Eindhoven University of Technology

MASTER

# Living a Healthy Life

## Evaluating the Effect of a Rasch-based Recommender System on Healthy Lifestyle Choices with a Gamified mHealth Tool

Arreman, Dennis

*Award date:*
2022

Link to publication

**Disclaimer**
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

**Take down policy**
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 27. Mar. 2025

Eindhoven,
July 8th, 2022

# Living a Healthy Life: Evaluating the Effect of a Rasch-based Recommender System on Healthy Lifestyle Choices with a Gamified mHealth Tool.

by Dennis Arreman

identity number 0935169

in partial fulfilment of the requirements for the degree of

**Master of Science**
**in Human Technology Interaction**

Supervisors:
Dr. Ir. M.C. Willemsen
Dr. P.M.E.V. van Gorp
Dr. Ir. K.C.H.J. Smolders

**Abstract**

As sedentary- and other unhealthy behaviors are becoming a significant problem in modern-day society, we need to look for solutions that are able to effectively motivate us in improving our lifestyle. Theory shows that in order to push us to our limits, goals or tasks should be at the verge of our limits, while still feeling achievable. However, as no individual is the same, a "one-solution-fits-all" does not exist. We therefore need to find methods of personalizing these goals, catering them to the individual. Recommender systems are already used in this exact context, where a scale based on Rasch modeling has successfully been applied in order to give appropriate recommendations. The current paper investigates the feasibility of creating a uni-dimensional Rasch scale based on healthy lifestyle activities and using this scale in a recommender system. Our initial analyses showed that we were able to successfully fit a Rasch scale based on historical data. We continued with the recalibration of this scale based on survey data (N=38). This recalibrated scale was used in establishing lifestyle recommendations. We evaluated two levels of relative difficulty in order to quantify how these recommendations should be tailored. Results indicate that creating a Rasch-scale in which a healthy lifestyle is the uni-dimensional construct is feasible. Results of the study further indicate that there is an negative effect of level of difficulty of the recommendations on satisfaction and motivation of users. This is in line with previous studies, but further studies are necessary to confirm this finding.

**Key words**: Healthy Lifestyle, Recommender Systems, mHealth, Rasch modeling, Psychometrics

**Acknowledgements**

During the writing of this Master' Thesis, the realization hit me that this will be the last thing I will produce as a Human-Technology Interaction student. Although I make this sound like a bad thing, I am positive and enthusiastic for my future, as my Master's program has brought me so much, both in theory and in practice. Nevertheless, with this preface, I would like to show my appreciation to the many people that helped me to get to this point in my Thesis.

Martijn, I want to thank you for the many, many meetings we have had during the last couple of months. Some more serious than others, but always productive. I realize we have had more meetings than we both would have wanted, but I hope that my enthusiasm and positivity made me bearable.

Pieter, although we did not have as much contact as we both would have wanted, thank you for your critical view during the times we did converse. From a scientific standpoint, the fact that your view is so different from mine is absolutely fantastic. From a personal standpoint, it was sometimes frustrating, but very important and helpful in the end.

Raoul, I would like to thank you for your accessibility and interest during my thesis. Without your help during the set-up of the experiment, this thesis could not have been possible. Also, your research is one of the main reasons I became even more enthusiastic about the current study.

Thank you to my dearest Dominique, which supported me throughout the whole process, and for sometimes giving me a kick in the ass when I needed it.

Finally, I want to thank you, dear reader, for showing interest in my Master' Thesis. I hope that, with reading this thesis, you experience the same enthusiasm I have had throughout the whole process.

## Introduction

Globally, 1 out of 4 adults do not meet the recommended levels of physical activity by the WHO. (World Health Organization, 2020) This is problematic, as sedentary behavior imposes a large number of health risks, where the WHO even states that insufficient physical activity can lead to a 20% to 30% higher death risk compared to a sufficiently active person. But not only physical activity has a positive influence on general health; dietary intake is also a large contributor to a healthy lifestyle. (World Health Organization, 2010) In order to get control of their lifestyle, society needs to look (and is looking) for solutions to give individuals control over their choices regarding physical activity, dietary intake as well as other factors contributing to a healthy general lifestyle.

In this search, the use of mHealth application can offer a solution. The range of information these applications can monitor can vary from physiological information such as heart rate and blood pressure to physical information such as kilometers walked or meters swam or detailed nutrient intake. These applications give their users control and insights into their personal lifestyle and can give recommendations based on the recorded information. Although mHealth apps show high potential in driving behavioral change, long-term effects are insufficiently studied or unclear. Mainly, the retention rate for these kind of applications is relatively low. For mHealth applications with more than 50000 active users, only 38% of these report more than 1000 monthly users (Medica Magazine, 2018), and the average retention rate for mobile applications across all industries is around twenty percent after 90 days (Brown, 2018).

Although retention rates are a known problem for mHealth applications, research suggests that personalizing the "healthy" recommendations that an mHealth app gives can be a significant factor in motivating users, both short- and long-term. (e.g. Adams et al. (2017) and Nuijten et al. (2022))

The current research sets out to create a novel approach to catering healthy lifestyle recommendations. In order to do this, the system used in producing recommendations is based on the relation between attitude and behavior as described by the Rasch model (Kaiser et al., 2010) Rasch-modeling finds its origin in the Item-Response Theory (IRT) (An & Yung, 2014) and is a model for analyzing mostly categorical data. It is able to directly compare the difficulty of a given measure, item or activity to the current behavior of an individual and in turn enables a tailored method of recommending activities to an individual based on their abilities or preferences.

Another important question is then raised; should the recommendations given to an individual be above one's ability or below? Flow Theory suggests that there needs to be a between the difficulty of an activity and the capability of an individual, but does not state what this balance should be (Csikszentmihalyi, 2009). Goal-Setting Theory and literature suggests that setting a specific goal is effective in provoking significant behavioral change, and that these goals should be "achievable" but does not say what this "achievability" represents (Locke et al., 2015; McEwan et al., 2016). Studies

investigating goal difficulty tailoring suggest that goals that are on the verge of one's capabilities are more effective (Adams et al., 2017; Chapman et al., 2016; Moon et al., 2016), while other studies show that people tend to view easier goals as more appropriate (Starke, 2014).

The current study sets out to investigate the feasibility and effectiveness of a Rasch-based recommender system on healthy lifestyle choices. These recommendations are not only physical activity-related, but also includes decisions regarding dietary intake or general lifestyle choices such as doing groceries by foot or bike. This translates into the following main research question, around which this thesis is based on;

*RQ: "What is the effect of a Rasch-based recommender system on healthy lifestyle choices?"*

## Background

### mHealth Applications and Recommender Systems

According to the World Health Organization, 28 percent of the global adult (i.e. over 18 years of age) population engages insufficiently in physical activity. This number is higher in high-income countries, where the average level of inactivity is twice as high compared to low-income countries. Although health benefits of an active lifestyle have been proven, the level of global inactivity has only been rising, with a rise of five percent between the years 2001 and 2016 (World Health Organization, 2020).

Individuals who are becoming aware of the dangers and negative outcomes of their increased sedentary lifestyle are in search of tools which can aid them pursuing a healthier lifestyle. In light of this, the use of mHealth applications is highly promising. mHealth applications regard applications which are capable of monitoring (and in some instances sharing with either their social contacts or their medical experts) health information via either self-reports or with the use of mobile technologies such as wearables or smart phone applications. These applications can not only be used for monitoring heath-related information, but can also use this information to motivate their users in changing their behavior (Milne-Ives et al., 2020).

An excellent example of a healthy lifestyle-oriented application is GameBus. GameBus is a project which aims at creating a novel application aimed at gamifying the experience of monitoring and improving various lifestyle-related aspects of one's life such as physical activity or dietary intake. This done by using the concept of "challenges"; a time-restricted period in which users are given a number of activities which they can perform to their own liking. Challenges are mostly created for a group of individuals, which range from a group of friends to an entire faculty of an university. Each performed activity gives the user a number of points, set by the organizer of the challenge, which can be anyone (for example, someone from the management of the faculty or a company doctor), while the aim of the challenge is to score the most amount of points. ("GameBus", 2016)

***"A platform that rewards teams for playing together healthy social, cognitive and physical activities in a personalized gaming experience"*** ("GameBus", 2016)

However, a current limitation of the setup of the GameBus application and its "challenges" is that the activities and the points rewarded are set by an individual and are equal for all participants. This results in unequal opportunities for every participants to achieve the same number of points. For example, someone who does not have a bike cannot receive any points for any bike-related activities. Moreover, one cannot assume that all participants are equally fit or physically healthy; someone may have difficulties running more than 1 kilometer, either because they are not fit enough or they have a medical condition. Therefore, to give good and appropriate recommendations for activities, recommendations should be personalized and be tailored

to the individual's personal capabilities. The current research draws upon recommender system literature to personalize these recommendations. The next section will discuss in detail what recommender systems are and how they achieve this personalization

## Recommender Systems

Recommender systems can be very useful when aiming to create a personalized experience. These systems use a user's information regarding their current preferences or attitudes (e.g. past purchases on a website (Rana & Jain, 2012) or songs listened to (Liang & Willemsen, 2019) but can also incorporate personality traits (Ferwerda & Schedl, 2016; R. Hu & Pu, 2009) to give appropriate advice. This advice can, for example, include recommendations on what song to listen to (Liang & Willemsen, 2019) or what movie to watch (Gomez-Uribe & Hunt, 2015). The next section will discuss some successful recommender system interventions, namely in the energy-saving and the nutrition intake domains and then continues with a short overview of the current literature in the healthy lifestyle domain.

### *Energy-Saving Behavior*

Starke et al. (2017) have developed a recommender system of 79 distinct energy-saving measures and have proven that it can effectively be used in order to give recommendations on which measures someone can take based on their current attitudes and preferences. It draws upon research by Knijnenburg et al. (2014), which proved that recommender systems could effectively be used in helping people discover new, effective energy-saving measures that accurately matched their profile. Furthermore, their research concludes that the balance between the energy-saving ability of the individuals and the required effort or behavioral cost of a measure should be taken into account when tailoring advice, stating that measures that match one's current preferences or attitude are most effective.

### *Dietary Intake*

In a study by Schäfer and Willemsen (2019), researchers created a recommender system that aids the user in striking a healthy balance between all nutrients. They were successful in creating a recommender system that used both the user's food preferences as well as current nutrient intake to give recommendations on recipes that matched the user's food preference while increasing nutrient intake for nutrients that were lacking in the user's eating habits. In their research, the tailored system was more successful in provoking positive behavioral change as well as slowing the decline in system interaction over the course of the two-week intervention study.

### *Healthy Lifestyle*

Most current literature on lifestyle recommender systems aim to use the combination of a physical activity- and a food recommender system to create a "healthy

lifestyle" system (e.g. "Runner" (Donciu et al., 2011) or "Shade" (Faiz et al., 2014)). However, most of these systems relate to very niche user groups (e.g. users with a medical condition (Faiz et al., 2014; Ferretto et al., 2020) or professionals (Donciu et al., 2011)). Dharia et al. (2016) have proposed a system that suggests personalized workout and dietary recommendations that uses contextual data of users, that include past activities as well as their current physical fitness and preferences. Recommendations are then given based on both the user's preferences, as well as the preferences of other, similar users. However, our literature research showed no recommender system that uses such a framework. In the current study, we aim to investigate the feasibility of a system that uses the contextual data as Dharia et al. (2016) suggest in a generalized (i.e. for every individual) setting.

## Goal Difficulty and Perceived Capability

The notion of the balance between the perceived challenge and an individual's capabilities is not unfamiliar when investigating the effectiveness of behavioral interventions. For example, we already discussed Starke et al. (2017), in which the researchers found that measures that matched the user's profile were most effective. Research widely suggests that the the extent to which a task is challenging is an essential factor when aiming to provoke behavioral change. However, research on the optimal balance between the capabilities of the individual and the difficulty of an activity yields varying results. (e.g. Bonenfant (1971) and Sporrel et al. (2021)) We therefore continue by describing two influential theories in the domain of behavioral change and discussing their influence and view on what this balance should be.

### *Goal-Setting Theory*

The theory of goal-setting is one the most frequently used methods found in physical activity interventions aimed at provoking behavioral change (Michie et al., 2018). Locke (1968) states that goal-setting theory (GST) is based on the idea that the process of setting goals is highly effective in reaching a user-set level of proficiency in a given activity. They explain that it focuses an individual's attention towards their goal and thus in turn increasing tenacity towards this goal.

Locke et al. (2015) states that the ratio between a set goal and the performance of an individual can be moderated by distinct 'goal attributes', such as goal clarity , feedback, task complexity, and challenge. Moreover, the study points out that several individual difference variables may significantly affect the performance on the goal. More specifically, they mention that a higher commitment to the goal, ability and self-efficacy, perceived importance and anticipated satisfaction maximise the effects of goal setting.

In a systematic review and meta-analysis by McEwan et al. (2016) on 45 studies, the researchers found a medium positive effect of goal setting on physical activity. Furthermore, the researchers found that effectiveness is present irrespective of goal communication method (e.g. in-person or via technology) and intervention duration. In

line with Locke et al. (2015), no significant changes in effectiveness were found when controlling for sample characteristics such as age, weight or prior activity status. All these findings are encouraging, as they suggest that an intervention using specific goal-setting could be very effective in provoking behavioral change.

### *Flow Theory*

Csikszentmihalyi's Flow Theory discusses the idea that an individual can be fully immersed when performing an activity, and creates feelings of immediate enjoyment, full involvement and a high level of focus (Csikszentmihalyi, 2009). An example of such an activity is performing one's hobby, such as drawing. Although the theory is relatively old, it began to become prevalent between 1980 and 1990. It can be applied to not only lifestyle interventions, but is also found in areas such as habit learning or occupational therapy. Current Flow Theory research is interested in how it can be used in order to emphasize positive experiences. In the end, Flow Theory can be applied when one aims to create an experience that is so enjoyable that it creates happiness and positive affect in the long run. Moreover, studies found that challenges and/or tasks that are above an individual's average capability nurture positive affect.

Csikszentmihalyi poses that, in order to achieve flow, three distinct conditions should be met; the activity must have clear goals and progress, feedback should be immediate and there is a requirement for a balance between the perceived challengingness and perceived capabilities of the individual. We find the former two in multiple goal-setting or behavioral change theories (e.g. in GST (Locke & Latham, 2002)). However, in the scope of the current study, the latter is especially interesting.

