

## MASTER

### Supporting the selection of health improvement measures by means of a recommender system

Elsten, T.J.M.M.

*Award date:*  
2012

[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.

Eindhoven, August 2012

**Supporting the selection of  
health improvement measures  
by means of a  
recommender system.**

By: T.J.M.M. Elsten

Identity number: 0646822

in partial fulfillment of the requirements for the degree of

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

Supervisors:

Dr. ir. M.C. Willemsen (Martijn) - Eindhoven University of Technology, department of IE&IS.

Dr. W.A. IJsselstein (Wijnand) - Eindhoven University of Technology, department of IE&IS.

Dr. C. Veenhof (Cindy) – NIVEL Research Institute, Utrecht.



Eindhoven University of Technology: School of Innovation Sciences.  
Master Human-Technology Interaction. (HTI)

Keywords: recommender system, health, intervention, lifestyle, tailored advice



## Abstract

Modern day society is subjected to several threats on a daily basis, the threat that we are addressing in this work is visible on the streets every day, has affected 40% of the Dutch population so far and is still gaining ground. The creeping danger we are talking about is of course obesity. The worst thing about overweight and obesity is probably that it is a direct consequence of our own unhealthy lifestyle, in other words; we only have ourselves to blame for it. The good news on the other hand is that all of our western prosperity has also provided us with huge amounts of information on how to change our unhealthy lifestyles and live healthier lives.

The large quantities available and unstructured nature of this information makes it particularly difficult to find and select which information is useful and personally relevant. This research therefore introduced a recommender system to the situation. Recommender systems are software systems that are designed to facilitate the decision process of its user by filtering large quantities of data and recommending only a few personally relevant items. In order to allow for the use of a recommender system in this situation we first gathered and structured as much information on healthy living as possible. This information was formulated as clear and small healthy assignment and inserted into the recommender.

The goal of this research was to investigate whether recommender systems could be of added value to the domain of healthcare and help people to change their lifestyle. In order to do so we built two different versions of a lifestyle recommender system that was later titled the *Lifestyle Coach*. The first recommender system offered the user the possibility to enter preferences regarding relevant attributes into the system following which the system provided its user with tailored recommendations. The second system on the other hand provided its user with random (non-tailored) recommendations.

We designed two different between-subject studies in order to test the effect of the above mentioned recommender systems on the lifestyle of the user. The first study was based on a single interaction session with the Lifestyle Coach and measured the effect on participants' lifestyle one week after. The second study consisted of three weekly cycles and thereby focused more on the temporal effect that the Lifestyle Coach had on participants' lifestyles.

We measured user experience in both studies prior to, and after interaction with the Lifestyle Coach. Furthermore, we recorded the number of items that participants in this experiment selected and removed while interacting with the Lifestyle Coach. We also measured how many of the selected items were actually being executed by our participants.

Analysis of our results showed that our participants used our recommender systems very differently. We found evidence that some participants did benefit from the tailored recommendations. However, the observed behavior of our participants mainly suggested that they did not prefer to make decision regarding their lifestyle in an explicit way as was facilitated by our attribute-based recommender. Instead it seems that more implicit recommendations (such as those by collaborative recommenders), simply based on their browsing behavior would be more desirable in this case.

## Preface

This thesis marks the end of what has been not only an incredibly interesting and challenging project but also the end of my career as a student. A career that has been responsible for shaping me both professionally as well as socially into the person I am today. Starting of as a physicist - believing only what I could translate into numbers and formulas – into a technical psychologist that is now (better) able to understand and study what happens outside the world of physics, where technology actually meets humans.

During my master study in particular I developed an (unhealthy) appetite for engaging in every project, activity or adventure I could get my hands on; the subject of this thesis being the perfect example. Ever since the graduation project of my physics degree I knew that I wanted to apply my knowledge of technology to improving peoples' health and performance. Learning of something called recommender systems therefore sparked my interest and made me wonder whether this could be the Holy Grail in improving our population's general health<sup>1</sup>. I learned from personal experience that plenty of people at least try to live healthy(er) but often do not know to do this or how to combine this with their daily routine. I figured that if I had to spend half a year of research for my master thesis, it might as well be on this subject.

