CHAPTER 6

GENERAL DISCUSSION
The aim of the Active2Gether project described in this dissertation was to develop and to evaluate an app-based intervention targeting physical activity behaviors in young adults. The aims of this dissertation were therefore fourfold; 1) To gain insights in the publicly available apps that promote physical activity; 2) To gain insights in the preferences of features of such smartphone apps of young adults; 3) To develop a mHealth intervention promoting physical activity among young adults; 4) To evaluate the intervention to inform further development and future mHealth interventions. In this final chapter, the main findings, methodological challenges, and the implications for future research and interventions/practices are summarized and discussed.

Discussion of the Main Findings

Insights in the publicly available apps that aim to promote physical activity

Since health promoting apps – including physical activity apps – were relatively new in 2012, we aimed to do an inventory of the state-of-the-art of smartphone apps promoting physical activity. An inventory and analysis of the two largest app stores for the presence and contents of available physical activity promotion apps was conducted to inform the development of an app-based intervention. To do so, we conducted two reviews analyzing the apps from two different perspectives: the first focused on identifying and reviewing behavior change techniques (BCTs) that were included in existing physical activity apps, whereas the second review focused on the technological possibilities and opportunities in physical activity apps.

The first review (described in Chapter 2), identified 64 unique apps that targeted physical activity, used tailored feedback and were available in iTunes and Google Play Store. The use of BCTs in these apps were reviewed based on an established taxonomy of such techniques. On average, these apps included five different BCTs and none of the apps used less than two or more than eight. There was no observed difference in the number of identified BCTs between free-of-charge apps and apps that require a fee. The most frequently used BCTs in apps were goal setting, self-monitoring and feedback on performance, which is equivalent to BCTs most frequently used in other types of physical activity promotion interventions. The review also indicated that apps have the potential to provide tailored feedback and to integrate BCTs. However, based on these findings, it was clear that there is room for improvement regarding the number of applied techniques. Especially inclusion of BCTs is warranted that were hardly implemented to date but are potentially effective, such as motivational interviewing, stress management, relapse prevention, self-talk, role modeling, and prompted barrier identification.

The second review (described in Chapter 2) covered 169 unique apps available in the App Store and Google Play Store that aimed to promote physical activity. This review aimed to investigate the use of...
AIMS

The aim of the Active2Gether project described in this dissertation was to develop and to evaluate an app-based intervention targeting physical activity behaviors in young adults. The aims of this dissertation were therefore fourfold; 1) To gain insights in the publicly available apps that promote physical activity; 2) To gain insights in the preferences of features of such smartphone apps of young adults; 3) To develop a mHealth intervention promoting physical activity among young adults; 4) To evaluate the intervention to inform further development and future mHealth interventions. In this final chapter, the main findings, methodological challenges, and the implications for future research and interventions/practices are summarized and discussed.

DISCUSSION OF THE MAIN FINDINGS

Insights in the publicly available apps that aim to promote physical activity

Since health promoting apps – including physical activity apps - were relatively new in 2012, we aimed to do an inventory of the state-of-the-art of smartphone apps promoting physical activity. An inventory and analysis of the two largest app stores for the presence and contents of available physical activity promotion apps was conducted to inform the development of an app-based intervention. To do so, we conducted two reviews analyzing the apps from two different perspectives: the first focused on identifying and reviewing behavior change techniques (BCTs) that were included in existing physical activity apps, whereas the second review focused on the technological possibilities and opportunities in physical activity apps.

The first review (described in Chapter 2), identified 64 unique apps that targeted physical activity, used tailored feedback and were available in iTunes and Google Play Store. The use of BCTs in these apps were reviewed based on an established taxonomy of such techniques. On average, these apps included five different BCTs and none of the apps used less than two or more than eight. There was no observed difference in the number of identified BCTs between free-of-charge apps and apps that require a fee. The most frequently used BCTs in apps were goal setting, self-monitoring and feedback on performance, which is equivalent to BCTs most frequently used in other types of physical activity promotion interventions. The review also indicated that apps have the potential to provide tailored feedback and to integrate BCTs. However, based on these findings, it was clear that there is room for improvement regarding the number of applied techniques. Especially inclusion of BCTs is warranted that were hardly implemented to date but are potentially effective, such as motivational interviewing, stress management, relapse prevention, self-talk, role modeling, and prompted barrier identification.

The second review (described in Chapter 2) covered 169 unique apps available in the App Store and Google Play Store that aimed to promote physical activity. This review aimed to investigate the use of
technical features in these apps that can be used to monitor or encourage physical activity. To do so, we have developed a scoring tool to assess these features. This review showed that on average eight out of 37 features were found in these apps. The features that were used most were ‘user input’ (i.e. to log activities or to create a personal profile), ‘a textual/numerical overview of the user’s behavior and progress’, ‘sharing achievements or workouts in internal or external social networks’, and ‘general advice on physical activity’. The features that were used least often were ‘adaptation to the user’ (e.g. does the app automatically adjust the goals to the user’s behavior), ‘integration with external sources’, and ‘encouragement through gamification’, some form of punishment or ‘the possibility to contact an expert through the app’. There were no differences found between apps from the Google Play or the App Store, nor between free and paid apps. The results indicated that apps can be improved substantially in terms of their utilization of the possibilities that current mobile technology offers.

In summary: publicly available physical activity apps mainly use goals-setting and self-monitoring and feedback on performance techniques like the ones most frequently used in other types of physical activity promotion interventions. Although apps clearly have the possibility to include theory-based BCTs, they often fail to use the full potential of possible BCTs as well as the technological possibilities. Very few apps dynamically adapt the content and feedback and link with external monitors. We concluded that in to go beyond existing apps, the Active2Gether intervention should include at least BCTs as self-monitoring, goals-setting and feedback on performance and it should include the new technology to dynamically adapt the content of the app to the user’s needs and preferences.

**Insights in the preferences of smartphone features of young adults**

Physical activity apps were relatively new in health promotion when the studies in this thesis were conducted and little was known about how the target population – young adults – used apps in general and physical activity apps in particular. To gain more insights on how their preferred or more ideal physical activity app would look like, we conducted focus group discussions (described in Chapter 3).