### Tailoring Goal Difficulty

Goal-Setting theory states that the challenge of a task should be at the verge of one's own capabilities, and performance on these tasks stabilizes or even decreases when these capabilities are exceeded or when tasks are too easy (Locke & Latham, 2002). This notion is also present in Flow Theory, as it requires a balance between the perceived challengingness of an activity and the perceived capability of the individual. Moreover, Flow Theory suggests that a person's perceived capability changes over time as one becomes more proficient in an activity (Csikszentmihalyi, 2009). This thus also suggests that, in order to prolong engagement and to keep performance on an activity high, the goal complexity and/or challengingness should be adapted to the individual's skills.

In the scoping review by Sporrel et al. (2021), the researchers explored the different design characteristics of persuasive strategies used in PA mobile health interventions. It further examined the extent to which previous mHealth intervention studies sufficiently examined the effects of different design characteristics in order to sufficiently persuade the user. In summary, monitoring, reminders, sharing and social comparison were insufficiently studied to draw conclusions about implementation. Furthermore, rewards are an effective method in order to promote PA but the effect of a given value or type (financial, points etc.) is unclear. Also, as rewards mainly promote

extrinsic motivation and long-term behavioral change is mainly correlated with intrinsic motivation, the effect of rewards on long lasting behavioral change is uncertain.

The results of the study further suggest that goal setting has a significant effect of the intervention on promoting PA. Among other results, the study revealed that system-set goals are more effective than user-set goals. Interestingly however is that most users chose to set a manual goal instead of letting the intervention set a goal for them, implying that users mostly prefer to set their own goals. Furthermore, in line with GST, the study also demonstrates the fact that when goals are catered to the physical capabilities and/or context of the individual, effectiveness is likely to increase in comparison to generic goals.

Results further discuss the fact that adaptively tailored goals are more effective than generic, static goals. This is mainly seen in a study by Adams et al. (2017), in which researchers concluded that adaptive goal setting resulted in a slower decline in PA throughout the study period, resulting in a higher improvement in PA at the end of the study period. Extending on this, results further suggest that goals that are on the verge of a user's capabilities appear to be more effective than goals that are set too easy or too hard. This is in line with Flow Theory, which suggested a balance between the perceived challengingness of a task and the perceived ability of the user. However, the exact challengingness of the task is unclear, as well as the difficulty increase when using dynamic goals. For example, Moon et al. (2016) found that an increase of 20% and 40% is most effective, while Chapman et al. (2016) found that a doubling of difficulty is most effective.

### Goal Personalization

Based on the previous literature, we have concluded that lifestyle interventions highly benefit from setting goals that are achievable and measurable. It also suggests that retention rates can possibly be improved by personalizing the experience and goal setting to the user (Sporrel et al., 2021). Therefore, in order for an intervention to effectively provoke behavioral change, the goals set should be personalized to the individual's situation. For example, someone who is more proficient in running a marathon will have a lower challenge in running-based challenges than a user who does not perform any physical activity outside daily activities.

Nuijten et al. (2022) studied the effect of personalizing goal setting in a gamified mHealth intervention. The aim of the research was to reinforce the notion that goals catered to a user's capabilities are more effective in provoking behavioral change.

In the eight-week study, participants participated in a health promotion campaign which was aimed at specifically promoting walking, bike rides and other sport sessions. Users could track their performance using a the GameBus mHealth tool, which also enabled users to compare their performance to other peers. The two randomized groups consisted of 1) the control group, which had their goals set by national guidelines and 2) the intervention group, in which goals were adaptively set based on self-perceived capabilities and self-selected goals.

Similar to the aforementioned study by Sporrel et al. (2021), engagement inevitably declined. However, similar to Sporrel et al. (2021), the rate of decline is lower for the intervention group, suggesting that a dynamic goal setting strategy is effective in prolonging engagement. This effect was especially observed when personalizing the frequency of activities, as results shows that personalizing the frequency of the sports sessions resulted in significantly higher engagement levels.

Although this paper is very insightful, as its work suggests that personalizing goals result in prolonged engagement and higher effectiveness, the personalization method used is difficult to translate to a generalized setting. Not every individual has a specific goal in mind, nor does everyone want to limit themselves to three distinct activities. Furthermore, the method used in adaptively making goals more difficult does not account for non-linear changes in an individual's capabilities, as it currently follows a linear increase. Finally, Sporrel et al. (2021) already showed that system-set goals are more effective than self-set goals, which are used in the current study. Therefore, we aim to improve on this research by improving on the personalization method used.

Summarizing, multiple studies have already shown that personalization of goals or recommendations can have a significant positive effect on positive behavior change. We also discussed the importance of the balance between the capabilities of a person and the difficulty of a task, where theory suggests that goals should be at, or just over, one's own capabilities to be most effective, whereas previous studies show higher effectiveness for tasks that are relatively easy (e.g. Starke et al. (2017). Moreover, it is unclear what attributes or personal characteristics play a role in motivating an individual in becoming more fit or performing a given activity.

In order to overcome these issues in the energy-saving domain, Starke et al. (2015) created a recommender system based on Rasch modeling, which encapsulates both persons and energy-saving measures on one scale. Similar accomplishments have been achieved for recommender systems for nutrition assistance systems (Schäfer & Willemsen, 2019) and lifestyle changes (Radha et al., 2016). Rasch modeling is driven by personal data and thus supplies the user with recommendations that are specifically catered to their personal capabilities and attitudes. This method of personalizing goals and recommendations comes with significant advantages in comparison to different methods or personalization. For example, the Rasch modeling does not assume that all activities are equally difficult for everyone. It provides a direct comparison between the difficulty of an activity to the ability of an individual and can therefore accurately determine whether it is probable that this individual will complete the activity. Furthermore, it eliminates the need for extensive surveys or tracking systems, as only a small amount of personal behavioral data is needed in order to be able to create a broad range of recommendations. For example, Starke et al. (2020) asked users for a set of 13 behaviors whether they performed these or not and was thereafter able to create an (accurate) recommender system from a set of 79 different behaviors. We therefore aim to create a similar system in order to personalize the lifestyle recommendations given by applications such as GameBus. The next section will discuss the psychometric Rasch

model and its design implications for a recommender system.

### The Rasch Model

Rasch modeling finds it origin in item-response theory (IRT) and is commonly used for analyzing categorical data. The premise of Rasch modeling is to compute the probability that a person performs a given behavior. This is done by quantifying a latent trait of the individual based on the assumption that a individual's response to any particular item is a function of (the difference between) both his/her capabilities and the difficulty of the item. For example, in Starke (2019), the researchers aimed at quantifying energy-saving behavior (the latent trait) by asking a set of energy-saving behaviors (such as turning the lights off when leaving a room) whether participants already performed these or not. Another example can be found for assessing mathematical performance of students, where this performance is evident when assessing their performance on mathematical assignments with varying difficulties (Bond & Fox, 2007; Galli et al., 2008).

Existing literature already proves that Rasch-based recommender systems can produce very promising results. The concept of a Rasch-model can be generalized as the probability than an individual performs a given item, given the difference between the ability of an individual and the difficulty of an item. This can be formalized as the probability $P$ that person $p$ performs a measure $m$ as the difference between the person $p$'s ability $\theta$ and the measure $m$'s difficulty $\delta$:

$$P(X_{pm} = 1) = \frac{e^{\theta p - \delta^m}}{1 + e^{\theta p - \delta^m}}$$

This inherently suggests that when an item is exactly equal to an individual's ability, the probability that the individual performs this item is exactly 50 percent.

As Sick (2010) demonstrate in figure 1, the model fit can be visually assessed by constructing Item-Characteristic Curves (ICC). In these curves, the x-axis represent the latent trait that is modeled in the Rasch model on a logit scale and the y-axis represents the probability that the individual performs or achieves the item. For example, when considering an individual with an ability of 0.5 logits, figure 1 shows that there is a 20% chance this individual performs D(1), while there is an almost 100% chance they perform D(-1)

### Design Implications

Starke et al. (2020) already showed that by asking a small subset of items, it is possible to create a extensive recommender system of energy-saving measures based on a Rasch model. Furthermore, Nuijten et al. (2022) showed that personalizing physical activity goals are effective in prolonging engagement and increasing positive behavioral change. However, the latter only limited itself to three distinct activities. Moreover, it also adapted its recommendations on a linear increase over time towards a user-set goal. Although their work resulted in useful insights, the current study proposes a significant

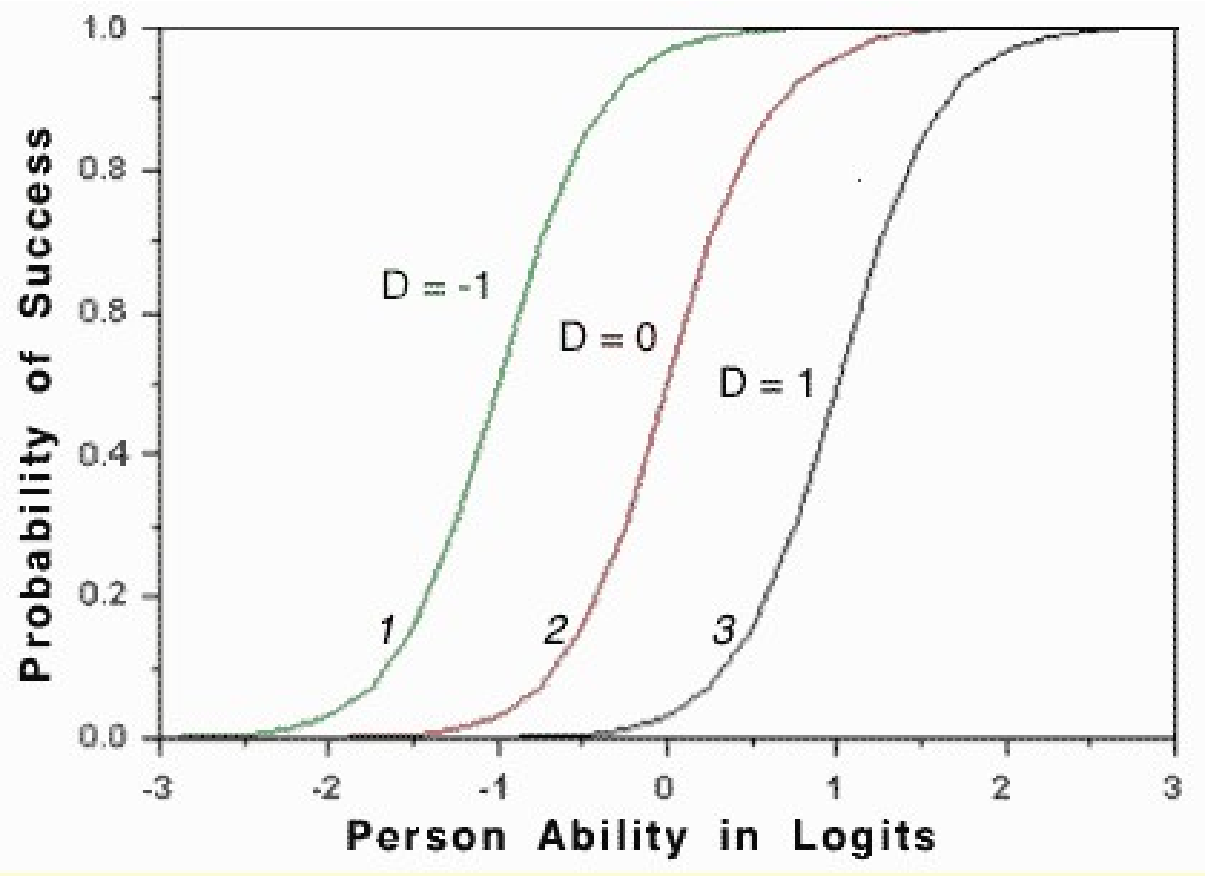


**Figure 1**
*(Hypothetical) Item Characteristic Curves. Retrieved from Sick (2010)*

improvement to their study. Mainly, the current study aims to use a Rasch-scale in order to determine the current activity levels of individuals, whereas Nuijten et al. (2022) asked for an estimation of steps taken, kilometers biked and number of sports sessions on a daily to weekly basis. Rasch modeling will result in a more accurate representation of the current level of activity of the individual, and will furthermore enable a broader range of- and more appropriate recommendations, as seen in Starke et al. (2020).

However, in contrast to the work by Starke et al. (2020), where the researchers used energy-saving measures from a previous study (Knijnenburg et al., 2014), no pre-existing calibrated dataset of measures exist that can be used in creating and calibrating a Rasch-scale in the current domain. Therefore, the present research sets out to use a broad range of registered activities from previous studies to create a list of healthy lifestyle choices that will be used in creating and validating a Rasch scale, and using this scale in a recommender system. These activities are adapted from studies used in Nuijten (2022) and include both food- and physical activity related activities.

**Sub-Research Questions and Research Setup**

Summarizing, the current research sets out to create a general healthy lifestyle recommender system based on both physical activity and dietary intake recommendations. This system can be used by mHealth applications to motivate their users on a personal level to improve on their lifestyle choices. For this purpose, we use the gamified mHealth tool GameBus as used in Nuijten et al. (2022).