I realize that not every student has had the luxury of being able to choose their own research topic as detailed as I have done and for that I am very grateful. In particular I thank Roland Friele at the NIVEL institute for this, without involvement of the NIVEL institute this study would have never taken place (perhaps it would have, but not by my hand). Furthermore I thank Martijn Willemsen and Cindy Veenhof for supporting this research from the very beginning. I experienced their supervisory role as being very personal and pleasant, without their feedback this thesis would definitely not have been as good (off course this is for the reader to judge) as it is now.

Besides my supervisors, I also thank my fellow graduation friends with whom I shared an office for six months. As the only man in the room I feared to be slowly transformed into a soft metro-man, nonetheless the opposite is true and I remain manly as ever (also for the reader to judge off course).

I particularly thank my lovely girlfriend Marlies, who was not only unfortunate enough to have to share the graduation room with me during the day but also had to live with me the rest of the time in our apartment. I assume that my graduation stress and the many evening hours that I invested in my thesis have at least to some extent affected her user experience<sup>2</sup> with me, however strictly non-empirical evidence tells me that her choice satisfaction has remained unchanged, at least mine has.

Before I conclude my mini-essay on my career as a student I finally thank my family. I do not specifically thank them for their support during the final half year of my study (although it has most certainly been there), I thank them in general for their facilitation of my

---

<sup>1</sup> This study has shown me that most likely there is no such thing as a Holy Grail, nonetheless I remain to have a keen interest in anything that comes close.

<sup>2</sup> For more information on the subject of user experience, read the rest of this thesis.

development into the person I am now. Thanks to them I have always been able to choose the direction into which I wanted to go, which is invaluable to any young researcher and any person in general.

What lies ahead of me now is either an exciting career as a young professional or an indefinite period of unemployment. Nonetheless I will continue my efforts in improving my own health and that of those around me.



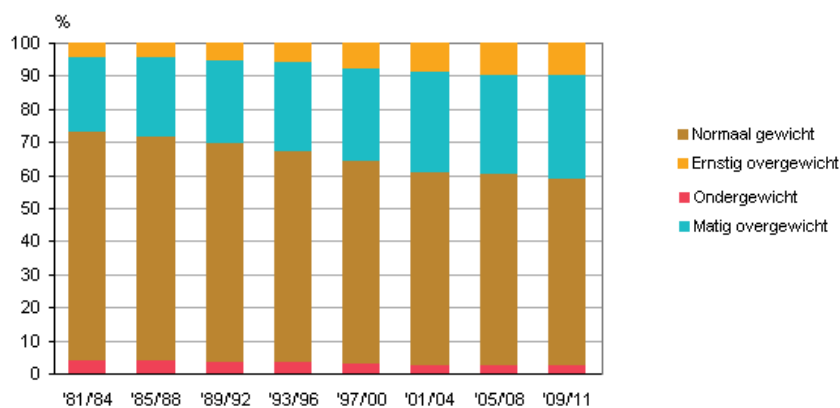
# Contents

1. Introduction.....	1
1.1 Behavioral theories.....	2
1.2 Tailored health intervention studies .....	3
1.3 Recommender systems .....	6
1.4 Healthy decision making .....	8
1.5 This research .....	10
1.6 Hypotheses .....	11
2. Method.....	14
2.1 Design.....	14
2.2 Participants.....	14
2.3 Procedure.....	14
2.3.1 Study 1 .....	15
2.3.2 Study 2.....	16
2.4 Materials.....	17
2.4.1 Database with healthy measures .....	17
2.4.2 Software.....	20
2.5 Measures.....	22
2.5.1 Study 1 .....	23
2.5.2 Study 2.....	24
2.6 Analysis.....	25
3. Results.....	26
3.1 Study 1.....	26
3.1.1 The effect of tailoring .....	26
3.1.2 The analysis of realistic choice behavior.....	31
3.1.3 Discussion.....	34
3.2 Study 2.....	35
3.2.1 The effect of tailoring .....	36
3.2.2 Discussion.....	39
3.2.3 Temporal effects of the Lifestyle Coach .....	39
4. General discussion .....	42
4.1 Main findings .....	42
4.2 Implications.....	45
4.3 Conclusion & future research.....	47