We aimed to explore the use and appreciation of and the preferences for various features of a physical activity app. Thirty Dutch students (aged 18–25 years) used a physical activity app – the *Nexercise App* - for three weeks and subsequently attended a focus group discussion (k = 5). As expected, based on the popularity of health and fitness apps, participants expressed positive attitudes toward the physical activity app. Participants who reported to meet physical activity guidelines thought that such apps are more useful for others than for themselves, stating that physical activity apps such as *Nexercise* could raise awareness for those who are not physically active, but that such apps are not suitable for themselves. Those who did not meet physical activity guidelines highlighted a need for a personal coach function in such apps to help them achieve their self-determined goals, whereas those who met the guidelines preferred detailed training information, such as how to intensify their training sessions.
Almost all participants preferred a companion website that could give detailed and general information about their behaviors. Based on these focus group results we concluded that apps aiming to promote and assist increasing physical activity in young adults should provide personalized and tailored feedback and include a coaching function. Furthermore, such apps should include an easy-to-use design and the option to customize the application. Preferred features to be included in physical activity apps are ranking features, a coaching feature to motivate users during the exercise and to provide feedback afterwards, and the possibility to set goals and to work exercise according to a schedule. In addition, participants preferred a website that accompanies the app to provide overviews of their results and progress. There appears to be little need for a sharing feature to post results through social media.

The focus group discussions further pointed out that the participants preferred the BCTs that are already most likely to be present in existing physical activity apps. However, as they also wanted the app to provide personalized and tailored feedback and to be able to customize the app, we wanted to learn to what characteristics the feedback should be tailored, i.e. what the preferred tailoring variables are. In other words: we wanted to explore if the specific BCTs used, should be tailored to personal characteristics of the users, such as personality traits, level of self-efficacy and/or sport identity. To achieve that, we conducted an online survey to assess young adults’ ratings of BCTs applied in apps and analyze associations of these ratings with personal characteristics (described in Chapter 3). The cross-sectional online survey was conducted among healthy 18-30-year-old adults (N=179). They rated a selection of important and often-used BCTs (i.e. goal setting and goal reviewing, feedback and self-monitoring, and social support and social comparison) based on their experiences, wishes and/or requirements in a smartphone physical activity app. Ratings of the BCTs addressing self-efficacy were relatively high, in contrast to the ratings for BCTs addressing social support. Agreeableness was positively associated with the ratings for ‘goal setting and goal reviewing’. While neuroticism was inversely associated with ratings on ‘feedback and self-monitoring’ and self-efficacy was positively associated with ratings on ‘feedback and self-monitoring’. To conclude, ratings of various self-regulation BCTs in a smartphone app were high in a selected group of highly educated and physically active young adults. BCTs addressing social support were less appreciated. Differences in ratings of BCTs due to differences in personality and exercise self-efficacy between young adults should be considered.

**Development of an intervention aimed at promoting physical activity among young adults**

The results of the systematic review (described in Chapter 2) revealed that publicly available apps almost never include connection to an external sensor, such as a fitness tracker. Furthermore, the systematic review pointed out that dynamic tailoring by iterative assessments and providing feedback...
Chapter 6

– i.e. longitudinal feedback – was associated with larger effect sizes compared to static tailoring indicating the need for an ongoing assessment of the physical activity behavior and associated feedback and advice. This is in line with the results of the focus group discussions (described in Chapter 3), where participants who did not report meeting the physical activity guidelines preferred a personal coach feature in their app to provide feedback and advice, taking into account changes made over time. For Active2Gether, the Fitbit One was selected to assess physical activity behavior in daily life because; a) the activity monitor wirelessly uploads the activity data to the user’s account, which is accessed through a smartphone application or website; b) It assesses daily activities, allows the user to monitor his/her progress and provides real-time feedback (e.g. current amount of steps taken, stairs ascended and a growing or shrinking flower depicts the activity level) to encourage the user to be more physically active; c) The Fitbit One and its web interface include BCTs typically used in physical activity interventions. However, before using such a new device for research purposes it is important to know the validity of the Fitbit One among young adults. We conducted a validation study that compared Fitbit One measurements with ActiGraph GT3X+ in terms of assessing physical activity in different time intervals, to assess the construct validity of the Fitbit One for providing real-time physical activity feedback among young adults (described in Chapter 4). The results indicated that although the Fitbit One systematically overestimates physical activity as assessed by the ActiGraph, Fitbit One did show excellent associations with ActiGraph measures for step counts, and fair to good associations for time spent in moderate, vigorous physical activity and combined moderate-vigorous physical activity. These measurement properties of the Fitbit One were similar for the short time intervals (i.e. minutes and hours) and longer time intervals (i.e. days), which suggests that the Fitbit One is relatively well-suited for providing real-time feedback on steps activity and for self-monitoring. However, to evaluate the efficacy of the Active2Gether intervention regarding minutes of moderate-vigorous physical activity an additional activity monitor is needed, such as the ActiGraph.

We used a 7-step systematic approach to develop the Active2Gether intervention (described in Chapter 4). The Active2Gether app is a web-based application suitable for Android phones running on a 4.0 version or higher. The app specifically targets active sports participation, stair use, and active transport to increase levels of overall weekly moderate-vigorous physical activity. The theoretical framework was largely based on Social Cognitive Theory and a selection of BCTs were selected (e.g. self-monitoring, goals setting, social comparison, modelling, action planning, prompt instructions, barrier identification). Since dynamically tailoring by iteratively assessing and providing feedback has been associated with larger effect sizes compared to static tailoring, we aimed to provide personalized, real-time and context specific feedback continuously assessed by an activity tracker. Using dynamically and real-time feedback is relatively new in health promotion and is still lacking in
publicly available apps, but the results of the focus group discussions indicate that participants would appreciate such feedback. To provide such feedback, we designed a model-based reasoning system.

In the attempt to go beyond the state of the art of existing apps the content was highly tailored and personalized, using a reasoning engine that generates conclusions from information stored in the different databases. Combining the data that has been collected via the monitoring components with information about the user’s context, the reasoning engine enables personalization of the coaching strategies to try to realize the most positive effect on behavior change.

