95% CI	*F*	*p*
1st Survey	CBT	19 (*30.6*)	5.33 (*1.54*)	4.59 - 6.06	.47	(*.71*)
Content	ACT	29 (*46.8*)	5.80 (*1.26*)	5.32 - 6.27		
	PPI	13 (*21*)	5.63 (*1.94*)	4.46 - 6.80		
	ALL	1 (*1.6*)	6.44	6.44		
Feedback	Video	18 (*29*)	5.45 (*1.41*)	4.75 - 6.15	.33	(.72)
	Avatar	19 (*30.6*)	5.85 (*1.60*)	5.08 - 6.62		
	Text only	25 (40.3)	5.98 (*1.50*)	4.97 - 6.21		
Design	Bricks	14 (*22.6*)	5.39 (*1.39*)	4.59 - 6.20	.43	(*0.65*)
	Bike	34 (*54.8*)	5.61 (*1.52*)	5.08 - 6.14		
	Calendar	14 (*22.6*)	5.92 (*1.55*)	5.02 - 6.82		
Total	ALL	62	5.63 (*1.49*)	5.25 - 6.01	1.15	(*.35*)
2nd Survey	CBT	20 (*32.3*)	6.03 (*1.49*)	5.34 - 6.73	1.15	(*.32*)
Content	ACT	16 (*25.8*)	5.35 (*1.29*)	4.66 - 6.03		
	PPI	14 (*22.6*)	5.39 (*1.79*)	4.31 - 6.47		
Feedback	Video	13 (*21*)	5.76 (*1.81*)	4.67 - 6.85	.31	(*.73*)
	Avatar	22 (*35.5*)	5.75 (*1.42*)	5.11 - 6.40		
	Text	15 (*24.2*)	5.38 (*1.52*)	4.58 - 6.18		
Design	Bricks	16 (*25.8*)	6.00 (*1.72*)	5.05 - 6.95	1.44	(*.25*)
	Bike	22 (*35.5*)	5.24 (*1.52*)	4.56 - 5.91		
	Calendar	12 (*19.4*)	5.93 (*1.13*)	5.21 - 6.65		
Total	ALL	50	5.64 (*1.52*)	5.20 - 6.08	.87	(*.62*)

Note: n= Sample Size; SD = Standard Deviation

Table 14. *Frequency Distributions of and Differences between Least-Fit Features per App Domain based on TWEETS results of 1st and 2nd Survey.*

Domain	Feature	*n* (*%*)	mean (*SD*)	95% CI	*F*	*p*
1st Survey	CBT	25 (*40.3*)	4.12 (*1.18*)	3.64 - 4.61	2.30	(*.09*)
Content	ACT	27 (*43.5*)	3.60 (*1.44*)	3.03 - 4.17		
	PPI	9 (*14.5*)	4.84 (*1.61*)	3.61 - 6.08		
	ALL	1 (*1.6*)	6.44	6.44		
Feedback	Video	25 (*40.3*)	3.95 (*1.56*)	3.30 - 4.59	.22	(*.80*)
	Avatar	3 (*4.8*)	3.61 (*1.43*)	.05 - 7.16		
	Text	34 (*54.8*)	4.11 (*1.31*)	3.65 - 4.56		
Design	Bricks	12 (*19.4*)	3.91 (*1.41*)	3.01 - 4.80	3.52	(*.02**)
	Bike	11 (*17.7*)	4.70 (*.84*)	4.14 - 5.26		
	Calendar	38 (*61.3*)	3.77 (*1.40*)	3.31 - 4.23		
	ALL	1 (*1.6*)	7.33			
Total	ALL	62	4.02 (*1.41*)	3.66 - 4.38	1.63	(*.10*)
2nd Survey	CBT	18	4.50 (*1.68*)	3.67 - 5.33	1.23	(*.30*)
Content	ACT	22	4.39 (*1.57*)	3.70 - 5.09		
	PPI	13	5.32 (*2.20*)	3.99 - 6.66		
Feedback	Video	21	4.91 (*1.88*)	4.06 - 5.77	1.66	(*.20*)
	Avatar	9	5.29 (*2.03*)	3.74 - 6.85		
	Text	23	4.18 (*1.54*)	3.51 - 4.84		
Design	Bricks	10	5.94 (*1.37*)	4.96 - 6.93	9.04	(*.00*)*
	Bike	17	5.27 (*1.72*)	4.39 - 6.16		
	Calendar	26	3.76 (*1.52*)	3.15 - 4.37		
Total	ALL	53	4.66 (*1.78*)	4.17 - 5.15	1.94	(*.05*)*

Note: n= Sample Size; SD = Standard Deviation; * p≤ .05

Appendix D

Bar charts about the distributions of app versions per rank of the first and second survey.