However, as such a recommender system has only been applied in specific settings (i.e. medical conditions or professionals), we cannot draw from previous research whether such a system is feasible. Neither can we confidently say that a Rasch scale can be accurately created and calibrated. Therefore, before we can investigate the effects of a Rasch-based recommender system, we need to investigate whether such a scale is possible in the healthy lifestyle domain;

*SQ1: "To what extent is it feasible to create a Rasch-scale with a healthy lifestyle as the unidimensional construct?"*

The current research follows a similar setup as that found in Starke et al. (2015). We first perform a pre-study in which we perform a Rasch analysis on a large set of registered activities to create and calibrate a Rasch scale to determine whether such a scale is feasible. After the scale has been calibrated, we measure the scale fit and reliability. However, the data on which this scale is based is not inherently suitable for Rasch scale calibration. It is sampled from different studies with different main goals and participants (see Table 14.1 in Nuijten (2022)). The studies also include activities that are not applicable to everyday life (e.g. some studies focused on pre adolescents). We therefore continue by asking participants to fill in a questionnaire in which they are asked for each of the activities found in the first scale whether they already perform these in their daily lives. We use the results of this questionnaire to create and calibrate a second Rasch scale. This second Rasch scale will be used to give recommendations in a two-week period. Since there is no research investigating whether such a system is appropriate, we aim to quantify the perceptions participants have regarding the recommendations given. In order to get an accurate representation of the appropriateness of our recommendations, three subjective measures are chosen. First, we want to have insights into whether people are satisfied with their received recommendations, as previous studies have shown that satisfaction has a positive effect on provoking behavioral change (Starke et al., 2017), as well as on both short- and long-term engagement (Zou et al., 2019). Second, we want to find out whether the received recommendations give the person a feeling of motivation, as individuals who are more motivated are likely to interact with the system in the long-term (Zhao et al., 2020). We furthermore make the intuitive assumption that individuals who feel motivated by their recommendations will also report a higher satisfaction with their recommendations and in turn will complete more activities. Finally, in line with our

literature overview, we want insights into how much effort the participants feel the activities take to complete, as we want to find out whether our system does not recommend activities that are either significantly above or below the capabilities of the individual, as well as find out whether this affects the effectiveness of our recommender system. In order to do this, we conduct a questionnaire at the end of the study period that uses an adapted version of the questionnaire found in Starke (2019). This questionnaire consists of questions regarding the perceived effort the recommendations require, the extent to which the given recommendations are motivating and the extent to which participants are satisfied with the recommendations;

*SQ2: "What is the effect of Rasch-based goal personalization on the satisfaction, motivation and perceived effort of recommendations?"*

In order to determine whether recommendations should be either above one's own capabilities (as both Goal-Setting Theory and Flow Theory suggest) or below (as found to be more appropriate by participants in the energy-saving domain (Starke et al., 2020)), we assign participants to either of these two conditions;

*SQ3: "How should healthy lifestyle recommendations be tailored in order to be most effective?"*

## Methods Outline

The current study follows a relatively convoluted approach. Therefore, this preface outlines the next sections, which describe the methods used in the study. The overall study can be divided in two distinct steps.

The first step (Study 1) relates to the creation of a first Rasch scale (further labeled as Scale 1). It will first describe the data that will be used in creating Scale 1 and how it will be cleaned and transformed to a usable dataset. It will then continue with describing the process of creating the Rasch scale. However, as will be apparent in the next section, the scale created in this study is not directly suitable to create and calibrate an effective Rasch scale that can be used for recommendations. Therefore, in order to effectively answer our research question regarding the effectiveness of a recommender system based on a Rasch scale in the healthy lifestyle domain, we first need to re-calibrate this scale.

The second step (Study 2) relates to the recalibration of Scale 1 (further labeled as Scale 2) as well as implementing this scale in a recommender system. We therefore further divide Study 2 in two distinct steps; Study 2a relates to the recalibration of Scale 1. It begins by describing the survey that will be used in order to obtain the data necessary to recalibrate Scale 1 and presents and discusses the scale that will be used in Study 2b. Study 2b covers the implementation of Scale 2 in a recommender system, as well as the methods of the experiment to answer SQ2 and SQ3.

## Study 1: Creating a one-dimensional scale of healthy lifestyle choices

In the following section, the creation of Scale 1 will be discussed in detail.

### Creating the base Rasch Model

The data set retrieved from Nuijten (2022) consists of 543 unique users and 184 unique activities, with a total of 582123 registered activities. These activities include both physical as well as food-related activities. Examples of the first range from "Walk at least X km" to "Participate in a game of dodgeball" or even "Empty out the dishwasher". The latter include activities such as "Eat a piece of fruit" or "Cook a healthy dish for dinner". These activities were included in various other studies in the period between 2018 and 2021. This data however also includes unfinished tasks, system-related entries, duplicate entries and other unusable data. Figure 2 illustrates the data cleaning and transformation process.

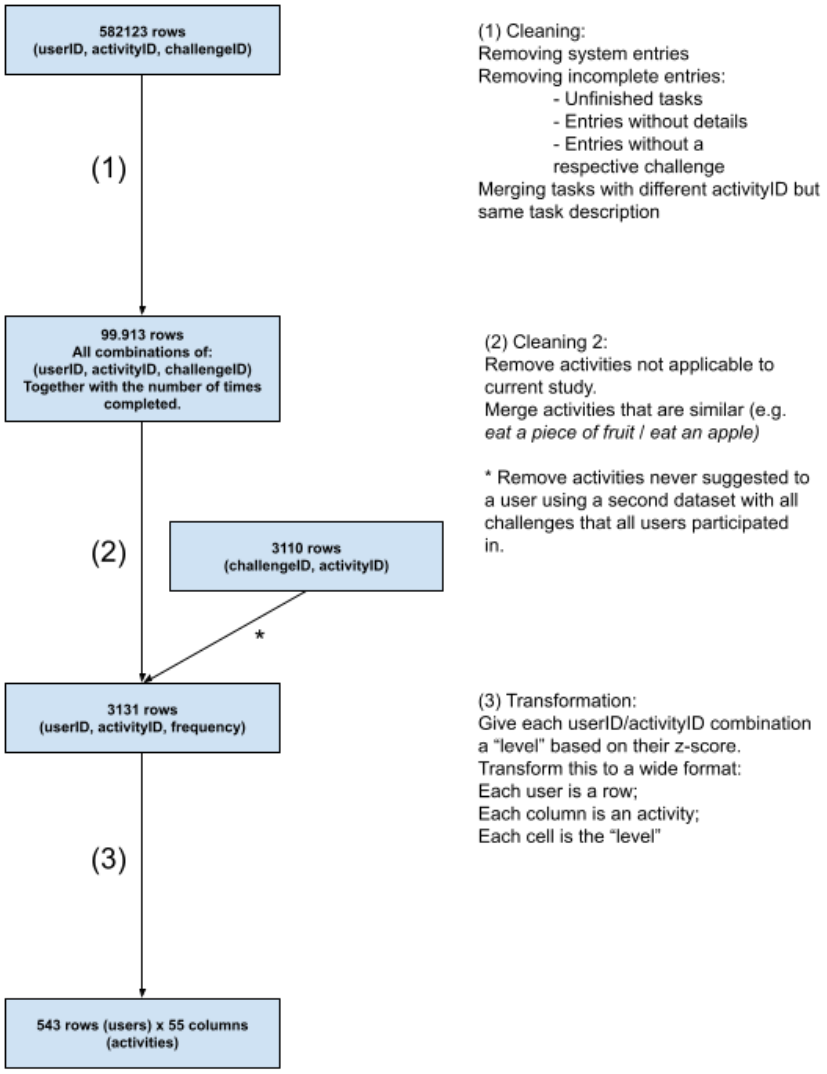


**Figure 2**
*Flowchart illustrating data cleaning and transformation*

All data cleaning and transformation has been performed using the Python 3.9 programming language. [1] The method used in determining the activity level (step 3 in Fig. 2) is given in pseudocode in Alg. 1. To ensure that outliers (i.e. individuals who have an extremely high effect on the mean of activity for a given activity, for example, on average, "Walk at least 500 meters" is completed 3.6 times, while one individual had 281 registrations.) are accounted for and do not negatively affect the validity of the Rasch model, all entries are translated to its z-score. Creation of the Rasch model is performed using the Winsteps software, which is a versatile and well-documented Rasch Measurement & Analysis tool (Linacre, 2022b).

[1] This is mainly done using the `pandas` package, which enables data analysis methods not present in the base Python programming language.

**Algorithm 1** activityLevelFunction()

1: Create a list of all **unique** users
2: Iterate over this list, performing the following:
3: Create a list of all **unique** tasks this user participated in
4: Iterate over this list, performing the following:
   a: Calculate the z-score for this user for this activity. [a]
   b: Depending on this score, give the user an activity level for this activity [b]
   c: Add an entry to a different dataset with the user's ID, the activity's ID and the activity level.

[a] This is done by $Z = \frac{x-\mu}{\sigma}$, where $\mu$ is the mean number of times completed for this activity and $\sigma$ is the standard deviation.

[b] Depending on the amount of standard deviations the user is away from the mean; <-2=1, -2 to -1=2, -1 to 0=3, ..., >2=6

## Results

Scale 1 can be seen in Table 1. The final scale was fitted on 52 items and 539 users. Figure 3 shows the distribution of both activities and users in the Rasch model. We see that both distributions follows neither a normal nor an even distribution.

### *Misfit analysis*

The misfit analysis is focused on determining which activities possible suffered from being overfitted (i.e. fitting too perfectly in the model, which indicates that either an activity is redundant or user responses for a particular activity varied too little) or being underfitted (i.e. when the pattern of responses is unexpected, indicating noise in the data). Out of the 52 items, three items exceeded the prescribed infit guidelines of a MNSQ of 1.4 (see B1), indicating that most items do not suffer from either over- or underfitting.

### *Reliability analysis*

We find that the behavioral cost levels of the activities ranged from $\delta$ 1.73 to $\delta - 3.47$ (M = 0.00; SD = 1.57). We find a larger number of items in the positive direction from the mean than in the negative (38 and 14 respectively). To further investigate model reliability, Winsteps reports reliability statistics. Item reliability was high ($\alpha = 0.95$), suggesting that the order of behavioral cost levels for the activities is appropriate and reproducible (Bond & Fox, 2007). In order to further study the reliability of the model, Winsteps also reports separation indices, which refer to the number of statistically different levels of both person or item difficulty that can be distinguished in the data. Linacre (2006) states that separation values of >2 are satisfactory. Our model shows high item separation (4.31), thus both suggesting that the model can determine a sufficient number of groups between items as well as indicating that the person sample is large enough to confidently confirm the hierarchy

of the items (Linacre, 2022a). Finally, person reliability and separation were relatively low but acceptable (0.62 and 1.28 respectively). High person reliability implies that there is a good diversity in the ability levels of the persons and the number of items per person is sufficient to create a acceptable model. We therefore conclude that our current model required more registered activities per person or a more diverse sample with respect to ability levels of persons to confidently determine the ability of the users (Bond & Fox, 2007).

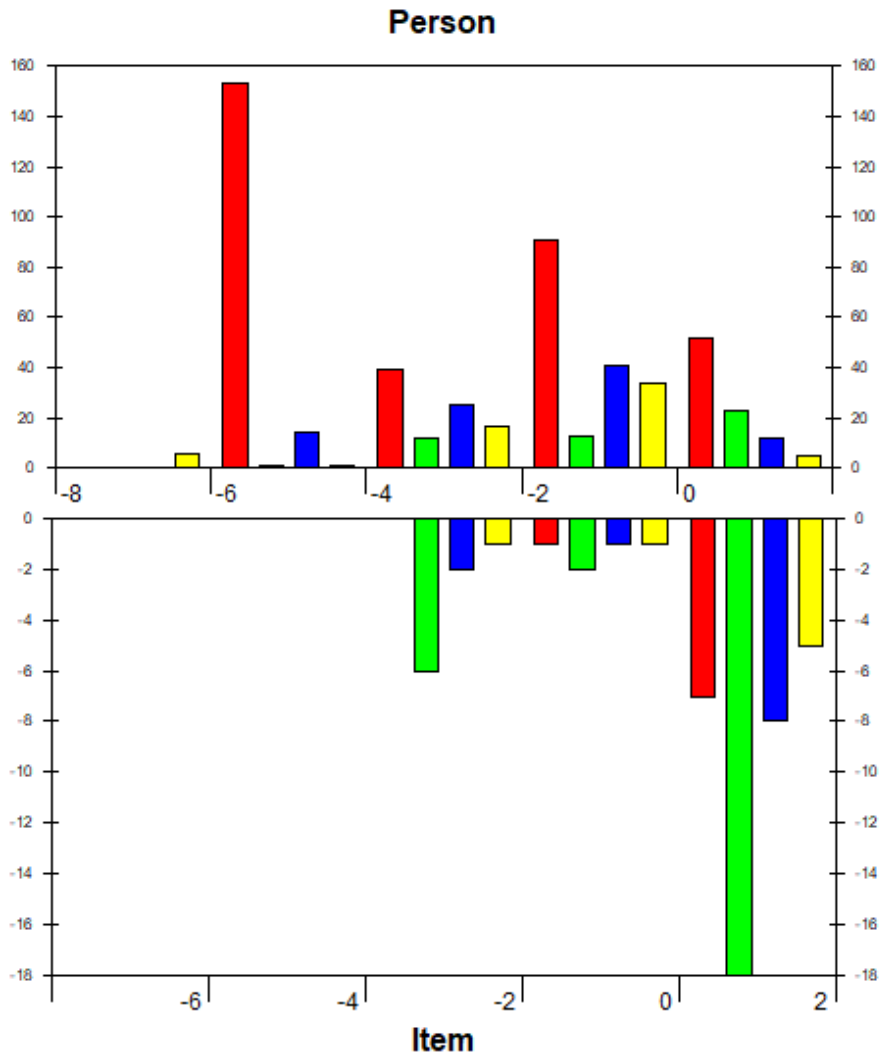


**Figure 3**

*Barplot depicting the distribution of both users and activities over the first Rasch scale. The x-axis represents the difficulty/ability scale, the y-axis represents the number of items/persons.*