The result of the systematic development was a web-based android app that sends daily messages tailored to physical activity behavior determinants and that encourages and motivates users to be more physically active, i.e. to participate in sports and use active transport and the stairs more often. The app displays the weekly goal, Fitbit data (i.e. number of steps and stairs ascended) and it shows three graphs containing weekly steps data, coaching data, i.e. weekly minutes of sports, active transport or number of stairs ascended and the last graph compares the users’ activity data to other users.

**Evaluation of the Active2Gether intervention to inform further development and future mHealth interventions**

As part of the development of the Active2Gether intervention, the app-based intervention was evaluated with respect to its efficacy and its usability. The efficacy of the Active2Gether intervention was explored in a quasi-experimental trial in which the Active2Gether intervention was compared with a ‘light’ version of Active2Gether — without tailored feedback and coaching messages — and with the Fitbit app (Chapter 5). Furthermore, the Active2Gether intervention was evaluated with respect to its usability (i.e. adherence, interaction rates) and users’ appreciation (Chapter 5).

The first study aimed to explore whether the two versions of the Active2Gether apps — the version with (Active2Gether-Full) and without (Active2Gether-Light) a coaching — were more effective in increasing levels of physical activity among young adults than an existing self-monitoring app (the Fitbit app). The secondary aims of the study were to examine whether the intervention was effective in changing levels of relevant behavioral determinants of physical activity and how participants used and evaluated the app. No evidence for significant superior intervention effects of Active2Gether on increased physical activity or more positive determinants of physical activity as compared to the ‘light’ version or Fitbit were found. The majority of the Active2Gether app users used the app at least several times per week, but was not satisfied with the app. In addition, a substantial number of participants experienced technical problems. The evaluation study that was conducted differed substantially from the original protocol, which had impact on randomization and power. Thus, the results of this study
should be interpreted with caution. The violations of the original study protocol and its consequences were described in detail in Chapter 5.

The second study aimed to evaluate the use (i.e. adherence, interaction rates) and the users’ appreciation of the Active2Gether app. In summary, significant differences were seen in duration of usage of the Fitbit in favor of the two Active2Gether apps. Furthermore, significant differences were seen with respect to the appreciation between the Fitbit app and the two Active2Gether apps; the Fitbit app scored higher in terms of ‘Satisfaction’, ‘User friendliness’, ‘Perceived effectiveness’ and ‘Professionality’. Reasons for the relatively low scores could be that users of the two Active2Gether apps reported quite some technical problems (54 and 45 percent respectively). Lastly, the results revealed that users of the Active2Gether-Full app appreciated coaching (on top of self-monitoring functionalities), but are also critical regarding its implementation (in terms of the number and content of the messages). It is important to find a balance in the number of messages sent; too many messages were perceived as annoying, while too few might not be effective.

METHODOLOGICAL CHALLENGES

We used a stepwise approach to develop the intervention. During this process, we encountered several methodological challenges on which we based a list of recommendations and important points for consideration for developing future mHealth and particularly app-based interventions.

Challenges in conducting physical activity research among young adults

The work described in this dissertation focused on young adults and despite the well-designed recruitment strategies, selection bias may have resulted – also because the original design of the studies could not always be realized – in a lower external validity (generalization of the results). In all studies, a selection of eligible participants was included that may have resulted in a study population that was not representative for the broader population. Therefore, generalization of the Active2Gether results (described in Chapters 4 and 5) should be done with care. The participants who enrolled for the Active2Gether studies were mainly highly educated Dutch young adult women (described in Chapters 3, 4 and 5). The setting was the same for all studies, namely at the campus of the VU University Amsterdam, making it relatively easy to recruit (VU-university) student-participants, but it was more difficult to recruit other participants. However, as the Active2Gether study was aimed at testing efficacy and not primarily at testing more general effectiveness, the studies described in this dissertation did not specifically target lower educated young adults. It is clear that the findings of the current work should not be generalized to all young adults.

In all studies that were included in this dissertation a relative large majority of the participants were female. Therefore, the results are more relevant for female young adults than for males. We were not
able to test this more explicitly because of the number of participants of which in particular the male participants was too low to allow explorations of interaction by gender and/or stratified analyses. This over-representation of women is in line with other studies on health related mHealth interventions. A recent systematic review reported that in 14 out of 19 studies the majority of the participants was female.\textsuperscript{157} We expected that young men would be interested in participating in research that made use of physical activity trackers and apps, as they are expected to be more interested in technology and IT, but this was supported by the participation rates of the studies presented in this dissertation. The approached men frequently reported that they already were sufficiently active or that they were not interested in this type of research, and the generally higher interest of women in health-and-weight-related issues appears to have been more relevant. Future research should more actively involve men and develop recruitment strategies that appeal to them.

**Challenges in study designs**

Interventions should be developed by means of a stepwise and systematic – such as proposed by Intervention Mapping\textsuperscript{6} – approach to increase the likelihood of effect; such an approach was followed to develop the Active2Gether intervention, including a number of studies to inform the development of Active2Gether using different study designs and research methods. Below we describe the challenges we encountered while applying various research designs and provide some recommendations for future studies.

**Challenges in conducting a systematic review on publicly available apps**

As the topic of the current work was relatively new and innovative in the field of physical activity promotion we first conducted systematic reviews (described in Chapter 2) focusing on physical activity apps to explore the landscape of the state-of-the-art of physical activity promotion apps. There was no standard protocol on how to search for apps in app stores and how to review their content and features. As we were the first to rigorously screen physical activity apps, we aimed to follow the standard review process as far as possible. However, the first challenge was already in identifying and screening relevant apps in app stores, where other search terms were needed than are used for reviews of electronic databases of scientific literature such as PubMed. Furthermore, using relevant search terms in App Store and Google play revealed for some search terms over 10,000 apps per term making it practically impossible to screen all possible apps. Consequently, we needed to adjust the standard screenings process to make the review feasible. Making such concessions in the screening process, may have led to an exclusion of potentially relevant apps. In addition, scoring apps was relatively new and there was no protocol on how to score apps. The previous published reviews scored the apps based on their descriptions\textsuperscript{36,60} whereas we downloaded the apps and explored the apps for behavior change strategies embedded in content or algorithms. However, we might have missed
features that only become apparent after a longer use of the app than we were able to do for the review. Also, despite the well-described scoring protocol, the results are evidently to a certain degree dependent on the reviewers’ interpretation.