5. Acknowledgements.....	49
6. References.....	50
7. Appendices.....	53
7.1 UX questionnaire.....	53
7.2 Fruit, vegetable and fat intake questionnaire .....	54
7.3 SQUASH questionnaire .....	56
7.4 Screenshots of the Lifestyle Coach software .....	57

## 1. Introduction

Worrisome reports regarding health research and warnings of an increasingly overweight population are more and more prevalent in current day media than ever before. Unfortunately this is not without reason. A longitudinal study (Figure 1) by the Statistics Netherlands agency (CBS) over the past 30 years shows that the number of people with overweight has increased by more than 10% resulting in 40% of the Dutch population being overweight. At the same time, the number of adults with extreme overweight has even doubled. Most likely this is a consequence of the also continuously increasing number of overweight kids over the past 30 years.



**Figure 1 – Change in the distribution of underweight (ondergewicht), healthy weight (normal gewicht), overweight (overgewicht) and obese (extrem overgewicht) people over the Dutch population. (Source: CBS)**

In parallel with the increasing number of overweight people, the availability of information to battle this unhealthy trend has also been increasing over the past decades. This increase is not merely due to a growing interest in health issues by the scientific community. Media such as the Internet have the capability of spreading information on health issues among large audiences and penetrate into almost all layers of society. The enormous reach and ease of accessibility of the Internet and all the information it contains has led many experts, self-proclaimed experts and even complete novices to ‘contribute’ to the world wide web, sometimes resulting in an incomprehensible amount of information on almost every subject. There is plenty of information available on how to live a healthy life or how to change one’s lifestyle. The excess or overload of information leads to increased difficulty in finding information that is personally relevant. (Berland et al., 2001) This information overload is typically a challenge that can be addressed by using eHealth interventions and tailored health solutions in particular.

The term eHealth generally refers to efforts in which expertise on health and ICT are combined in a joined effort to produce intelligent solutions to health problems (Eysenbach, 2001). Since the eHealth domain is still in its infancy, the full extent to which technology can enrich the field of health and medicine is not yet known. Even so, a reasonable body of literature has been developed over the past 2 decades, focusing on the use of technology and technological interventions to change or modify unhealthy behavior. Most efforts that make

use of technological interventions work through theories that have emerged in social psychology. Long before the use of technology in behavioral research, these theories have proven their value in understanding, explaining and predicting human behavior to a certain extent.

### 1.1 Behavioral theories

In order to better understand and value the literature on intervention studies in the health domain, the theories that are most widely used in health improvement studies are discussed briefly below.

#### *Social Cognitive Theory*

Research by Miller & Dollard (1941) identified social cognitive theory (SCT) as the process of learning through observing, otherwise referred to as vicarious learning. According to their theory, people can learn new behavior through observing, cognitively processing this observation and then repeating it. A few decades later, Albert Bandura (1998) applied this theory of social learning to the domain of health promotion and argued that measures should be aimed mainly at changing collective social attitude towards healthy behavior rather than focusing on the individual.



Figure 2 - Illustration of the stages of change in the transtheoretical model by Prochaska. (TTM)

#### *Transtheoretical model*

The transtheoretical model (TTM) has a fairly different approach and involves 5 stages of change that describe how an individual's attitude and behavior can change from being reluctant to perform a certain behavior into actually engaging in that particular behavior. (Prochaska & DiClemente, 1982)(Figure 2) When using the TTM, one steadily attempts to change the views that the person under study holds in regard to the pros and cons of the specified behavior. When designing lifestyle interventions, effectiveness of intervention programs will increase when measures are matched to a person's stage of change; this is why many intervention programs nowadays make use of this principle.