| Activity | $\delta$ | MS | ZSTD |
|---|---|---|---|
| walk at least 3 km. | 1.73 | 0.95 | 0 |
| walk at least 1 km. | 1.68 | 0.98 | 0.1 |
| eat a healthy lunch. | 1.62 | 0.83 | -0.30 |
| eat a healthy snack. | 1.62 | 0.77 | -0.50 |
| go for a "roll" for at least 15 minutes (skeelering, ice skating etc.). | 1.50 | 1.66 | 1.50 |
| go for a bike ride of at least 8 km. | 1.49 | 1.36 | 1.10 |
| have a sports sessions of at least 30 minutes. | 1.30 | 1.17 | 1.00 |
| go for a bike ride of at least 1 km. | 1.22 | 0.68 | -1.80 |
| go for a bike ride of at least 4 km. | 1.18 | 1.33 | 1.70 |
| go for a bike ride of at least 2 km. | 1.08 | 0.62 | -0.60 |
| go for a run of at least 15 minutes. | 1.06 | 0.76 | -1.20 |
| go for a bike ride of at least 3 km. | 1.05 | 1.51 | 1.50 |
| walk at least 2 km. | 1.00 | 0.94 | -0.30 |
| take a picture outside. | 0.91 | 1.49 | 1.00 |
| perform at least 15 sit-ups. | 0.88 | 0.98 | 0.10 |
| take a "work walk". | 0.88 | 0.92 | 0.30 |
| participate in a ball sport for at least 25 minutes. | 0.87 | 1.19 | 0.80 |
| go for a swim of at least 15 minutes. | 0.86 | 0.99 | 0.10 |
| take a screenshot of an app that shows 15k steps on one day. | 0.83 | 1.27 | 0.60 |
| walk at least 250 meters. | 0.75 | 0.69 | -2.40 |
| go for a bike ride of at least 15 minutes. | 0.73 | 0.83 | -1.00 |
| perform at least 30 lunges. | 0.71 | 0.35 | -0.90 |
| walk at least 500 meters. | 0.64 | 0.71 | -0.30 |
| eat a piece of fruit. | 0.63 | 0.90 | -0.50 |
| take a picture of yourself in sporting clothes in a park or forest. | 0.60 | 0.76 | -0.40 |
| take a picture of a healthy dish. | 0.60 | 0.47 | -0.80 |
| take a picture of a "healthy" shopping cart. | 0.59 | 1.18 | 0.50 |
| participate in fitness, yoga of dance for at least 25 minutes. | 0.58 | 0.94 | -0.30 |
| take a picture of yourself on your bike with work colleagues. | 0.56 | 0.86 | 0 |
| make a healthy salad. | 0.56 | 1.04 | 0.20 |
| participate in aerobics or a different gym sport for at least 30 minutes. | 0.56 | 1.34 | 0.70 |
| take a picture of yourself on your bike on your way to the gym. | 0.48 | 0.71 | -0.60 |
| make a healthy sandwich. | 0.45 | 0.82 | -0.10 |
| take a picture of yourself at the top of the stairs in the Atlas building. | 0.43 | 1.36 | 0.80 |
| do your groceries by foot. | 0.41 | 1.39 | 1.00 |
| take a picture of yourself while participating in a group physical activity. | 0.30 | 0.73 | -0.70 |
| do at least 15 reps of deskercise. | 0.30 | 1.35 | 1.80 |
| do your groceries by bike. | 0.15 | 0.68 | -1.10 |
| participate in a game of dodge ball. | -0.30 | 1.85 | 1.90 |
| eat a healthy dish. | -0.68 | 1.15 | 0.50 |
| go for a short walk. | -1.22 | 0.68 | -1.40 |
| take a sweaty picture with your friends. | -1.22 | 0.46 | -3.50 |
| go for a run of at least 10 minutes | -1.53 | 0.46 | -3.50 |
| perform the following yoga pose: | -2.44 | 0.86 | -0.80 |
| go for a walk with your dog. | -2.59 | 1.24 | 1.10 |
| take a selfie in a park or forest. | -2.92 | 1.07 | 0.40 |
| go for a short bike ride. | -3.23 | 0.88 | -0.30 |
| participate in a game of checkers. | -3.25 | 1.36 | 0.80 |
| drink a cup of tea. | -3.25 | 0.88 | -0.10 |
| go for a bike ride. | -3.30 | 1.06 | 0.30 |
| perform at least 20 push-ups. | -3.39 | 0.60 | -0.80 |
| participate in an activity that makes you sweat. | -3.47 | 0.73 | -0.40 |

**Table 1**

*All activities included Scale 1. All infit statistics (difficulty ($\delta$) , mean square (MS) and ZSTD (standardized as a z-score) are included as well. More difficult items are at the top.*

## Conclusion

In conclusion, Study 1 produced a diverse unidimensional scale of 52 healthy lifestyle related activities, relating to both physical activity as well as dietary intake choices (see Table 1). We were futhermore able to relatively accurately determine the ability of the 539 persons. The latter is particularly encouraging as, on average, every individual only partook in 12 unique activities. Therefore, with a relatively low amount of data per user, we were able to create and calibrate a Rasch scale with activities from the healthy lifestyle domain. However, the dataset used comes with significant limitations. First of all, the data is comprised of data from a number of different studies. This inherently means that not all data is registered with the same aim in mind. Moreover, not all data is retrieved from the same population (for example, data is registered by both (pre) adolescents as well as adults). Therefore, in order to use the data to effectively create a scale, the data will first need to be recalibrated.

Important to note is that not every activity that is currently present in the scale will be used in Study 2. This is due to some activities being present in the current list of activities that can be merged. For instance, the activities "go for a bike ride" and "go for a short bike ride" are very similar. Furthermore, due to very small differences in behavioral cost levels between some similar activities, some activities are dropped (e.g. cycling 1km, 2km, 3km and 4km have a respective behavioral cost levels of 1.22, 1.08, 1.05 and 1.18). Therefore, of the 52 activities, only 43 are present in Study 2.

Nevertheless, the goal of this study was not to create a scale appropriate for recommendations, but only to investigate the feasibility of such a scale as well as creating a scale or list of a large number of appropriate activities that could be used for recommendations. Therefore, this study nevertheless suggests that a Rasch scale with a healthy lifestyle as the unidimensional construct is feasible and creates a stepping stone on which Study 2 builds upon.

# Study 2: Implementing a Rasch-based Healthy Lifestyle Recommender System

## Study Procedure

In Study 2, we implement the Rasch scale created in Study 1 in a healthy lifestyle recommender system. However, as discussed, the data used in the scale cannot directly be used in creating an effective scale. Therefore, we first re-calibrate the scale on a new sample, after which this re-calibrated scale is used in order to give recommendations. To that end, the first step of Study 2 (Study 2a) relates to the re-calibration of the scale created in Study 1, creating a Rasch scale used that can be used for recommendations. This re-calibration is done by introducing the activities from the current scale to a sample of participants. In the second part of Study 2 (Study 2b), the same participants will receive a number of recommendations for healthy lifestyle choices. This system is implemented in the gamified web application SamenGezond, which is based on and similar to the GameBus application discussed before and used by Nuijten et al. (2022)

# Study 2a: Creating and Calibrating an Appropriate Rasch Scale

## Participants and Research Setup

In Study 2a, we re-calibrated the scale created in Study 1 to be suitable and appropriate for giving accurate recommendations. Participants were told they were invited to test and participate in a new tool which tailored healthy activities to their own capabilities, for which they were first required to fill in a survey regarding their current lifestyle. In total, 193 participants were approached of which a total of 45 responded to the survey. All 193 approached participants were part of a health-related research community.

### *The Survey*

To assess the current lifestyle of every participant, each participant was asked for every activity from the scale from Study 1 to which extent they perform these in their daily lives. For every activity, participants could indicate "never", "rarely", "sometimes", "often" or "not applicable". Examples of not applicable activities could be emptying out a dishwasher when a participant did not have a dishwasher or going to one's gym by bike when a participant did not have a gym membership. At the end of the questionnaire, participants were asked to create an account on the SamenGezond platform, on which their personalized recommendations are given in Study 2b.

Participants who submitted a complete response to the survey entered a 1 in 5 raffle to win a 10 euros giftcard.

# Results

## Onboarding questionnaire

Out of the 45 responses, seven responses were incomplete. All other 38 responses were complete and were included in the analysis. Of the 43 activities that were presented, only two activities were indicated to not be applicable to every participant. "Empty the dishwasher" and "Go to your gym by bike" were filled in by respectively 27 and 29 out of 38 participants, which seems quite logical, as these are two activities that are not applicable to every individual (not everyone has a dishwasher nor has a subscription to a gym).

## The Rasch model

Based on the results of the study, we were successful in creating a new Rasch scale. [2] The resulting model fit the standard guidelines, achieving high levels of reliability. Table 2 summarizes the behavioral cost levels of all activities that were present in the onboarding survey as well as their infit statistics. The table is ordered from low to high behavioral cost levels ($\delta$). A barplot of the distribution of both persons and activities is shown in Figure 4.

The barplot shows that the distribution among the population follows a relatively normal distribution, which is supported by a Shapiro-Wilk test for normality ($p = 0.26$). The distribution of items does not follow a normal, but a bimodal distribution, with two peaks at around both .75 and -.75 ability marks. This could be explained by the fact that the set of activities can be distinguished between both physical activities and food-related activities, resulting in two distinct means for the two groups.

[2] The Rasch model is again created and analyzed using the Winsteps software (Linacre, 2022b)

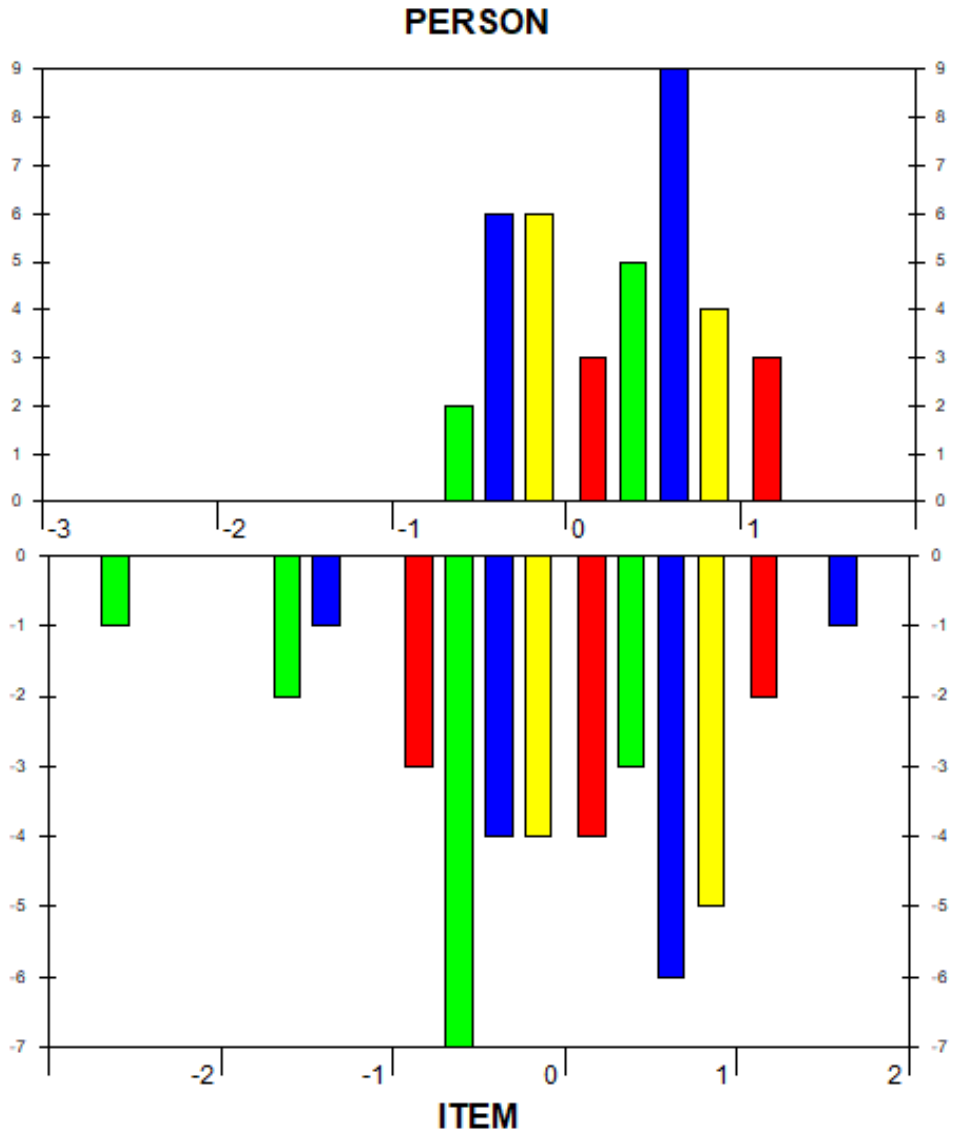


**Figure 4**

*Barplot depicting the distribution of both users and activities over the Rasch scale on which recommendations will be based. The x-axis represents the difficulty/ability scale, the y-axis represents the number of items/persons.*

### *Misfit analysis*

Similar to Study 1, we again perform a misfit analysis. No activities exceeded the infit guidelines for overfitting with no items having a MNSQ of lower than 0.5, indicating that items are generally not redundant (i.e. there is no to little overlap between items) and there is sufficient variation between user responses. Five activities exceeded the infit guidelines, having an infit MS higher than 1.4. This indicates that these items are susceptible to underfitting (i.e. noise in the data that is not modeled). However, due to the already limited number of activities and to retain resolution in the item separation, no items were excluded.

### *Reliability analysis*

The $\delta$-levels of all activities ranged from 1.63 to -2.61 (M = -0.06; SD = 0.86), which includes all person ability levels ($\theta$-levels of 1.08 to -0.66, M = 0.25; SD = 0.52). Winsteps indicated a high item reliability of 0.93 and item separation of 3.65, which are again appropriate statistics (Boone, 2016). Our model furthermore reports a high person reliability of 0.88 and person separation of 2.75. These statistics all indicate that the model can accurately and confidently be used for measurements and recommendations and is likely to be replicated in future studies.

## Conclusion

Study 2a resulted in a unidimensional scale of 43 healthy lifestyle activities, which is suitable for giving recommendations. In contrast to the data used in Study 1, the current study produces a Rasch scale consisting of healthy lifestyle choices that is based on (self-reported) data of an appropriate sample.

To answer our first sub-question, namely *"To what extent is it feasible to create a Rasch-scale with a healthy lifestyle as the unidimensional construct?"*, we discuss our results.