There is a need for a more standardized protocol to screen apps for evidence and behavior-change theory-based quality. As there is no existing protocol or taxonomy, researchers develop their own protocols and subsequently the results are more difficult to compare. Such a protocol should include information on the procedure to score apps (i.e. score apps based on their description or based on the downloaded app, how to use an app before scoring it) and a clear definition of a BCT that can be applied in an app. Additionally, recently more attention has been given to engagement elements and it can be valuable to add items to the protocol addressing engagement elements, such as gamification and design elements.

**Challenges in assessing the target population’s preferences**

We used focus group discussions to gain in-depth knowledge from young adults on the usage of apps and feature preferences (described in Chapter 3). Since this was a topic not yet often studied, we regarded focus group discussions as a suitable research approach to explore the students’ opinions, beliefs and experiences regarding physical activity apps 236, accepting that the results of such a focus group study are not readily generalizable to a population at large. Additionally, and subsequently, we investigated the same topic quantitatively using a cross-sectional survey. An additional aim of this survey study was to explore whether personality characteristics and ratings of BCTs were associated and thus whether we should further tailor the BCTs to personality characteristics. For this research aim, conducting a cross-sectional survey was – we believed – the most suitable study design, given the timeframe and as a next step after the focus group investigation. However, the results of the survey provide insights in associations and is not valid to establish causal relations. Furthermore, this survey focused on the participants’ ratings based on their previous experiences, but it did not focus on how participants used certain features that include BCTs. Thus, future studies should examine how and when participants use certain app features and how that is associated with health behavior, such as physical activity.

**Challenges in validating a commercially available activity tracker**

We conducted a validation study (described in Chapter 4) to validate the Fitbit One against the ActiGraph GT3X+ in daily life. Participants wore an ActiGraph GT3X+ and a Fitbit on an elastic belt on the right hip for 7 consecutive days. The aim was to assess the relative validity of the Fitbit One for steps, and levels of moderate, vigorous and moderate-vigorous physical activity. As little is known about the validity of the Fitbit One in validly monitoring physical activity using shorter time intervals relevant for real-time feedback and instant behavioral insights to its users, we aimed to validate the
Fitbit using shorter time intervals, i.e. minute, hour and day. However, while conducting the validation study we encountered several challenges. Firstly, the Fitbit assesses step activity and number of stair climbed, however no accelerometer that is frequently used for research purposes assesses the number of stairs ascended. Thus, we were not able to assess the relative validity of the Fitbit for assessing the number of steps climbed. Secondly, the Fitbit allows users to choose their wear position of the Fitbit, for example on a belt, in the front pocket or on a bra. However, for practicality reasons we asked and instructed all participants to wear the Fitbit on the right hip. It remains unclear how valid the Fitbit is with respect to assessing physical activity while wearing the Fitbit in the front pocket or on a bra. Thirdly, Fitbit does not provide raw data and it remains unclear how the Fitbit algorithms define activity levels as moderate, vigorous or moderate-vigorous intensity. Lastly, the Fitbit was compared to the ActiGraph, i.e. another accelerometer. The gold standard to assess total energy expenditure in free-living conditions is the doubly labeled water (DWL) technique, but for practical reasons (i.e. costs, availability) we used the ActiGraph as a reference measurement instead of DWL.

We were not the first study examining the validity of the Fitbit, and various studies have been published validating commercially available activity trackers, including the Fitbit. Despite the fact that most studies had in general the same research aim, i.e. assess the validity of one or more commercially available activity trackers, and had in general the same outcome variables (step count and/or energy expenditure) various research protocols were applied, the majority of the earlier studies were conducted in a controlled lab-setting using different golden standard measures, i.e. room calorimeter, (video) observations and a mobile metabolic analyzer. Although the choice for controlled lab-settings is essential for first insights in accuracy and validity of new devices, these studies generally fail to provide an environment that resembles real-life situations. In the case of commercially available activity trackers it is essential to know the accuracy and validity in real-life situations. Studies validating the Fitbit in daily life differed however with respect to the duration of the study, the reference device, data handling and data analyses. Consequently, it is difficult to compare the results of these studies and to compare the validity of the Fitbit with other commercial devices that were validated using different study protocols. Thus, for such studies validating commercially available activity trackers in daily life, better standardization regarding the number of days of data collection, the preferred reference measurement, and data handling and analyses, is warranted.

**Challenges in developing an app-based intervention to promote physical activity**

In the current work, we combined knowledge and expertise from different research fields. We aimed to surpass existing physical activity apps and interventions by combining theory-based coaching strategies, with model-based prediction to provide a truly personal intervention to promote physical...
activity (described in Chapter 4). Throughout the process of developing an app-based intervention we encountered several challenges.

Firstly, the development of the app-based intervention took more time than anticipated. Furthermore, information and artificial intelligence technology and their possibilities are constantly and rapidly changing and evolving. Thus, possibilities and preferences that were assessed at the beginning of the development (described in Chapter 3) might be outdated and/or no longer preferred today. Because of such rapid changes, protocols that allow shorter cycles of developing and testing are needed. For example, in software development agile methods are used to develop software programs. Using an iterative process with periods varying from a day to a month, smaller products are built in groups. Thus, instead of building the intervention at once and test it at the end, an iterative process may better allow researchers to build smaller parts of the intervention and to test it and then progress from there towards a more comprehensive intervention. This approach allows researchers to tackle small technical problems at an early stage in the development process. Furthermore, Van Gemert-Pijnen et al. suggest that the development is a continuous and participatory process and that stakeholders should be involved in the development process.

Secondly, the main focus was to develop the intervention content based on health behavior theory and scientific evidence, informing a health behavior change predictive model for provision of highly tailored feedback. Consequently, less attention was paid to app design and aesthetics. As app design and aesthetics may have a strong influence on user engagement, future studies should focus more strongly on a collaboration with app-designers and app-programmers who have experience in designing and building engaging apps.

**Challenges in evaluating an app-based intervention to promote physical activity**

When the intervention has been developed it is necessary to evaluate its efficacy or effects. The randomized controlled trial (RCT) is regarded as the gold standard for such evaluation, since it allows analysis of the causal relation between intervention and effects. However, evaluating the efficacy of an app-based intervention using RCTs also comes with challenges.