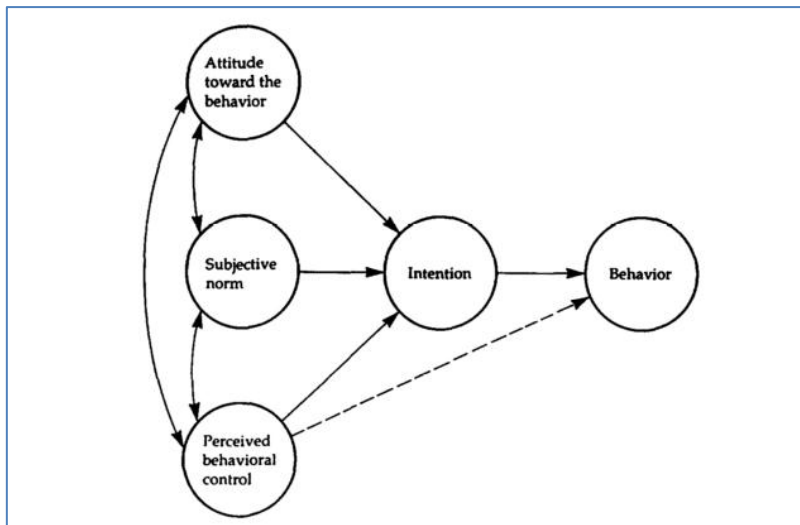


Figure 3 – Schematic overview of the theory of planned behavior by Ajzen.

### *Theory of planned behavior (TPB)*

Perhaps one of the most well known theories in behavioral science is Icek Ajzen’s Theory of planned behavior (TPB). Ajzen’s work (Ajzen, 1991) provides a model that predicts one’s behavioral intentions by taking attitudes toward the behavior, subjective norms with respect to the behavior, and perceived control over the behavior into account. Figure 3 shows how these variables relate to each other and together predict behavioural intentions. A literature review by Godin & Kok (1996) shows that the model performs very well in predicting behavioral intentions in the health domain. Attitude and perceived behavioral control appeared to be the main predictors of behavior. The TPB aims to predict behavior whereas the TTM explains how behaviour is changing or can be changed.

### **1.2 Tailored health intervention studies**

Ever since the birth of computerized tailored health interventions, this field of research has been subjected to continuous development and redesign resulting in three distinguishable generations to date. First generation interventions refer to leaflets, flyers or newsletters that are specifically designed to address a certain target group and are distributed accordingly. In this case experts use the computerized tailoring technology to generate printed tailored materials that are subsequently distributed amongst users. The second generation of computerized tailored health interventions on the other hand allowed the user to interact with the technology themselves by means of websites and handheld computers to interact with it themselves. This generation benefitted of the increasing availability and functionality of computer technology. The third and current generation takes the possibilities that are offered by current-day technology even further by using the omnipresence of the Internet and allowing for temporal and spatial triggering of information.

The technology used in this research has its roots somewhere in between the second and third generation; the literature review below will therefore mostly discuss work of a similar nature.

A study by Kreuter, Oswald, Bull, & Clark (2000) was among the first to provide proof for the benefit that tailoring intervention materials can have on personal health. In this study, the authors tested printed tailored intervention material that focused solely on weight loss and compared this with generic intervention material. By determining a goodness-of-fit measure for each applied intervention, Kreuter et al. (2000) were able to determine if non-tailored materials were maybe by chance still personally relevant to the participant. The goodness-of-fit variable allowed them to answer the question whether handing out random intervention material to individuals would not lead to similar effect sizes as applying tailored materials. Their results show that this is not the case, tailored materials do have an edge on generic health materials even though they can be personally relevant simply by chance. Latter research is an example of a first generation health intervention strategy.

About a decade ago, Brug, Campbell, & van Assema (1999) published their literature review that assessed 8 studies in which computer tailored health information was combined with behavioral change models in order to change people's nutritional intake. This is an example of a second-generation health intervention. The most important determinants for effective interventions that are reported in this literature review are threefold: (1) motivators and reinforcers should be personally relevant (2) personalization should be applied to self-assessment and self-evaluation (i.e. participants should be able to monitor and adjust the progress of their intervention) (3) people should be allowed to actively participate in the given intervention (Contento, Balch, Bronner, Lytle, Maloney, Olson & Swadener (1995) cited in: Brug et al., (1999)). Tailoring the information therefore increases the effectiveness of a certain intervention. The computer-tailored intervention studies that have been reported in the literature review use variables such as socio-demographics, health status, attitudes and perceptions towards a certain behavior to individualize the feedback they provide. Since most of the reviewed studies used a stages-of-change based model, tailoring was done based on the stage a person is in. Following this, users were taken through the consecutive stages until they reached the end goal; their desired behavior. Although results were not yet conclusive, they did point to the conclusion that tailored materials were evaluated better than non-tailored ones. Personalization of intervention materials leads to the information being read, remembered and experienced as more personally relevant. (Brug et al., 1999)