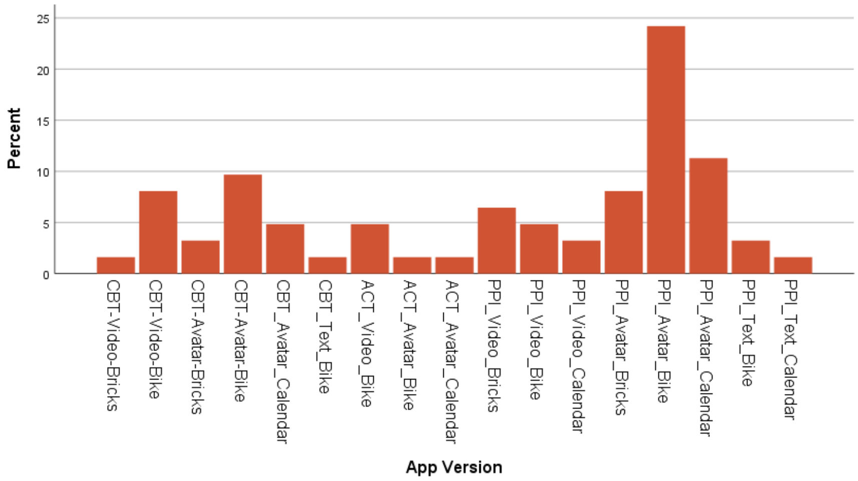


Figure 17. Frequencies of second-best-fit app versions based on feature preferences in the 1st survey.

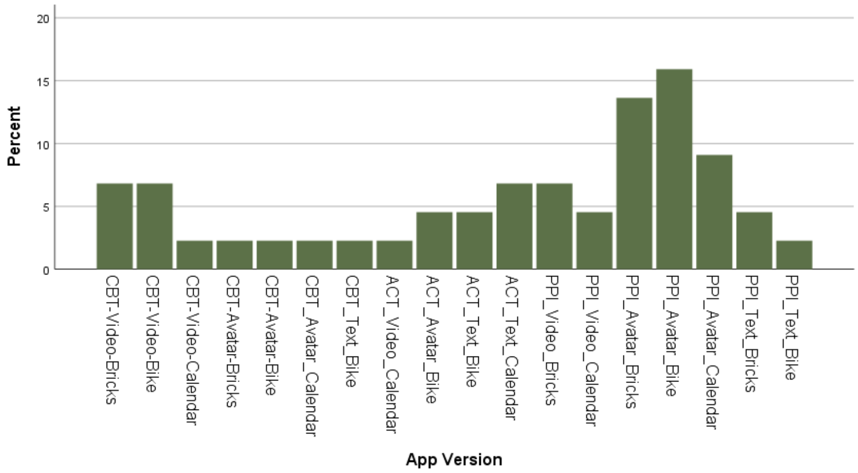


Figure 18. Frequencies of second-best-fit app versions based on app preferences in the 2nd survey.

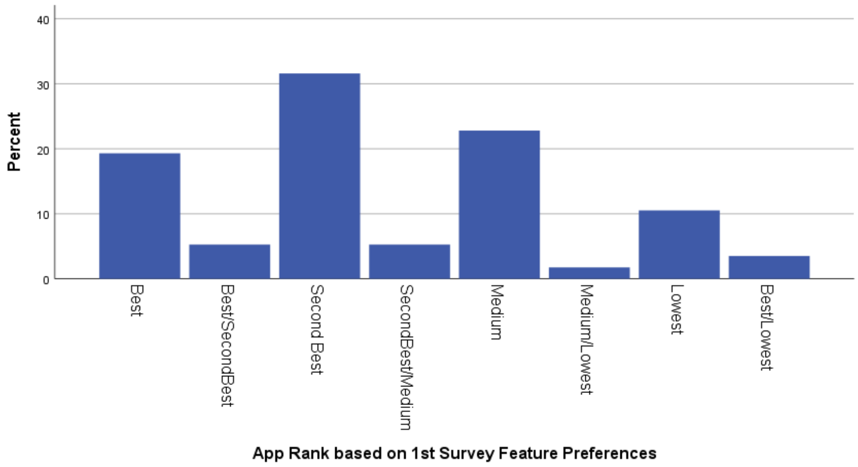


Figure 19. Frequencies of which app versions were second most preferred in the second survey.