The first challenge was choosing what an appropriate control group would be; we choose to compare our intervention with two other interventions instead of a no-intervention or wait list control group. We argued that we wanted to develop an intervention that should do more than already available or less advanced interventions. In the trial described in Chapter 5, we compared the newly developed Active2Gether intervention therefore with more basic version of Active2Gether and to the Fitbit app. The comparison between the two Active2Gether versions provided information regarding the additional effects of the coaching part of Active2Gether; the comparison with the Fitbit control arm
provided information about the effects as compared to an existing ‘usual care’ physical activity monitoring and promotion device.

As already extensively discussed in Chapter 5, participants were not assigned to the three conditions based on true randomization. One reason for this was that the two versions of the Active2Gether app (Active2Gether-Full and Active2Gether-Light) were not available for iPhone users, so iPhone users were automatically assigned to the Fitbit condition. Also, the proportion of Android users who registered for the study was lower than expected, so – to maintain a balance between the three conditions – they were randomized over the two Active2Gether conditions, rather than over all three conditions. This is an obvious and important violation of the RCT-design and thus the design became in fact a quasi-experimental trial. Since iPhone users may systematically differ from Android phone users in study-relevant characteristics, for the study described in Chapter 5 the interventions should also have been ready for use on iPhones or the study should have been conducted among Android users only. However, recruitment of participants proved more difficult than anticipated, and to achieve a sufficient sample size, it was decided to include iPhone users as well. Even though we decided to include iPhone users, the number of participants willing to participate was still lower than envisioned. Despite our efforts for participant recruitment, fewer people than expected were willing to participate due to lack of interests, lack of time and the perceived burden for the participants. Due to the small sample, the statistical power of the results is too low to reject the null hypothesis.

The second challenge in evaluating an app-based intervention was that the trial could not be truly controlled, i.e. study participants may have used other apps or intervention activities to monitor and help them increase or maintain physical activity levels next to the Active2Gether app or Fitbit app. Furthermore, participants may not have used the app as frequently as expected and necessary to induce behavior change. The aim of the trial was to test the intervention effects for the three intervention groups and to compare them in a real-life setting. For example, all participants received a Fitbit that needed to be synchronized through the Fitbit website or the Fitbit app, some participants that were assigned to one of the two Active2Gether conditions used the Fitbit app as well. To assess, whether users used other apps including the Fitbit app, questions about app usage were included in the 12-weeks follow-up questionnaire. However, we relied on the self-reported answers of the users, who might have provided socially desired answers that may have biased the results. Furthermore, from the user experience evaluation we know that the Active2Gether app was not used as intensively as meant. Consequently, the trial was not truly controlled and it might be that the interventions effects in these two groups were influenced using the Fitbit app and/or low user engagement.

Thirdly, app-based interventions offer the possibility to deliver just-in time interventions that are relevant for the user’s situation for that specific moment. However, our trial examined whether the
total intervention was effective over the intervention period, but it did not provide information about possible effectiveness of specific real-time feedback and advice moments.\textsuperscript{179} Ecological momentary assessment in such a trial setting may help to assess potential specific effects throughout the intervention period.\textsuperscript{180} In the Active2Gether intervention group, participants received tailored coaching messages using adaptive modeling, whereas participants in the Active2Gether- Light group did not receive any messages. Comparing the Full group with the Light group therefore provided information about whether sending tailored coaching messages on top of the monitoring and social-comparison features had an additive effect on physical activity. However, it then remains unclear whether a potential additive effect was due to just receiving messages or to the fact that those messages were tailored to the characteristics of the user.

App-based interventions offer the possibility to provide adaptive and just-in-time interventions to promote health behaviors. As already mentioned, such interventions are relatively new in health promotion and a recent review concluded that the majority of studies on interventions targeting health behaviors in adults reported significant improvements in health behaviors.\textsuperscript{157} These results indicate that app-based interventions can be used to promote health behaviors such as physical activity. However, little is known about the longer-term effects and again these results only examine the overall intervention effects and it remains unclear what parts of the intervention are effective and what works for whom. To improve the efficacy of app-based interventions we need studies with longer follow-up periods and that examine what parts of the interventions works and for whom. Study designs with the participants as their own controls offer possibilities to examine the intervention effects of the app features and allow researchers to adapt the intervention content while examining the effects. Potential study designs that could answer these types of questions are the Sequential Multiple Assignment Randomized Trial (SMART)\textsuperscript{247}, Single Case Designs\textsuperscript{248, 249}, or micro-randomized controlled trials.\textsuperscript{179}

Lastly, we aimed to evaluate how users of the Active2Gether app – both Full and the Light version - used the app and evaluated the app. To do so, we used survey data and some system use data. However, app-based interventions offer the possibility to capture how the intervention is used by each user and can offer insights in usage patterns, exposure to the intervention and adherence.\textsuperscript{245} Information on such engagement measures are important, as e- and mHealth interventions report low levels of adherence and high levels of attrition. In addition, a consumer survey showed that 26 percent of health apps is only used once after downloading, and 74 percent of health app users indicated to have stopped using the app within ten times of using it.\textsuperscript{230} In the Active2Gether intervention, over 92 percent of the coaching messages were received successfully by the users and the majority of the Active2Gether users – both Full and Light – were still synchronizing their Fitbit after 12 weeks (see \textbf{Chapter 5 - Figure 5.1}), indicating adherence to the intervention. Thus, to evaluate the user
engagement in app-based interventions and to gain insights in reasons for the high rates of attrition, future studies should include measures of engagement, such as qualitative approaches (e.g. semi-structured interviews\textsuperscript{245, 250}, think aloud\textsuperscript{251}, focus groups\textsuperscript{245, 252}), questionnaires (e.g. user engagement scale\textsuperscript{245, 253}, digital behavior change intervention engagement scale\textsuperscript{245, 254}), system use data (frequency, intensity, time, type of engagement\textsuperscript{245, 255}) in their evaluation studies.

**Challenges regarding measurement methodologies**

A variety of measurement instruments were used and combined in the studies presented in this dissertation. To interpret the outcomes of the measurement instruments it is important to know whether the instrument is adequate for its purpose and how it compares with similar measures.