Model statistics were good; item and person reliability as well as their separation statistics were high. We can therefore confidently conclude that it is feasible to construct a Rasch scale with a healthy lifestyle as the unidimensional construct that can effectively be used in order to give recommendations. The scale created in Study 2a can therefore accurately be used in Study 2b, in which participants will receive recommendations based on their ability on this scale. The high model statistics ensure that the recommendations given are accurate with relation to the ability of the participants.

| Activity | $\delta$ | MS | ZSTD |
|---|---|---|---|
| perform some form of physical activity in which you "roll" (e.g. skeelering, ice skating etc.) | 1.63 | 1.17 | 0.6 |
| participate in a ball sports (soccer, volleyball) for at least 25 minutes? | 1.10 | 1.66 | 2.5 |
| go for a swim? | 1.10 | 0.90 | -0.40 |
| go for a run in a park/forest? | 0.96 | 1.05 | 0.30 |
| play at least one game of a ball sport (soccer, dodgeball etc.) | 0.95 | 1.94 | 3.00 |
| perform at least 20 push-ups? | 0.93 | 1.10 | 0.60 |
| perform at least 30 lunges? | 0.90 | 1.09 | 0.50 |
| perform some form of yoga? | 0.86 | 1.05 | 0.30 |
| participate in a mental board game? (e.g. checkers, chess etc.) | 0.70 | 1.51 | 2.10 |
| perform at least 15 sit-ups? | 0.63 | 0.91 | -0.40 |
| do at least 15 squats? | 0.63 | 1.14 | 0.70 |
| go for a long bike ride? (>17.5km) | 0.61 | 0.89 | -0.50 |
| follow a biking route? | 0.58 | 0.96 | -0.20 |
| perform at least 30 minutes of aerobics or a similar gym sport? | 0.52 | 1.17 | 0.90 |
| perform deskercise? (i.e. stretching out, do light workouts behind your desk while working/studying) | 0.49 | 1.12 | 0.70 |
| do at least 25 minutes of fitness? | 0.32 | 0.98 | -0.10 |
| use stairs as an exercise? | 0.30 | 1.03 | 0.20 |
| participate in some form of physical group activity? (e.g. group sports lessons, morning running classes) | 0.24 | 1.10 | 0.60 |
| take at least 15.000 steps on a day? | 0.22 | 0.79 | -1.20 |
| participate, with friends, in a physical activity that makes you sweat? | 0.11 | 0.86 | -0.70 |
| do your groceries by foot? | 0.03 | 0.90 | -0.50 |
| visit a/your local park/forest? | -0.06 | 0.63 | -2.30 |
| go for a walk during your work/studies? | -0.09 | 0.89 | -0.50 |
| eat a healthy salad? | -0.20 | 0.76 | -1.30 |
| go to your gym by bike? | -0.20 | 1.24 | 1.10 |
| take the stairs instead of the lift on work/on campus? | -0.35 | 0.96 | -0.10 |
| drink a cup of tea? | -0.41 | 0.68 | -1.70 |
| go for a leisurely walk? | -0.41 | 0.68 | -1.60 |
| empty the dishwasher? | -0.43 | 1.13 | 0.60 |
| perform any sports for at least 30 minutes? | -0.51 | 1.12 | 0.60 |
| eat a healthy sandwich? | -0.52 | 0.56 | -2.30 |
| do your groceries by bike? | -0.56 | 1.43 | 1.80 |
| eat a healthy snack? | -0.64 | 0.53 | -2.30 |
| go to work/study by bike? | -0.68 | 1.54 | 2.00 |
| go for a short bike ride? | -0.71 | 1.28 | 1.20 |
| perform some form of physical activity that makes you sweat? | -0.73 | 0.88 | -0.50 |
| do you shop for healthy groceries? (no ready-to-go meals, sugar-rich drinks etc.) | -0.81 | 1.05 | 0.30 |
| eat a piece of fruit? | -0.94 | 0.76 | -0.90 |
| eat healthy for lunch? | -0.99 | 0.91 | -0.20 |
| eat a healthy dish as dinner? | -1.33 | 0.75 | -0.70 |
| eat a dish which you find healthy? | -1.56 | 1.11 | 0.40 |
| go outside? | -1.65 | 0.67 | -0.70 |
| take your dog for a walk? (if you have no dog, indicate "Not applicable") | -2.61 | MIN | MIN |

**Table 2**

*All activities included Scale 2. All infit statistics (difficulty ($\delta$) , mean square (MS) and ZSTD (standardized as a z-score) are included as well. More difficult items are at the top.*

## Study 2b: Investigating the Effectiveness of Rasch-based Healthy Lifestyle Recommendations

The current section starts with a description of the general procedure of the experiment and describes the process of matching a participant to their corresponding set of activities. It then continues with a description of the different metrics that were measured. It ends with a description of the survey that is conducted at the end of the study.

### Research Setup

In Study 2b, the scale created in Study 2a is implemented in a healthy lifestyle recommender system. As discussed in the literature overview, research gives no clear answer as to whether recommendations should be challenging or easy to perform. Theory and literature suggests that the first should be more effective (Adams et al., 2017; Csikszentmihalyi, 2009; Locke & Latham, 2002), however, some studies show that people have a tendency to view easier recommendations as more appropriate (Starke et al., 2020).

To that end, the current study followed the following procedure; first, all participants from Study 2a received an invitation (by email) in which they were told that they have received a list of personal recommendations for activities on the SamenGezond platform and that they are free to complete these activities to personal liking for the next two weeks. These recommendations were based on 1) their ability level, which is determined in the Rasch analysis in Study 2a and 2) their experimental condition, further labeled as "easy" or "hard". Participants were randomly assigned to either of the conditions. Depending on this condition, participants will receive recommendations which are either below ("easy") or above ("hard") their current ability level.

One week after the two-week activity period ended, we conducted a survey to determine the appropriateness of the recommendations. This survey will include an adapted version of the survey used in Starke et al. (2015).

#### *Matching Participants to their Recommendations.*

During the Rasch analysis in Study 2a, we have determined the current ability level for each participant. In order to determine which recommendations each participant should receive, we divided the list of 43 activities into thirteen subsets of six activities and determined for each subset the mean difficulty.

This process went as follows: starting from the low end of scale (i.e. the items with the lowest behavioral cost), the first group will consist of the first six activities, from which the next group will consist of the next six activities. This continues until the scale has fully been divided, after which a second number of groups will be created. These consists of the three items with the highest behavioral cost ("hardest") from subset 1 and the three items with the lowest behavioral cost ("easiest") from subset 2.

This again continues until the full scale has been divided (with the exception of six items; the three items with the lowest behavioral cost and highest behavioral cost). An abstract of this procedure is shows in Table 3.

| Group | Activity | Group |
|---|---|---|
| Subset 1. $\mu$ = X | Activity 1 | |
| | Activity 2 | |
| | Activity 3 | |
| | Activity 4 | Subset 7. $\mu$ = Y |
| | Activity 5 | |
| | Activity 6 | |
| Subset 2. $\mu$ = Z | Activity 7 | |
| | Activity 8 | |
| | Activity 9 | |
| | Activity 10 | |
| | Activity 11 | |
| | Activity 12 | |

**Table 3**
*Abstract of the procedure of dividing the list of activities into thirteen subsets.*

After the list has been divided, participants were matched to the group with a mean closest to their ability. Participants received the six activities in the set either above or below their matched set (depending on their experimental condition) as recommendations on the SamenGezond platform.

In total, of the 38 valid responses of the onboarding questionnaire, 26 participants created an account on the SamenGezond platform and were thus able to receive their appropriate recommendations. After the two-week period has ended, all participants (thus also participants that did not complete any activities) received the final questionnaire, in which they were asked for their perceptions on the received recommendations.

## Measures

### *Objective Measures*

To determine the effectiveness of the recommendations, we kept track of the number of activities that participants registered. This allows for comparison between the difficulty groups and enables analysis of which factors affect performance of the participant.

### *Subjective Measures*

Several subjective metrics regarding the participants' perceptions on the recommendations were measured in the survey at the end of the study. The background as to why these measures were chosen was discussed at the end of the Background section. These measures consist of the perceived effort a recommendation takes as well

| Survey Item | Factor Loading |
|---|---|
| *Perceived Effort ($\alpha$ = .55), AVE = 0.36* | |
| The recommended activities would be hard to perform | 0.97 |
| It would take little effort to perform the recommended activities. | -0.30 |
| These activities would push my limits | 0.22 |
| *Satisfaction ($\alpha$ = .81), AVE = 0.39* | |
| I am happy with the activities recommended to me. | -0.56 |
| I would like to see different activities than the ones that were recommended to me. | 1.00 |
| The activities recommended to me suit me well. | -0.32 |
| The activities would be fun to perform. | -0.37 |
| *Motivation ($\alpha$ = .60), AVE = 0.26* | |
| The recommended activities would be applicable to my personal situation | 0.59 |
| The recommended activities would motivate me to perform them | 0.43 |
| The activities would help me in achieving my health goals | 0.52 |

**Table 4**

*Factors and factor loadings based on a maximum-likelihood analysis for survey items from Study 2b.*

as the extent to which the recommendations were motivating and satisfactory. The questions found in the survey were retrieved and adapted from Starke et al. (2017), in which internal consistency was confirmed and all items were validated. In the final questionnaire, participants received a total of nine survey items (three for each measure). Table 4 presents all survey items with their respective factor loadings.

Table 4 shows that the factor analysis on the subjective measures is unsatisfactory. Internal validity could not be verified due to low AVE indices. Moreover, internal consistency could only be verified for satisfaction, as both perceived effort and motivation show low Cronbach's alpha values. This may mainly be due to the small sample size (N=11), which is insufficient to conduct a statistically significant factor analysis (Mundfrom et al., 2005). We therefore retain the items as found in Starke et al. (2015) and average these with equal weights.

### *Adjustments to the Final Questionnaire*

Unfortunately, participant response rates during the two-week activity period were too low (N=4). This meant that, if no action was taken, no statistically significant conclusions could be drawn. Therefore, to alleviate this problem and to nevertheless receive an acceptable number of data points, the final survey received two additions. First, the survey showed the list of recommendations the participant received, regardless of whether they have performed any activities. This meant that participants that did not participate in the two-week activity period or did not log into the SamenGezond platform at all were still made aware of the recommendations they received and were thus able to give their perceptions on these. Second, all participants did not only receive questions regarding their received recommendations (labeled as *RList* in the next sections), but were also presented with two lists of new recommendations (labeled

as *NList* in the next sections). These lists consisted of the sets of activities that were both one group above or below their assigned recommendations. For example, when regarding the example in Table 3, if the participant received Subset 2 as their list of recommendations, the final section of the survey will also present the participant with the activities from subsets 1 and 3. In this way, we were still able to measure the effects of different difficulty levels, even though not all users engaged with the platform. [3]

[3] Since participants now participated in two studies, compensation was increased. Therefore, participants that responded to the final questionnaire could forfeit their 20% chance for the 10 euros giftcard and receive a 5 euros giftcard for certain instead.

## Results

Out of the 26 participants that were approached, 4 participants registered at least one completed activity. A total of 11 participants filled in the final questionnaire. Table 5 shows the statistics for the three subjective measures.

| Measure | Mean | SD | 95% Conf. Interval |
|---|---|---|---|
| Satisfaction | 3.42 | 0.17 | 3.05 - 3.80 |
| Perceived Effort | 3.09 | 0.19 | 2.68 - 3.50 |
| Motivation | 3.06 | 0.13 | 2.76 - 3.36 |

**Table 5**
*Mean, standard deviation (SD) and 95% confidence intervals for the three subjective measures. Note that the measures are based on a scale ranging from 1 to 5 and Perceived Effort is reversely coded (i.e. a higher mean value represents a lower perceived effort).*

### Structural Equation Model

As multiple metrics with multiple possible relations and interaction effects are measured, we organize all measures in a Structural Equation Model (SEM) found in Figure 5. This model relates to the survey results regarding the received recommendations (*RList*). Structural Equation Modeling is a powerful tool due to the fact that it is capable of evaluating both direct and indirect effects on casual relationships between variables (Fan et al., 2016).

The overall model had an statistically insignificant fit: $\chi^2(10)=15.890$ $p = 0.103$, $CFI = 0.950$, $TLI = 0.499$, $RMSEA = 0.172$ The low Tucker-Lewis Index (TLI) and high root mean square error of approximation (RMSEA) index shows that the model-data fit is not sufficient (L. T. Hu & Bentler, 1999). We see that the difficulty of the recommendations has a direct effect on the number of activities that were completed. (*coef.* = -0.556, $p<0.05$), which confirms that harder activities result in lesser activities completed. Figure 5 also shows that satisfaction has significant negative effect on the number of completed activities. This is strange, as previous literature shows satisfaction having a positive effect on the number of completed items (e.g. Starke et al. (2017)). We furthermore see a significant effect of motivation on satisfaction, indicating that people that see the recommendations as more motivating are also more satisfied with their recommendations.

However, the overall sample size of 11 (and N=4 for *# of completed activities*) is statistically too small to confidently draw conclusions from this model. See Appendix A for the full power analysis. Therefore, in order to draw statistically sound conclusions, we continue with an investigation on possible effects, disregarding the SEM model from Figure 5 as the ground truth.

The remainder of the current section consists of two parts: first, it describes the results from the survey that regard *RList*. It analyzes the effects of difficulty on the three subjective measures, as well as the effects both difficulty as well as the subjective

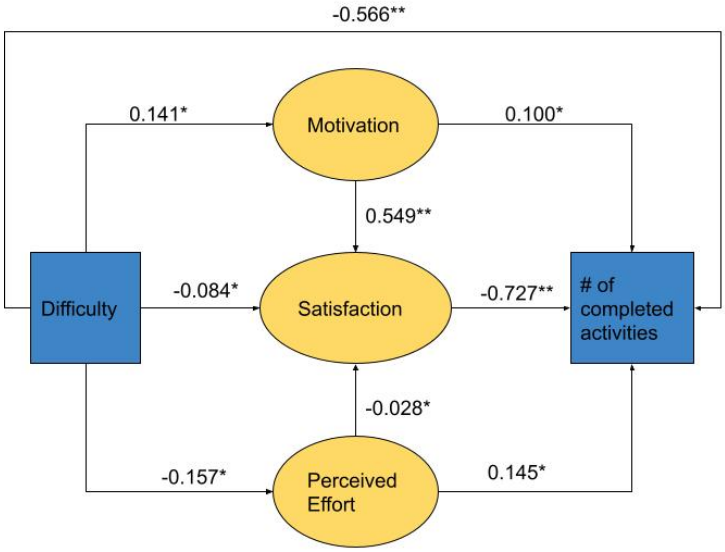


**Figure 5**

*SEM for Study 2. Number at the arrows represent the standardized coefficients of the effects (negative equates to a negative correlation). ** p<0.05, *p>0.05*

measures on the number of completed activities. The final segment reports the results from the items from *NList*.

### *Received Recommendations (RList)*

We find no significant differences for all measures between the two manipulation groups. (See Figure 6) Further analysis shows no effect of item difficulty on the number of completed activities. (*coef.* -3.4, p=.27), nor on the subjective measures (*Satisfaction: coef. -.32, p=.37; Perceieved Effort: coef. -.13, p=.64; Motivation: coef. .17, p=.68*)

Further analysis also shows no effect of the subjective measures on number of completed activities (*Satisfaction: coef. -4.2, p=.13; Perceieved Effort: coef. 1.34, p=.72; Motivation: coef. -2.34, p=.38*).

Finally, we find no effects of either motivation or perceived effort on satisfaction (*coef. .44, p=.13 and coef. .25, p=.56 respectively)*

### *New Recommendations (NList)*

We continue our analysis with the results from the second part of the survey, in which participants were presented two new lists of activities; one below and one above their received recommendations (*NList*). However, in contrast to the questions regarding *RList*, activities were presented separately, instead of an overall list of the six activities.

Figure 7 shows the means and significance of the difference between these for the three subjective measures. We find a significant difference between the two different difficulty levels with regards to satisfaction and motivation (p=0.02 and p=0.03 respectively). We find no significant effect for perceived effort (p=0.15), though the effect is in the direction we would expect.