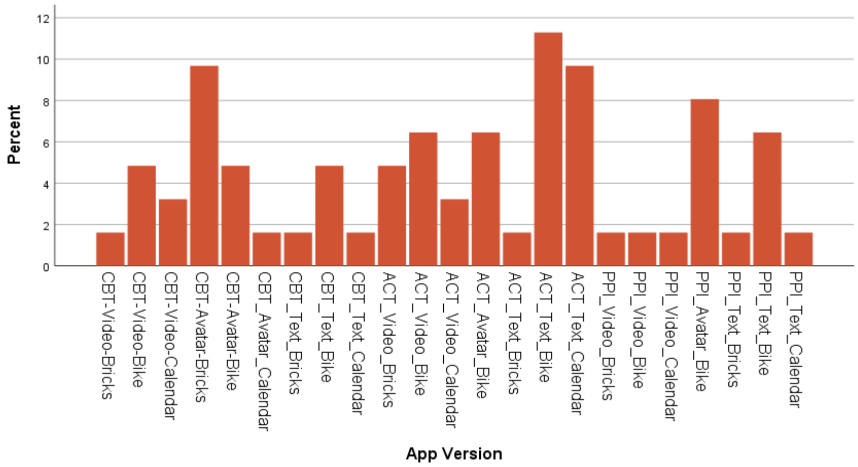


Figure 20. Frequencies of medium fit app versions based on feature preferences in the 1st survey.

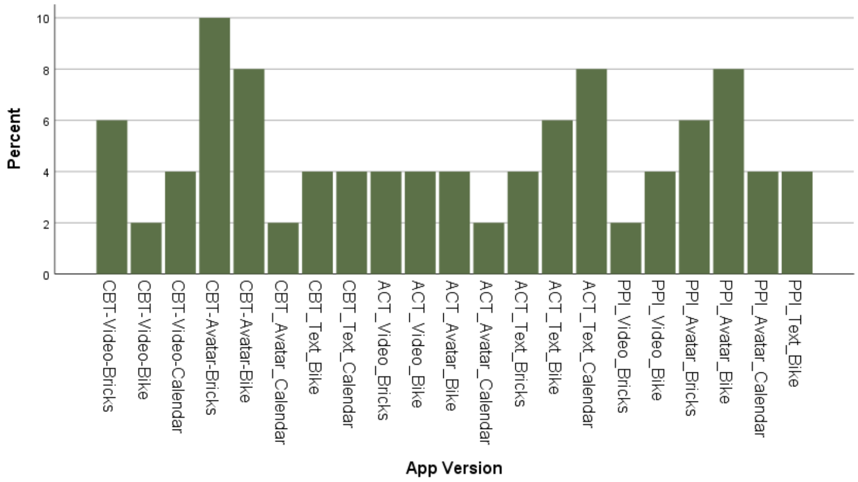


Figure 21. Frequencies of medium fit app versions based on app preferences in the 2nd survey.

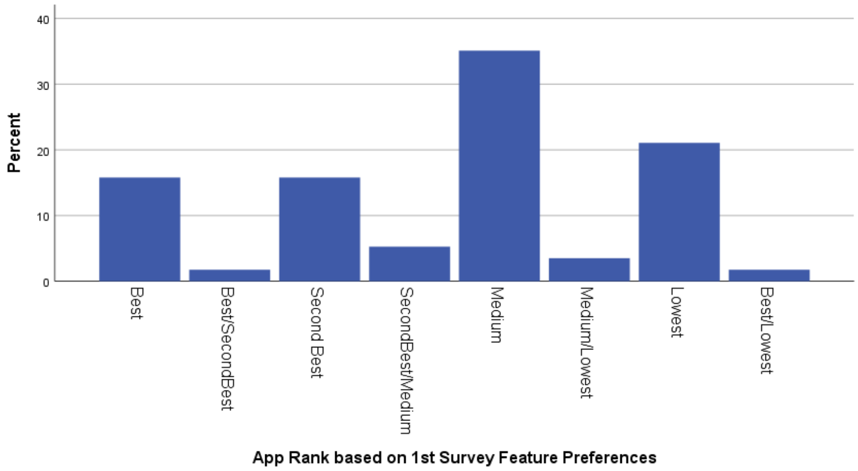


Figure 22. Frequencies of which app versions were medium preferred in the second survey.