**Challenges in assessing physical activity**

The target behavior in this dissertation was physical activity and that can be assessed using a variety of measurement techniques. In the Chapters 4 & 5 we used objective measurements, i.e. the ActiGraph GT3X+ and the Fitbit One. Both are tri-axial accelerometers that convert accelerations into step counts. Questionnaire were used to assess total physical activity and types of physical activity in Chapter 3 & 4.

As previously discussed, accelerometers have some limitations (i.e. limited in their ability to capture upper body movement, cycling and water activities) and might underestimate total daily activity and energy expenditure.\textsuperscript{198} Using accelerometers was done for pragmatic and financial reasons: indirect calorimetry – the golden standard – is a costly instrument and cannot quantify levels of physical activity but only total daily energy expenditure.\textsuperscript{256} A recent systematic review reported that the majority of the studies used ActiGraph accelerometers to assess intervention effects on physical activity in daily life.\textsuperscript{257} Thus the findings of the current work – especially those described in Chapter 5 – can be compared with other studies aiming to increase levels of physical activity, as all studies have the same limitations with respect to the use of accelerometers.

In the Active2Gether intervention (described in Chapter 4) participants received a Fitbit One to (self)-monitor their behavior, especially the number of daily steps taken. The validity and the limitations of the Fitbit One have extensively been discussed in Chapter 4 and the main issues were also discussed earlier in this general discussion.

The Active2Gether intervention (described in Chapter 4) aimed to increase total moderate-vigorous physical activity by targeting sport participation, active transport and stair usage. To do so, we used objective (i.e. Fitbit One) and subjective (i.e. questionnaires) measurements. The Fitbit One assesses step activity – representing overall physical activity used for self-monitoring purposes – and numbers
of stairs ascended. However, little is known about the validity of the Fitbit One and its ability to detect changes in stairs ascended.

The Fitbit One it is not able to assess minutes of sports participation and active transport. To assess weekly minutes of sports participation and active transport, participants answered questions regarding the total minutes of sports participation on the day before questionnaire administration as well as minutes of active transport. Such self-reports are likely to be biased due to over-reporting due to social desirability bias or recall bias.258

In Chapter 3, physical activity was measured using the Short Questionnaire to Assess Health-enhancing Physical Activity (SQUASH). Questionnaires are less accurate than objective measurements, but are easier to use in certain circumstances against relatively low costs. Two validation studies have been conducted earlier among Dutch adults examining the reproducibility and validity of the SQUASH that was used.118, 259 Spearman’s correlations of 0.58, 0.54 and 0.92 for the total activity score, moderate and vigorous activity respectively were reported for the reproducibility analyses118, these are considered to be ‘moderate’ and ‘very strong’.260 Whereas ‘moderate’ and ‘poor’ Intra Class Coefficients (ICCs, 0.64 and 0.38) for moderate-vigorous physical activity in Dutch women and men.259, 261 Furthermore ‘poor’ Cohen’s kappa scores of 0.29 and 0.04 were reported in Dutch women and men respectively, for meeting the physical activity guidelines.259, 262 Similar results were reported examining the construct validity. The Spearman’s correlation as a measure for the relative validity was 0.45 for the total activity score for men and women,118, and Spearman’s correlation of 0.41 and 0.17 were found in Dutch woman and men, respectively.259 The results of Nicolaou et al259 indicates that the SQASH might be more suitable for Dutch women than for men.

Challenges in assessing location data

The user’s location data (GPS coordinates) was collected with a GPS-app integrated in the Active2Gether app (described in Chapter 4). The location data was used to determine whether the user visited his/her significant locations (e.g. home, study/work place, sports club) to enable personalized location-specific coaching messages to the user. In addition, the participants’ significant places (e.g. home address, parental home, sports location, university) were assessed at baseline using a questionnaire specifically developed for this purpose. The questions focused on travel options from home to the significant locations, thus information about the active and non-active transportation options. This information was used to create personalized messages that are relevant to the user. There is room for improvement within the app, as the location data was not yet used to assess minutes of active transport. Furthermore, some users did not want to turn on GPS on their phone because it shortens their battery life or because of privacy reasons.
**Challenges in assessing psychological determinants**

Validated questionnaires were used to assess behavioral determinants, i.e. self-efficacy, outcome expectations, social norm, perceived barriers, and intentions (described in Chapter 3 & 4). These questions (described in Chapter 4) assessed the determinants with respect to moderate-to-vigorous physical activity in general and thus not with respect to the specific coaching domains, i.e. sports participation, active transport and stair usage. The single item questions (described in Chapter 4) assessing each of the behavioral determinants for the three specific coaching domains were specifically developed for the present studies and were not validated beforehand.

Assessing each of the behavioral determinants every week allowed us to examine whether the coaching messages were effective in changing the specific behavioral determinants and to examine associations with specific physical activity behaviors such as sports participation, active transport and stairs usage.

However, assessing the same determinants every week by asking the same questions, may have biased the results.

**Challenges in assessing the usability of an intervention**

A questionnaire was designed for the purpose of evaluating the Active2Gether intervention with respect to perceived effectiveness, user friendliness, user appreciation of the coaching messages (Chapter 5). When possible, questions were copied from other studies evaluating the usability. When necessary, questions were developed by the researchers. As the questionnaire was not validated, the results should be interpreted with caution. For future studies evaluating mHealth interventions it would be desirable to create a validated questionnaire to standardize user evaluations to compare results of different interventions. Eysenbach et al. provided a template to report such evaluations of web-based and mobile health interventions, but it mainly focuses on the description of the development and how it has been used and less on how to evaluate the usability of the intervention.

According to the law of attrition a substantial proportion of users drops out before completion or stops using the app. For that reason, it is important to obtain objective information on the usage of the app, e.g. number of messages sent, received and read, and number of log-ins. Furthermore, levels of engagement are associated with intervention effects. For that reason, it is important to assess levels of engagement. However, as there are various definitions and operationalization of engagement it is important to work towards more standardization.
Challenges regarding modern technologies in future research

Using the Fitbit One enables users as well as researchers to continuously assess daily activity on a minute-by-minute interval resulting in large datasets per individual. Such data enables identification of behavioral patterns and to link these patterns with other individual, social and contextual factors. However, a limitation was that the users needed to synchronize the Fitbit regularly through the Fitbit app or Fitbit synchronization tool to obtain the necessary data. For the Active2Gether users this might have been an additional burden that could be linked with attrition and undesired use of the Fitbit. For example, approximately 42 percent of the participants reported that synchronizing the Fitbit was bothersome/annoying, 41 percent reported that they experienced a shorter battery life of their phone due to the Fitbit app, and approximately 46 percent experienced some technical problems with the Fitbit app.