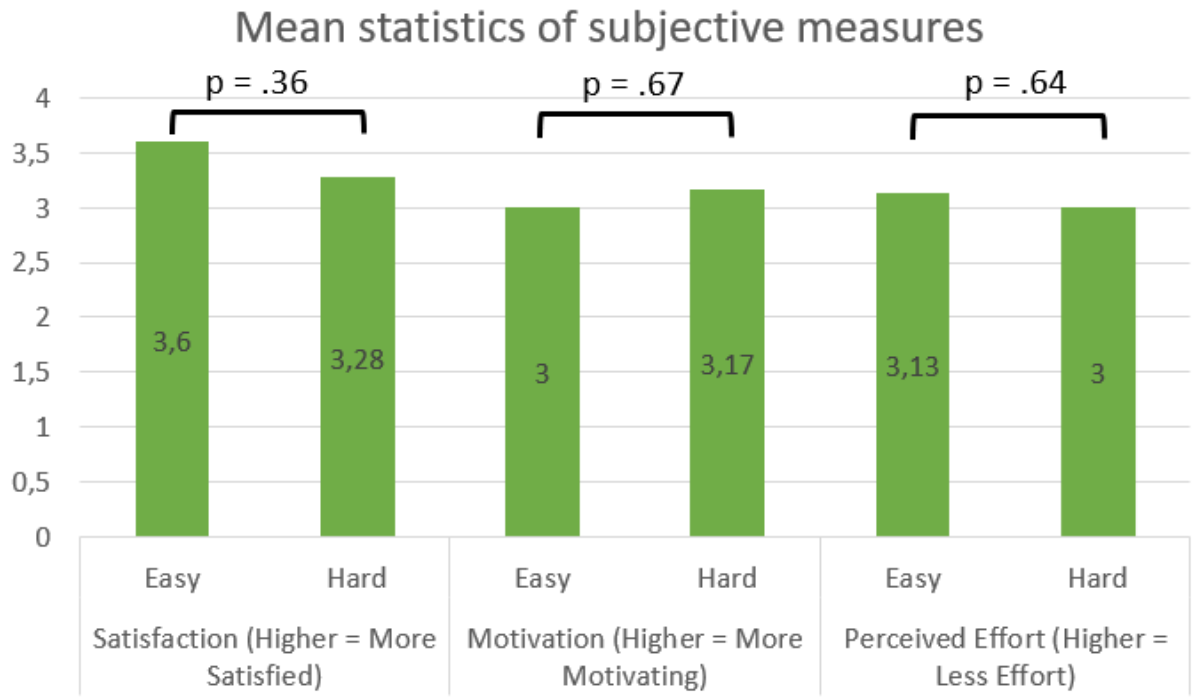


**Figure 6**

*Barplot summarizing the means of each subjective measure for both manipulation groups for the received recommendations (RList). Significance is shown on top. Note the reverse coding on Perceived Effort. N=11*

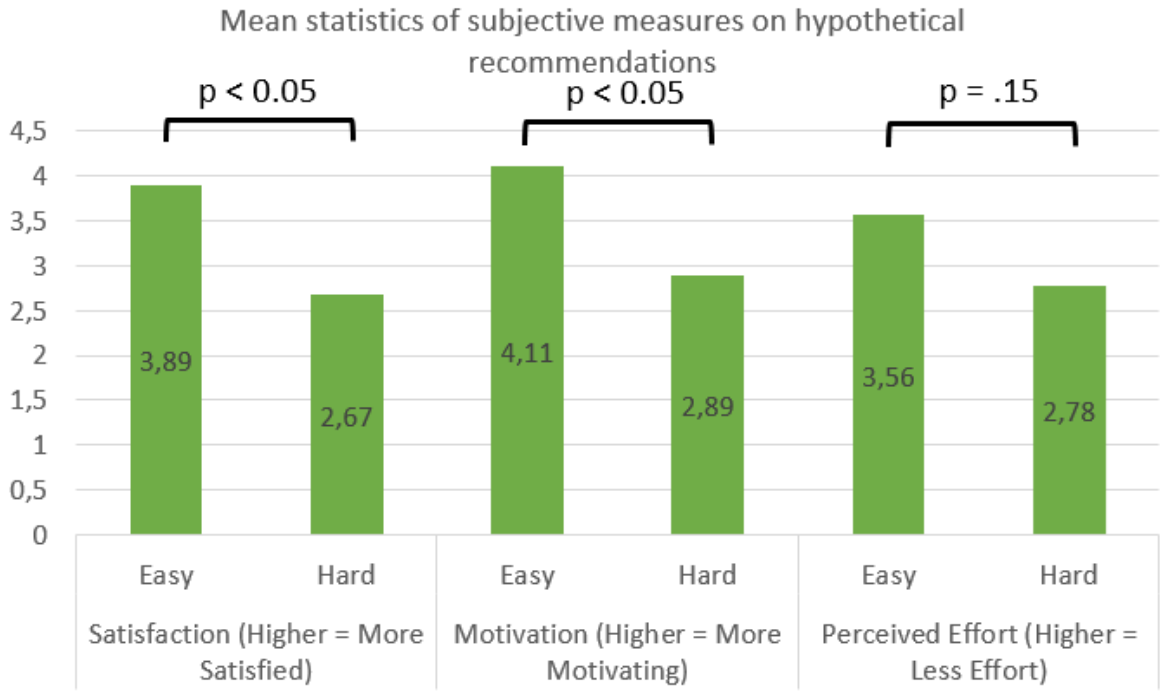


**Figure 7**

*Barplot summarizing the means of each subjective measure for both manipulation groups for the newly presented recommendations (NList). Significance is shown on top. Note the reverse coding on Perceived Effort. N=11*

## Discussion

We started the current study with the aim of creating a novel method of personalizing healthy lifestyle recommendations on the principle of Rasch analysis. We aimed to investigate whether such an method was even feasible, as this has not been done before in a generalized setting. We then continued with investigating the effect of such a system and finally investigated how these recommendations should be tailored.

### Reviewing the Research Question and its Sub-Questions

We found that it is in fact possible to create a Rasch scale on which a healthy lifestyle is the unidimensional construct. Activities could be arranged on their respective difficulties and individuals on their respective abilities. This scale can confidently be used in order to give lifestyle recommendations. We can therefore, with confidence, positively answer our first sub-question of *"To what extent is it feasible to create a Rasch-scale with a healthy lifestyle as the unidimensional construct?"* with the notion that this is, in fact, feasible.

To answer our second sub-question of *"What is the effect of Rasch-based goal personalization on the satisfaction, motivation and perceived effort of recommendations?"*, we cannot find a direct answer from our results. This is due to the fact that we have no baseline comparison, nor do we have a study that studied the same measures that have been measured in the current study. Therefore, we can only speculate on the following; on average, people seemed relatively satisfied with the received recommendations (see Figure 3), as on a 5-point Likert scale the average satisfaction was 3.44. Moreover, participants reported that the received recommendations were moderately motivating (with a mean of 3.09), as well as being perceived as moderately easy to perform (with a mean of 3.07, reverse coded). We therefore see modest positive effects of our Rasch-based personalization on these subjective measures, although these effects cannot be compared to a control condition.

Finally, to answer our third and final sub-question of *"How should healthy lifestyle recommendations be tailored in order to be most effective?"*, we first need to determine whether our manipulation was effective. Therefore, we compare the observed means of the three subjective measures between the two manipulation groups. We expect that the group that received recommendations that are above their ability level report a lower mean on perceived effort (due to reverse coding) than those receiving "easier" recommendations and report a lower mean on satisfaction and motivation (in line with previous studies). We find no significant effects for neither of the three measures between the two manipulation groups on the received recommendations (*RList*). However, when comparing the two additional lists of activities (*NList*), we find significant effects between the two difficulties with regards to satisfaction and motivation. These effects are in line with previous work, with both satisfaction and motivation being higher for the activities that are easier. Although this did not translate into a higher number of completed activities, we suppose that, similar to

previous studies, activities that are below one's ability are more effective in provoking behavioral change.

### *Study Limitations*

The model we have created in order to give activity recommendations comes with its own limitations. One of these is that the model is constrained with respect to the diversity and number of activities that can be recommended. The current system only has a total of 43 activities present, which does not encompass all possible activities. For example, one participant indicated that he/she hiked very frequently, but was not able to indicate this in the study.

Furthermore, the use of the SamenGezond platform comes with the constraint that this is a web application, which means that viewing and recording activities requires active participation and awareness of the fact that these activities should be registered (since participants are required to open the web application and register their activity every time one is completed). An improvement on how to lower the required effort to register an activity could be to use wearable technology to register physical activities.

One of the main limitations of the current study is the relatively small sample size. To effectively determine differences between the effectiveness of manipulating the difficulty of the recommendations, power analyses show a required sample size of around 100 (see Appendix A), while the current study obtained 38 (complete) responses to the first survey, of which only four proceeded to register at least one completed activity. This is at least partly alleviated by the adjustments made in the final questionnaire, since this resulted in a lower number of participants needed to show significant results, but the number of participants required was still higher than the 11 found in Study 2b.

Finally, the study limited itself to the population of only a database of individuals who have indicated that they are willing to participate in health-related studies, resulting in potential biases in the data of both surveys as well as the behavioral data from the activity period.

## Conclusion and Future Work

The current research presents a study that explored the possibility of a novel method of healthy lifestyle recommender system and investigated the effect of tailoring recommendations based on an individual's current ability.

In order to achieve this, we used pre-existing behavioral data recorded over several studies in which individuals were able to perform several activities. (543 users, 184 tasks) We used this data in order to create an initial scale on which both the activities and the individuals could be arranged on their respective difficulty and ability. The activities from this scale were used for conducting a survey (N=38), on which a new scale was built. Using this scale, we conducted a two-week study in which individuals received personal recommendations for healthy lifestyle activities they could perform, that were either above or just below their ability (N=4). We used this data, together with data of a survey conducted after the two-week period (N=11) to investigate the effects of difficulty tailoring.

Our results suggest that it is possible to create a Rasch scale that can be used for a healthy lifestyle recommender system. Our within-user results further suggest that people find easier recommendations as more appropriate, being both more motivating and satisfactory.

### Scientific Relevance and Design Implications

The current study creates a basis on which future studies on healthy lifestyle recommender system can build upon. It shows that such a system is feasible and can effectively be used in order to give accurate recommendations.

#### *Future Work and Scientific Relevance*

Future studies could expand on the current study, either replicating these studies on a larger, more diverse sample. This would increase the generalization of the results, as the sample used in the current study is relatively homogeneous. This would furthermore also enable more specific analyses, for example, performing studies on whether different kinds of activities (i.e. physical, dietary etc.) require a different approach (e.g. should physical activities be more challenging than dietary?).

Another interesting extension to the current study could be to investigate whether extending the activities included improves effectiveness. Examples of such an extension could be to implement social activities, such as visiting one's grandmother, or cognitive activities such as completing a puzzle or reading a book.

Furthermore, the current findings provides a significant improvement to the current knowledge on goal tailoring in the domain of recommender systems. These systems are already widely being studied in different areas such as energy-saving behavior (Starke et al., 2020) and dietary intake (Radha et al., 2016) and the current study extends on these studies by showing that such a system is feasible in the domain of lifestyle interventions. Not only does it prove the feasibility of such a system, it also

proves that it can be effectively be used in an already existing platform to (potentially) improve its effectiveness. It furthermore is in line with and extends on the current knowledge on the personalization of goals, suggesting that easier recommendations result in higher satisfaction as well as a higher level of motivation for their recipients. This knowledge can be used by applications aimed at providing their users with an accessible method of improving their lifestyle choices.

### *GameBus and Rasch-based Challenge Personalization*

In the literature overview, we touched upon the GameBus platform, a healthy lifestyle-oriented platform aimed at gamifying the experience of becoming more physically, socially and mentally fit. We also discussed a main limitation of the principle of challenges on which GameBus is based. Namely, every participant received the same activities and rewards. However, this results in unequal opportunities for each participant, as not all participants are equally fit or have the same abilities.

The current study suggests that the implementation of a recommender system in such a platform could be very beneficial. Such a system would create a tailored experience to every user of the application, creating equal opportunities for everyone. In the specific context of GameBus, the organizer of a challenge could set the relative difficulty the suggested activities should have in relation to the ability of the user instead of setting specific activities. For example, when the aim of the challenge is to simply make everyone participate, activities could be set to a relatively "easy" level, whereas when used in a more challenging situation (e.g. training for a fitness-related competition), relative difficulty could be much higher than one's ability. Users can also indicate that they are unable to complete an activity (i.e. they cannot go for a bike ride when they have no bike), which replaces the activity with a different one with a similar difficulty.

This would implement a Rasch analysis that continuously updates whenever a challenge has been completed, updating both item difficulties and user abilities with the results of the challenge. This would ensure that 1) item difficulties will be accurately determined and 2) user ability levels keeps being updated, accounting for improvements (or deterioration) of one's healthy lifestyle. Rewards for the given recommendations could be set appropriately, rewarding the participant with more points for completing an activity with a higher difficulty than one with a lower difficulty, thus challenging every participant to a similar degree.

## References


Adams, M. A., Hurley, J. C., Todd, M., Bhuiyan, N., Jarrett, C. L., Tucker, W. J., Hollingshead, K. E., & Angadi, S. S. (2017). Adaptive goal setting and financial incentives: a 2 × 2 factorial randomized controlled trial to increase adults' physical activity. *BMC Public Health*, *17*(1), 1–16. https://doi.org/10.1186/s12889-017-4197-8

An, X., & Yung, Y.-f. (2014). Item Response Theory : What It Is and How You Can Use the IRT Procedure to Apply It. *SAS Institute Inc.*, 1–14. https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch Model : Fundamental Measurement in the HumBond, T. G., & Fox, C. M. (2007). Applying the Rasch Model : Fundamental Measurement in the Human Sciences Second Edition University of Toledo.an Sciences Second Edition University of Toledo.

Bonenfant, J. L. (1971). Le professeur Carlton Auger (1912-1970). *Laval medical*, *42*(5), 423–427.

Boone, W. J. (2016). Rasch analysis for instrument development: Why,when,and how? *CBE Life Sciences Education*, *15*(4). https://doi.org/10.1187/cbe.16-04-0148

Brown, T. (2018). Pros & Cons of Artificial Intelligence in Medicine | Drexel CCI. Retrieved December 17, 2021, from https://drexel.edu/cci/stories/artificial-intelligence-in-medicine-pros-and-cons/

Chapman, G. B., Colby, H., Convery, K., & Coups, E. J. (2016). Goals and Social Comparisons Promote Walking Behavior. *Medical Decision Making*, *36*(4), 472–478. https://doi.org/10.1177/0272989X15592156

Csikszentmihalyi, M. (2009). *Flow: The psychology of optimal experience*. Harper; Row.