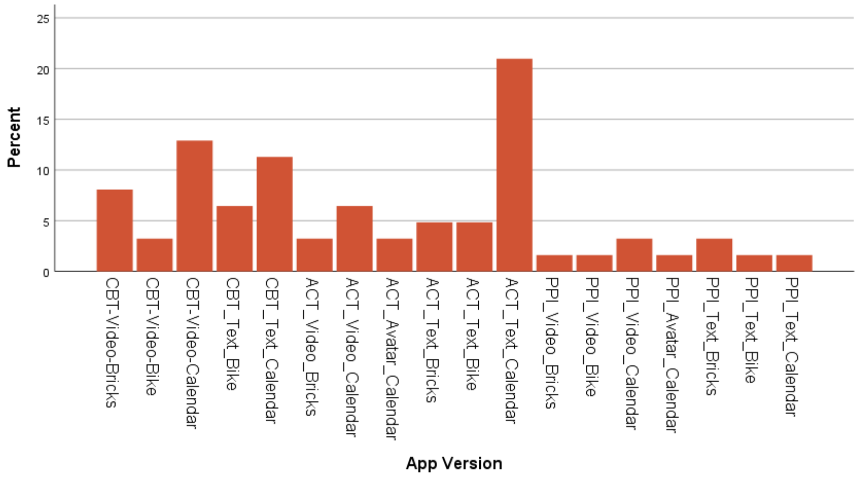


Figure 23. Frequencies of least fit app versions based on feature preferences in the 1st survey.

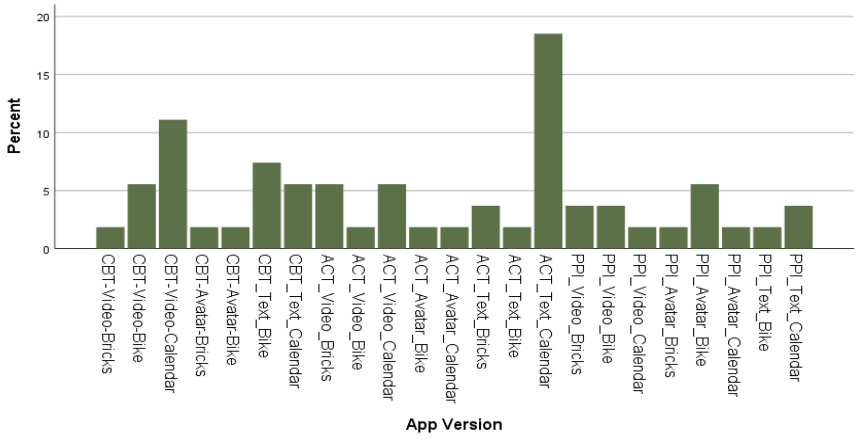


Figure 24. Frequencies of least fit app versions based on app preferences in the 2nd survey.

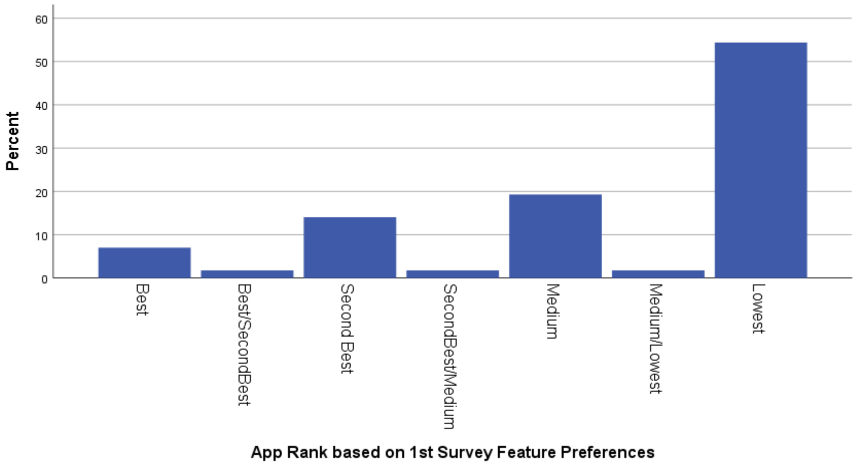


Figure 25. Frequencies of which app versions were least preferred in the second survey.