To limit the burden for the participants, future research should make use of activity trackers in smartphone to assess the participant’s behavior. Nowadays, smartphones automatically collect activity data and are able to collect other types of health-related data, such as heart-rate and blood pressure. Making use of the smartphones sensors allows researchers to collect and to combine increasingly sophisticated data that can be used to provide highly tailored and real-time advice. However, doing such kind of data collection has some disadvantages as well. Firstly, assessing physical activity with smartphones sometimes requires that the participants carry their smartphone with them all day. Additionally, to interpret and compare such phone-based data between and within participants, all participants need to carry their smartphones consistently in the same spot. Secondly, such using multiple apps or sensors also limits the battery life of the phone. Lastly, collecting all sorts of personal data from smartphones over a longer period of time for intervention or research purposes obviously poses very serious ethical challenges (i.e. privacy, data security and data ownership). Additionally, although the amount of data that can be collected using new technologies as smartphones and activity trackers is enormous, several studies have reported that the validity of some smartphone sensors and activity trackers are disappointing.

Challenges in statistical analyses

Challenges in analyzing survey data

In Chapter 3 & 5 we analyzed survey data, various aspects (e.g. behavioral determinants, user appreciation and experience) were assessed with questionnaires. For analysis of the data, sum scores were calculated. In Chapter 3 & 5 exploratory factor analysis were conducted to identify categories. However, summing the scores of the items within one category is based on some assumptions: a) all items are equality weighted and contribute equality to a construct, b) changes in a sum score is
dependent on the number of items, a sum score with larger number of items measuring the same construct will probably change a lot compared to a construct with sum score with only two items measuring the same construct.  

**Challenges for validation studies**

When using a new measurement instrument (e.g. Fitbit One) for research purposes, it is important to know its reliability (‘the degree to which the measurement is free from measurement error’ \(^{265}\)) and validity (‘the degree to which an instrument truly measures the construct(s) it purports to measure’ \(^{265}\)). In **Chapter 4** we aimed to examine the construct validity of the Fitbit One, as a gold standard was lacking. The construct validity is often considered to be less powerful than the criterion validity \(^{266}\), but as the ActiGraph is the next best measurement we assume that that analysis provide information on the validity of the Fitbit.

**Agreement analyses**

An aspect of testing the quality of a measurement instrument is the agreement that represents the lack of measurement error. Traditionally, Bland-Altman analyses are used to calculate differences between two measurements \(^{267}\), such as the difference between the ActiGraph and Fitbit measurements. However, in **Chapter 4** we used linear mixed model analysis with a four-level structure to examine the agreement between the Fitbit One and the ActiGraph. Additionally, Bland-Altman plots including the limits of agreement (LoA) were used to provide a visual representation of the systematic differences and to assess potential non-systematic differences between the ActiGraph and Fitbit.

Applying Bland-Altman analysis to validate activity trackers as the Fitbit has some challenges. Firstly, the analysis were originally developed to assess the agreement of two measurements done on one occasion, thus for independent data. \(^{268}\) Therefore, the traditional Bland-Altman analysis could not be applied for the repeated measures used in **Chapter 4**. \(^{268}, 269\) An alternative would be to calculate the average amount of steps taken per participant and to calculate the limits of agreement as used for independent data. \(^{268}\) It should be noted that the limits of agreement would be too narrow \(^{268}\), as can be seen in **Figure 6.1**. Another alternative would be to adjust for the non-independence of the observations (i.e. method where the through value varies or where the true value is constant), as proposed by Bland-Altman. \(^{268}\) In **Chapter 4**, we adjusted for the non-independence, assuming that the true value varies. \(^{268}\) In **Figure 6.1**, we plotted the physical activity data for steps per hour showing the means and limits of agreement calculated with the three methods; ignoring the dependency of the data, calculating the average steps per individual and applying the method where the true value varies. As can be seen, information gets lost when using mean values to examine the systematic differences,
but similar means and limits of agreement were seen for the adjusted and unadjusted means and limits of agreement.

Secondly, when the data is not normally distributed, the Bland-Altman analyses should be adjusted.\textsuperscript{270} This could be a challenge, as physical activity data is often skewed. In \textbf{Chapter 4}, all analyses were checked for normality of the data. However, previous studies used Bland-Altman plots for skewed data\textsuperscript{193} or did not report whether they checked for the assumptions that the analysis is based.\textsuperscript{187, 243}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{bland_altman_plot.png}
\caption{Bland-Altman plot for steps per hour for measurements with the ActiGraph and Fitbit with adjusted and non-adjusted means and limits of agreement and means per individual.}
\end{figure}

\textit{Note}. This figure shows a selection of the Bland-Altman plot shown in \textbf{Chapter 4} and zooms in in the data for showing the differences in limits of agreement between the adjusted and non-adjusted limits of agreement. The means and limits of agreement were calculated using the original dataset.

\textit{Note}. The x-axis shows the average number of steps and the y-axis shows the differences in steps between the two methods. The \textit{adjusted} mean and limits of agreements are plotted in \textit{orange} and the \textit{non-adjusted} mean and limits of agreements are plotted in \textit{blue}. The mean of all steps per day, hour and minute \textit{per individual} are plotted in \textit{red}.