Dharia, S., Jain, V., Patel, J., Vora, J., Chawla, S., & Eirinaki, M. (2016). PRO-Fit: A personalized fitness assistant framework. *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*, *2016-January*, 386–389. https://doi.org/10.18293/SEKE2016-174

Donciu, M., Ioniţă, M., Dascălu, M., & Trăuşan-Matu, Ş. (2011). The runner - Recommender system of workout and nutrition for runners. *Proceedings - 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2011*, (January), 230–238. https://doi.org/10.1109/SYNASC.2011.18

Faiz, I., Mukhtar, H., & Khan, S. (2014). An integrated approach of diet and exercise recommendations for diabetes patients. *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services, Healthcom 2014*, 537–542. https://doi.org/10.1109/HealthCom.2014.7001899

Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S. R., Park, H., & Shao, C. (2016). Applications of structural equation modeling (SEM) in ecological studies: an updated review. *Ecological Processes*, *5*(1). https://doi.org/10.1186/s13717-016-0063-3

Ferretto, L. R., Bellei, E. A., Biduski, D., Bin, L. C. P., Moro, M. M., Cervi, C. R., & De Marchi, A. C. B. (2020). A Physical Activity Recommender System for Patients with Arterial Hypertension. *IEEE Access*, *8*, 61656–61664. https://doi.org/10.1109/ACCESS.2020.2983564

Ferwerda, B., & Schedl, M. (2016). Personality-based user modeling for music recommender systems. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9853 LNCS*, 254–257. https://doi.org/10.1007/978-3-319-46131-1_29/COVER/

Galli, S., Chiesi, F., & Primi, C. (2008). The construction of a scale to measure mathematical ability in psychology students: An application of the rasch model. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *15*(1), 3–18.

*Gamebus*. (2016). Retrieved June 12, 2022, from https://blog.gamebus.eu/

Gomez-Uribe, C. A., & Hunt, N. (2015). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, *6*(4). https://doi.org/10.1145/2843948

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hu, R., & Pu, P. (2009). Acceptance issues of personality-based recommender systems. *RecSys'09 - Proceedings of the 3rd ACM Conference on Recommender Systems*, 221–224. https://doi.org/10.1145/1639714.1639753

Kaiser, F. G., Byrka, K., & Hartig, T. (2010). Reviving campbell's paradigm for attitude research. *Personality and Social Psychology Review*, *14*, 351–367. https://doi.org/10.1177/1088868310366452

Knijnenburg, B. P., Willemsen, M. C., & Broeders, R. (2014). Smart sustainability through system satisfaction: Tailored preference elicitation for energy-saving recommenders. *20th Americas Conference on Information Systems, AMCIS 2014*, (August).

Liang, Y., & Willemsen, M. C. (2019). Personalized recommendations for music genre exploration. *ACM UMAP 2019 - Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 276–284. https://doi.org/10.1145/3320435.3320455

Linacre, J. M. (1994). Sample Size and Item Calibration or Person Measure Stability. *Rasch Measurement Transactions*, *7*(4), 328. http://www.rasch.org/rmt/rmt74m.htm

Linacre, J. M. (2006). *A user's guide to winsteps ministep rasch-model computer programs. program manual 4.4.7.* https://archive.org/details/B-001-003-730

Linacre, J. M. (2022a). *Reliability and separation of measures: Winsteps help.* https://www.winsteps.com/winman/reliability.htm

Linacre, J. M. (2022b). Winsteps rasch measurement computer program (version 3.70.1). http://www.winsteps.com

Locke, E. A. (1968). Toward a Theory of Task Motivation and Incentives. *Organizational Behavior and Human Performance*, *3*(1000), 157–189.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717. https://doi.org/10.1037/0003-066X.57.9.705

Locke, E. A., Latham, G. P., Locke, E. A., & Latham, G. P. (2015). New Directions in Goal-Setting Theory New Directions in Goal-Setting Theory. *Psychological Science*, *15*(October), 265–268.

McEwan, D., Harden, S. M., Zumbo, B. D., Sylvester, B. D., Kaulius, M., Ruissen, G. R., Dowd, A. J., & Beauchamp, M. R. (2016). The effectiveness of multi-component goal setting interventions for changing physical activity behaviour: A systematic review and meta-analysis. *Health psychology review*, *10*(1), 67–88.

Medica Magazine. (2018). Diagnosing diseases with big data – MEDICA - World Forum for Medicine. Retrieved December 17, 2021, from https://www.medica-tradefair.com/en/News/Topic_of_the_Month/Older_Topics_of_the_Month/Topics_of_the_Month_2018/Big_data_in_diagnostics/Diagnosing_diseases_with_big_data

Michie, S., West, R., Sheals, K., & Godinho, C. A. (2018). Evaluating the effectiveness of behavior change techniques in health-related behavior: A scoping review of methods used. *Translational Behavioral Medicine*, *8*(2), 212–224.

Milne-Ives, M., LamMEng, C., de Cock, C., van Velthoven, M. H., & Ma, E. M. (2020). Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: Systematic review. *JMIR mHealth and uHealth*, *8*(3), 1–16. https://doi.org/10.2196/17046

Moon, D. H., Yun, J., & McNamee, J. (2016). The effects of goal variation on adult physical activity behaviour. *Journal of Sports Sciences*, *34*(19), 1816–1821. https://doi.org/10.1080/02640414.2016.1140218

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum Sample Size Recommendations for Conducting Factor Analyses. *International Journal of Testing*, *5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Nuijten, R., Van Gorp, P., Khanshan, A., Le Blanc, P., van den Berg, P., Kemperman, A., & Simons, M. (2022). Evaluating the Impact of Adaptive Personalized Goal Setting on Engagement Levels of Government Staff With a Gamified mHealth Tool: Results From a 2-Month Randomized Controlled Trial. *JMIR mHealth and uHealth*, *10*(3), e28801. https://doi.org/10.2196/28801

Nuijten, R. (2022). *Energize: Exploring the impact of gamification strategies on engagement with mobile health apps* (Doctoral dissertation) [Proefschrift.]. Industrial Engineering and Innovation Sciences. Eindhoven University of Technology.

Radha, M., Willemsen, M. C., Boerhof, M., & IJsselsteijn, W. A. (2016). Lifestyle recommendations for hypertension through rasch-based feasibility modeling.

*UMAP 2016 - Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 239–247. https://doi.org/10.1145/2930238.2930251

Rana, C., & Jain, S. K. (2012). Building a book recommender system using time based content filtering. *WSEAS Transactions on Computers*, *11*(2), 27–33.

Schäfer, H., & Willemsen, M. C. (2019). Rasch-based tailored goals for nutrition assistance systems. *International Conference on Intelligent User Interfaces, Proceedings IUI*, *Part F1476*, 18–29. https://doi.org/10.1145/3301275.3302298

Sick, J. (2010). Unidimensionality Equal item discrimination and error due to guessing. *JALT Testing & Evaluation SIG Newsletter.*, *14*(2), 23–29.

Sporrel, K., Nibbeling, N., Wang, S., Ettema, D., & Simons, M. (2021). Unraveling mobile health exercise interventions for adults: Scoping review on the implementations and designs of persuasive strategies. *JMIR mHealth and uHealth*, *9*(1). https://doi.org/10.2196/16282

Starke, A. (2014). With a little help from my friends investigating the effectiveness of Rasch-based energy recommendations with social endorsements With a little help from my friends : Investigating the effectiveness of Rasch- Rasch - based energy recommendations with soci.

Starke, A. (2019). *Supporting energy-efficient choices using Rasch-based recommender interfaces*.

Starke, A., Willemsen, M. C., & Snijders, C. C. (2020). Beyond "one-size-fits-all" platforms: Applying Campbell's paradigm to test personalized energy advice in the Netherlands. *Energy Research and Social Science*, *59*(March 2018), 101311. https://doi.org/10.1016/j.erss.2019.101311

Starke, A., Willemsen, M., & Snijders, C. (2017). Effective user interface designs to increase energy-efficient behavior in a rasch-based energy recommender system, 65–73. https://doi.org/10.1145/3109859.3109902

Starke, A., Willemsen, M. C., & Snijders, C. (2015). Saving energy in 1-D: Tailoring energy-saving advice using a Rasch-based energy recommender system. *CEUR Workshop Proceedings*, *1533*(October), 5–8.

World Health Organization. (2010). A healthy lifestyle - who recommendations.

World Health Organization. (2020). Physical activity.

Zhao, Z., Arya, A., Orji, R., & Chan, G. (2020). Effects of a personalized fitness recommender system using gamification and continuous player modeling: System design and long-term validation study. *JMIR Serious Games*, *8*(4), 1–27. https://doi.org/10.2196/19968

Zou, L., Song, J., Xia, L., Liu, W., Ding, Z., & Yin, D. (2019). Reinforcement learning to optimize long-term user engagement in recommender systems. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2810–2818. https://doi.org/10.1145/3292500.3330668

## Appendix A
## Power Analysis

To determine the total sample size for the calibration of Rasch scales, an estimate has to be made on the extent to which the estimate on the difficulty of an item has to be accurate. As healthy lifestyle choices have not been modeled to a unidimensional Rasch scale before, this estimate is hard to make. However, we aim to use a previous unpublished work by Starke et al., (n.d.), in which researchers provide estimates on the sample size relating to the required accuracy of the scale.

In order to be able to use this work for making an estimate of the required sample size however, we need to make some assumptions on the required accuracy of fit.

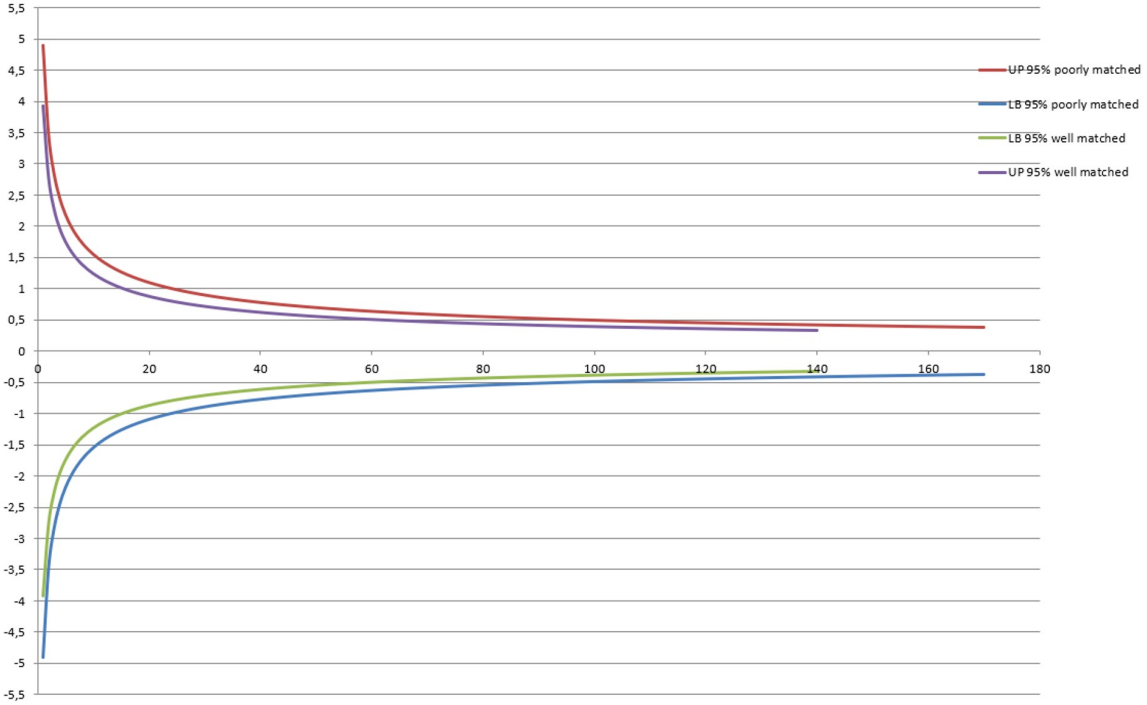


**Figure A1**

*Required sample size for Rasch scale calibration. Retrieved from Starke et al. (n.d.)*

As the nature of the study is mainly exploratory and the scale is used for recommendations, not precise measurements of attitude towards healthy activities, we do not require highly accurate fitting of both items and users on the Rasch scale. We therefore aim to construct and validate the scale based on an accuracy or +/- 1 logit, meaning we are aiming for medium-stake testing. Furthermore, we assume that we can create a medium- to well-matched scale of our unidimensional construct.

Based on these assumptions, based on literature of Linacre, 1994 and unpublished work of Starke, we can infer that we are aiming for a required sample size for Rasch scale calibration of about 60 participants.

The effect of difficulty on perceived effort is the most crucial factor in the power analysis as it is our experimental manipulation. We need to be sure that the difference in difficulty is strong enough to be perceived by the participants. In earlier work of

Starke et al. (2017) difficulty was manipulated between -1 and +1 logit and it's effect on perceived feasibility was measured. This led to a large effect size of d=0.6. We aim to have a similarly strong manipulation, and this would require about 100 participants to be able to show this effect size with a power of 0.9

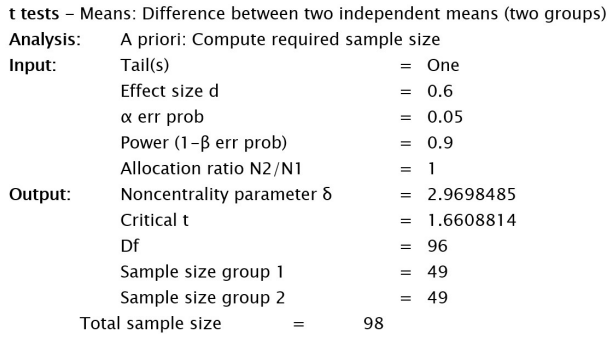


**Figure A2**

*Preliminary power analysis done before the start of the first study*

A sensitivity analysis shows that with around 150 participants we would also be able to show effect sizes of around 0.5

For SEM different guidelines are available. Some guidelines talk about 5-10 observations per item (we have about 15 items (3 constructs x 5 questions), so 75-150 participants should be sufficient. Earlier research in our lab with similar SEM models have shown that at least 100 participants are needed to build an adequate SEM model. In this study we aim for 150 participants.