\textbf{Association and Correlation analyses}

Traditionally, physical activity measurement have been validated with respect to their ability to assess weekly physical activity using correlation techniques (i.e. Spearman’s \( r \) or Pearson’s \( r \), intra class
correlation coefficient (ICC)\textsuperscript{266, 271}. Correlation parameters such as Spearman’s \( r \) or Pearson’s \( r \) show the degree to what two measurements are correlated.\textsuperscript{261, 272} For continuous variables the Pearson’s \( r \) can be used to examine whether two measurement instruments are correlated.\textsuperscript{266, 272} The Pearson’s \( r \) is a linearity index that provides information on the strength of a linear relationship, but it does not take systematic or fixed errors into account.\textsuperscript{260, 272} For that reason, the Pearson’s \( r \) can be used when there are only random errors, otherwise it will not provide a good indication of the reliability and validity. Additionally, Pearson’s \( r \) can be used for cross-sectional data, but it is an inappropriate technique for longitudinal data as it assumes the independence of the errors between observations.\textsuperscript{273} Therefore, examining the correlation between two measurements with continuous variables, researcher should apply other statistical techniques such as repeated measures correlation.\textsuperscript{273}

A more desirable parameter is the Intra Class correlation coefficient (ICC)\textsuperscript{260}, as it takes the correlation and the agreement between two measurements into account.\textsuperscript{261} Furthermore, ICC can be used for longitudinal data as well as there are several ICC formulas. It is important to select the correct formula and to report the type of ICC that has been calculated as different types of ICCs can lead to different estimates even when the same dataset is used.\textsuperscript{261} Furthermore, it is important to report a confidence interval for the ICC to provide information about the uncertainty of the estimate.\textsuperscript{261} Unfortunately, some studies examining the validity of the Fitbit and reporting the ICC often do not report the type of ICC nor the confidence interval.\textsuperscript{193} This makes it difficult to compare the results of different validation studies. Therefore, it is important to standardize the reporting of the results, i.e. by providing information on the type of ICC and providing a confidence interval.

Furthermore, ICC for time spent in moderate, vigorous and moderate-vigorous physical activity could not be calculated as this concerned dichotomous data. However it would have been possible to report Cohen’s kappa or to report the sensitivity and specificity.\textsuperscript{260} Thus to conclude, when validating modern activity trackers like the Fitbit One researchers should consider using smaller time intervals as days, hour or even minutes. Furthermore, researchers need to choose the appropriate statistical techniques, as the traditional techniques cannot be used for longitudinal data.

\textit{Challenges for evaluating mHealth intervention studies}

In \textit{Chapter 5} the intervention effects were evaluated using linear regression analyses. We examined whether weekly levels of physical activity and the behavioral determinants after twelve weeks differed between the intervention groups. Analyzing the intervention effects in that way provides information on the effects at the end of the intervention. Due to the small sample size in the two Active2Gether groups statistical power was an issue and therefore the results of the study should be interpreted with caution.
Chapter 6

The effectiveness of the intervention components such as the coaching messages were not examined due to the small sample size and attrition rates. With a larger sample size, it would have been possible to explore the mechanisms of underlying relationships between psychological determinants and physical activity behaviors by using mediation analyses or structural equation modeling, as we frequently assessed the psychological determinants and levels of physical activity.

**Challenges for analyzing studies using modern technologies**

Nowadays, physical activity data with smaller epochs allows us to identify patterns over time. Furthermore, new technologies as smartphones and wearable sensors enable researchers to frequently assess other types of data such as psychological determinants. By frequently assessing psychological determinants in combination with activity data, it is possible to explore the mechanisms of underlying relationships between psychological determinants and physical activity behaviors. Advanced statistical techniques can then estimate between and within participants associations between changes in determinants and changes in behavior. This provides insight in examining what works and for whom.

The rapid development and expansion of smartphones and wearable wireless devices (e.g. activity trackers, smart watches) enables assessing large volumes of data per individual. Especially when combining physical activity data with other types of data – location data, contextual data, heart rate, and psychological data – it gives the possibility to assess what works in certain situations or circumstances. However, analyzing such large volumes of data requires a critical selection of big data analytic methodologies and software used in analysis (e.g. predictive analysis, machine learning, data mining, network analyses).

**IMPLICATIONS FOR FUTURE MHEALTH RESEARCH**

Evidence and theory-based development of an ‘artificial intelligent’ app-based physical activity promotion intervention proved to be a complicated, interdisciplinary, and iterative process and the development took more time than anticipated. Furthermore, information and artificial intelligence technology and their possibilities are constantly and rapidly changing and evolving. Consequently, features that were not possible to incorporate when the endeavor began, may have been possible to include in interventions today. The implications for intervention development aiming to systematically develop an app-based intervention based on relevant theories and evidence should consider applying other approaches than what is considered state-of-the-art in health promotion intervention development such as proposed in intervention mapping, i.e. approaches with shorter cycles of developing and testing (e.g. control theory, use of factorial or fractionated evaluation designs).
while involving the target population and stakeholders as proposed by the holistic framework to improve the uptake and impact of eHealth technologies.244

Although receiving dynamically tailored feedback is an evidence-based approach that may even be considered as best-practice in health education, the evaluation of the Active2Gether interventions did not provide evidence in favor of this tailored intervention. There is an ongoing debate and different efforts are being undertaken to learn more specifically how interventions need to be tailored, e.g. to which factors (e.g. behavioral determinants, personal characteristics), using which BCTs where and when, and what aspects of design should be individually tailored.

Regarding the latter, in terms of design and layout, as well as certain aspects of functionality, apps designed by university parties and primarily and initially designed for research purposes to test efficacy of such apps and its characteristics, cannot fully compete with publicly available apps. The implication for health research is – and this has been argued before – to collaborate closely with professional app designers and companies and organizations that (aim to) spread and/or sell such apps, to minimize the probability of technical errors and to increase levels of engagement due to the aesthetics and design of the app. However, such collaborations comes with many challenges, such as regarding the actual purpose of such apps (e.g. health promotion or commercial value), (intellectual) property rights, open science vs commercial competition, privacy, potential conflicts of interest, and research ethics and integrity.

**GENERAL CONCLUSIONS**

The results of this dissertation indicate that young adults are interested in using apps to help them to increase their levels of physical activity. Such physical activity apps should be based on behavior change theory and should implement technology to provide higher and deeper levels of tailored feedback and coaching. Preferably, physical activity apps should include features that rank users’ performance and accomplishments as compared to peers, as well as goal setting, coaching and tailored feedback features. Opposite to expectations, social support features as implemented in current apps were not preferred by respondents. In addition, the results of the present studies revealed that users of the Active2Gether-Full app appreciated a coaching feature, but are critical about the implementation. No evidence for significant superior intervention effects of Active2Gether on increased physical activity or more positive determinants of physical activity as compared to a ‘lighter’ version of the same app or to a Fitbit-only intervention were found.