Concluding we aim for 150 participants which allows us to show effect sizes of our manipulation of around 0.5, build an adequate SEM model and have sufficient data to validate and fit a Rasch scale of sufficient precision.

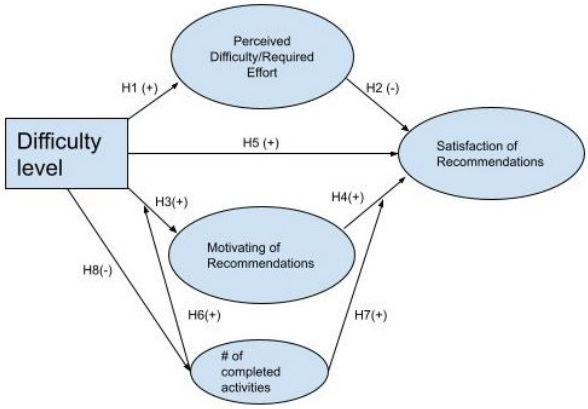


**Figure A3**

*Structural Equation Model for the study 2b including the expected directions*

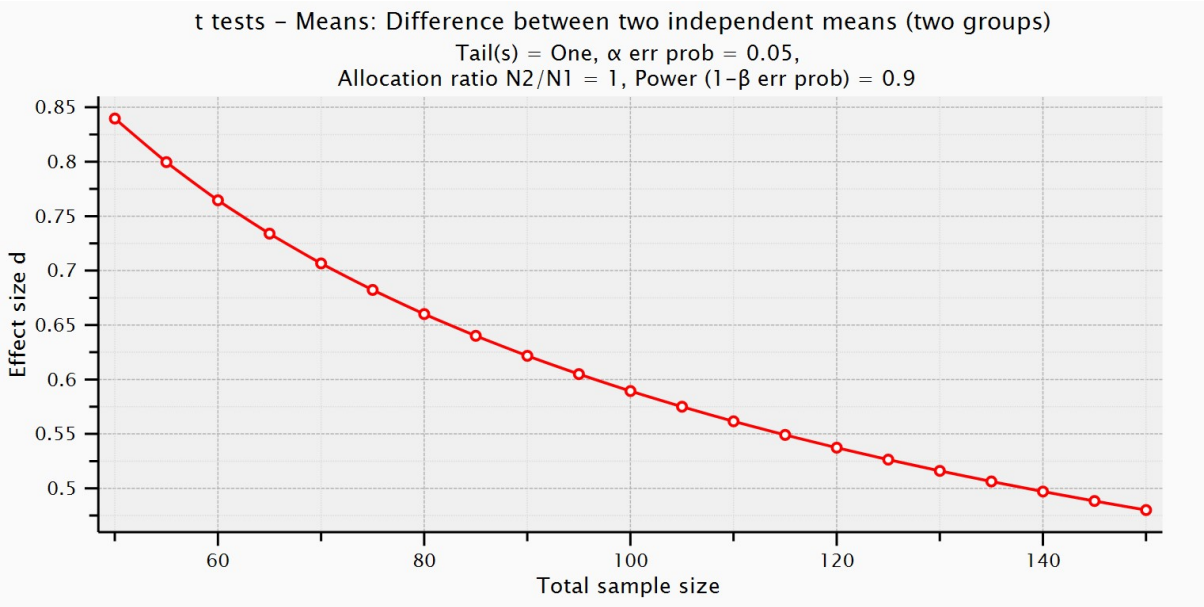


**Figure A4**

*Graph of the power analysis done before the start of the first study.*

## Appendix B
## Rasch Scale 1

| ENTRY NUMBER | TOTAL SCORE | TOTAL COUNT | MEASURE | MODEL S.E. | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PT-MEASURE CORR. | PT-MEASURE EXP. | EXACT MATCH OBS% | EXACT MATCH EXP% | Item | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 66 | 20 | .91 | .36 | 1.49 | 1.0 | 9.90 | 3.1 | A .28 | .42 | 46.2 | 62.2 | 38 Mooiste foto buiten | 0 |
| 53 | 429 | 138 | 1.50 | .22 | 1.66 | 1.5 | 9.81 | 5.6 | B .08 | .29 | 88.2 | 90.2 | 218 15min rollen (skeeleren, schaatsen etc.) | 0 |
| 29 | 42 | 12 | -.30 | .46 | 1.85 | 1.9 | 3.94 | 2.3 | C .45 | .58 | 28.6 | 39.4 | 166 Potje trefbal | 0 |
| 54 | 112 | 34 | .59 | .36 | 1.18 | .5 | 2.03 | 1.4 | D .43 | .52 | 66.7 | 75.7 | 219 Foto van een gezond winkelwagentje | 0 |
| 48 | 89 | 28 | .45 | .43 | .82 | -.1 | 1.79 | .9 | E .43 | .47 | 80.0 | 85.5 | 213 Maak een gezonde sandwich | 0 |
| 37 | 136 | 38 | 1.05 | .23 | 1.51 | 1.5 | 1.67 | 1.7 | F .31 | .48 | 45.5 | 52.5 | 202 3km fietsen | 0 |
| 24 | 50 | 14 | .41 | .34 | 1.39 | 1.0 | 1.64 | 1.1 | G .30 | .47 | 50.0 | 54.0 | 141 Fitpic boodschappen lopend | 0 |
| 18 | 31 | 9 | .56 | .39 | 1.34 | .7 | 1.55 | .8 | H .23 | .36 | 37.5 | 54.1 | 88 30min Aerobics/Gym sport | 0 |
| 25 | 25 | 7 | -3.30 | .48 | 1.06 | .3 | 1.48 | .8 | I .53 | .60 | .0 | 35.4 | 143 Stuk fietsen | 0 |
| 39 | 88 | 25 | 1.49 | .25 | 1.36 | 1.1 | 1.47 | .9 | J .38 | .47 | 55.0 | 47.5 | 204 8km fietsen | 0 |
| 5 | 49 | 14 | .43 | .36 | 1.36 | .8 | 1.41 | .8 | K .27 | .43 | 50.0 | 58.7 | 32 Fitpic bovenaan de trap in Atlas | 0 |
| 13 | 419 | 114 | .30 | .15 | 1.35 | 1.8 | 1.39 | 2.1 | L .36 | .58 | 34.3 | 60.6 | 60 15 reps deskercise | 0 |
| 14 | 17 | 5 | -3.25 | .66 | 1.36 | .8 | 1.37 | .7 | M .00 | .40 | .0 | 40.1 | 67 Potje dammen | 0 |
| 38 | 393 | 115 | 1.18 | .14 | 1.33 | 1.7 | 1.29 | 1.0 | N .50 | .58 | 41.8 | 50.7 | 203 4km fietsen | 0 |
| 46 | 71 | 22 | .83 | .36 | 1.27 | .6 | .65 | .0 | O .28 | .28 | 70.6 | 78.4 | 211 Screenshot van een app die 15k stappen laat zien | 0 |
| 51 | 124 | 34 | -2.59 | .27 | 1.24 | 1.1 | 1.21 | .6 | P .64 | .69 | 44.0 | 48.8 | 216 Maak een wandeling met de hond | 0 |
| 10 | 378 | 114 | .87 | .16 | 1.19 | .8 | .78 | -.5 | Q .45 | .47 | 66.7 | 69.2 | 49 25min willekeurige balsport | 0 |
| 12 | 420 | 123 | 1.30 | .15 | 1.17 | 1.0 | 1.17 | .8 | R .54 | .60 | 31.4 | 49.1 | 59 30min sporten | 0 |
| 22 | 43 | 12 | -.68 | .63 | 1.15 | .5 | 1.03 | .2 | S .75 | .77 | 62.5 | 61.0 | 122 Gezonde maaltijd | 0 |
| 52 | 135 | 37 | -2.92 | .28 | 1.07 | .4 | .81 | .3 | T .70 | .71 | 45.8 | 50.8 | 217 Maak een selfie in het park of in het bos | 0 |
| 49 | 151 | 47 | .56 | .25 | 1.04 | .2 | .27 | -.3 | U .39 | .38 | 82.9 | 79.3 | 214 Maak een gezonde salade | 0 |
| 6 | 44 | 13 | .56 | .36 | .86 | .0 | 1.02 | .4 | V .36 | .35 | 58.3 | 60.3 | 33 Fitpic op de fiets met collega's | 0 |
| 1 | 373 | 114 | .86 | .15 | .99 | .1 | .60 | -.6 | W .46 | .45 | 76.8 | 77.7 | 2 15min zwemmen | 0 |
| 47 | 78 | 24 | .88 | .37 | .98 | .1 | .75 | .3 | X .47 | .48 | 71.4 | 77.2 | 212 15 sit-ups | 0 |
| 42 | 114 | 34 | 1.68 | .28 | .98 | .1 | .76 | -.4 | Y .46 | .39 | 75.9 | 64.0 | 207 1km wandelen | 0 |
| 44 | 78 | 22 | 1.73 | .25 | .95 | .0 | .69 | -.4 | Z .47 | .40 | 59.1 | 53.9 | 209 3km wandelen | 0 |
| 43 | 426 | 122 | 1.00 | .13 | .94 | -.3 | .79 | -.8 | z .65 | .63 | 49.3 | 45.2 | 208 2km wandelen | 0 |
| 19 | 398 | 114 | .58 | .13 | .94 | -.3 | .83 | -.5 | y .56 | .54 | 59.6 | 60.7 | 98 25min fitness, yoga of dance | 0 |
| 32 | 86 | 25 | -3.23 | .34 | .88 | -.3 | .94 | -.1 | x .84 | .83 | 66.7 | 57.8 | 173 Korte fietstocht | 0 |
| 28 | 386 | 114 | .88 | .15 | .92 | -.3 | .74 | -.9 | w .54 | .50 | 70.7 | 66.3 | 159 Work Walk | 0 |
| 11 | 73 | 21 | 1.62 | .32 | .77 | -.5 | .90 | .0 | v .49 | .45 | 50.0 | 58.1 | 58 Gezond tussendoortje | 0 |
| 50 | 651 | 198 | .63 | .12 | .90 | -.5 | .47 | -1.4 | u .45 | .42 | 69.0 | 72.8 | 215 Eet een stuk fruit | 0 |
| 30 | 18 | 5 | -3.25 | .66 | .88 | -.1 | .76 | -.2 | t .60 | .52 | 50.0 | 40.1 | 170 Drink een kop thee | 0 |
| 45 | 793 | 229 | -2.44 | .18 | .86 | -.8 | .79 | -1.1 | s .91 | .90 | 68.0 | 60.9 | 210 Doe deze yoga-houding: | 0 |
| 4 | 73 | 21 | 1.62 | .32 | .83 | -.3 | .73 | -.4 | r .50 | .45 | 50.0 | 58.1 | 27 Gezonde lunch | 0 |
| 33 | 393 | 114 | .73 | .14 | .83 | -1.0 | .67 | -1.4 | q .60 | .52 | 64.6 | 62.0 | 174 15min fietsen | 0 |
| 3 | 47 | 14 | .60 | .33 | .76 | -.4 | .39 | -.1 | p .47 | .38 | 58.3 | 55.5 | 16 Fitpic in sportkleren in een park/in het bos | 0 |
| 26 | 378 | 114 | 1.06 | .16 | .76 | -1.2 | .67 | -.9 | o .54 | .47 | 69.7 | 68.7 | 144 15min hardlopen | 0 |
| 8 | 49 | 14 | .30 | .29 | .73 | -.7 | .41 | -.2 | n .53 | .43 | 58.3 | 50.8 | 44 Fitpic fysieke activiteit in een groep | 0 |
| 34 | 35 | 10 | -3.47 | .48 | .73 | -.4 | .59 | -.3 | m .74 | .65 | 60.0 | 35.2 | 183 Activiteit die je laat zweten | 0 |
| 17 | 49 | 14 | .48 | .34 | .71 | -.6 | .46 | -.6 | l .59 | .44 | 66.7 | 53.0 | 87 Fitpic op de fiets naar sportschool | 0 |
| 41 | 30 | 9 | .64 | .39 | .71 | -.3 | .32 | -.2 | k .38 | .31 | 50.0 | 49.5 | 206 500m wandelen | 0 |
| 40 | 1112 | 343 | .75 | .11 | .69 | -2.4 | .42 | -1.1 | j .53 | .51 | 64.2 | 63.3 | 205 250m wandelen | 0 |
| 35 | 413 | 123 | 1.22 | .14 | .68 | -1.8 | .58 | -1.5 | i .62 | .55 | 58.6 | 52.4 | 200 1km fietsen | 0 |
| 2 | 49 | 14 | .15 | .37 | .68 | -1.1 | .51 | -.8 | h .64 | .47 | 58.3 | 48.2 | 7 Fitpic boodschappen per fiets | 0 |
| 31 | 141 | 39 | -1.22 | .29 | .68 | -1.4 | .63 | -1.2 | g .80 | .74 | 59.3 | 55.7 | 172 Korte wandeling | 0 |
| 36 | 24 | 7 | 1.08 | .42 | .62 | -.6 | .30 | -.2 | f .53 | .44 | 66.7 | 63.3 | 201 2km fietsen | 0 |
| 20 | 15 | 4 | -3.39 | .73 | .60 | -.8 | .61 | -.7 | e .72 | .56 | 33.3 | 33.3 | 104 20 push-ups | 0 |
| 21 | 134 | 38 | -1.53 | .33 | .55 | -1.1 | .56 | -.9 | d .80 | .76 | 70.4 | 68.1 | 105 10min hardlopen | 0 |
| 27 | 717 | 220 | -1.22 | .19 | .46 | -3.5 | .53 | -2.6 | c .83 | .78 | 68.7 | 57.4 | 149 Bezwete foto met vrienden | 0 |
| 15 | 77 | 24 | .60 | .39 | .47 | -.8 | .22 | -.4 | b .44 | .38 | 85.0 | 79.9 | 68 Foto van gezonde maaltijd | 0 |
| 16 | 17 | 5 | .71 | .73 | .35 | -.9 | .22 | -.4 | a .60 | .51 | 75.0 | 67.0 | 75 30 lunges | 0 |
| MEAN | 202.7 | 60.1 | .00 | .32 | .98 | -.1 | 1.26 | .1 | | | 57.1 | 58.8 | | |
| S.D. | 232.9 | 70.4 | 1.57 | .16 | .32 | 1.0 | 1.82 | 1.3 | | | 17.8 | 12.8 | | |

**Figure B1**

*All items in the Rasch model from Study 1, ordered to fit (least fitting = highest). Items with a infit mean-square fit statistics (MNSQ) of higher than 1.4 are unproductive for the construction of the model*