In later work Brug, Oenema, & Campbell (2003) discuss in more detail the future of tailored health education and its possible edge on generic health education materials. Generic health information is generally speaking designed in one of two ways: information appealing to one or a few beliefs or attitudes *or* simply providing the public with as much information as possible. In the latter case, responsibility for finding relevant information is in the hands of the population itself. Both of these strategies differ in target-group size and impact due to the personal relevance and availability of the information. Tailoring the education materials can combine these two approaches by helping people find the information that is personally relevant to them from the large haystack of information available. Although some research suggests that tailored information would only produce its desired effect among higher educated people and women (Brug & Van Assema (2000) cited in Brug et al. (2003)), Brug et

al., (2003) pin this on the self-selection criterion that is used in most studies on tailored health interventions and present research of their own that shows otherwise. Many current intervention studies recruit through advertisements and therefore have limited power over the people that respond to them, often resulting in the overrepresentation of higher educated individuals and women. For example Brug, Steenhuis, van Assema & de Vries (1996) (cited in Brug et al. (2003)) showed positive health effects also for a largely male population in a workplace setting. On top of this, (Brug et al., 2003) mention that tailored materials may also reduce cost, accessibility and availability of health information in general.

An experimental study into the short-term effects of computer-tailored web-based nutrition information by Oenema, Tan, & Brug (2005) showed that tailored information indeed has a significant effect on the determinants of fat, fruit and vegetable intake. Effects on actual intake however are modest or absent suggesting that the intervention in this study mainly changed the attitudes of the participants and not their actual behavior. Evidence for behavioral change was only found in subgroups (people in a risk group) and only on fruit and vegetable intake whereas changes on fat-intake were absent. Manipulations were based on the stages-of-change model and questioned the participants step-by-step in order to take them through the process. Depending on their current situation and personal characteristics, the system provided them with future directions for healthy lifestyle behavior.

A literature review by Norman et al., (2007) studied 49 second generation eHealth intervention studies and reflected on the strengths and weaknesses of these studies. They also showed that there still is a lot of potential to be uncovered in this field. Close to half of the studies reviewed compared eHealth technology against a non-technology control group. Effect-sizes were usually found to be in the small to medium size range favoring technology. Common denominators for the success of the eHealth interventions were related to frequency of use and goal setting. Frequency of use – the digital equivalent of the dose in medicine – positively correlates with the effectiveness of behavioral interventions. An advantage of eHealth interventions is that this dose can be measured very accurately by means of page views and log files. A second strength of eHealth interventions was the ability to interactively change people's goals. Large goals can be broken down into smaller ones, making certain behavior more likely to change.

To conclude their review Norman et al., (2007) mention the potential of possible third-generation interventions that utilize mobile and handheld devices that are permanently connected to the Internet. These connected multi-sensory devices allow for adding new functionality such as temporal or spatial triggering of tailored messages, possibly increasing the effectiveness of eHealth interventions even more.

An example of a third generation intervention study is that of Kaptein, De Ruyter, Markopoulos, & Aarts (2012) that studied the effect of tailored text (SMS) messages on snacking behavior. Tailoring was done by employing different social influencing strategies. The type of influence strategy was based on participants score on the susceptibility to persuasion scale (STPC). The latter is a test determining which social influence strategy will be most effective. Participant in this research were subjected to tailored messages, random messages or contra-tailored messages. Results show that participants in the tailored treatment reduced their snacking more than participants in the random or contra-tailored treatment.

With the rise of these potent technological aids it is even possible to target multiple (possibly related) behaviors at the same time. A study by Vandelanotte, De Bourdeaudhuij, Sallis, Spittaels, & Brug (2005) investigated the targeting of multiple behaviors simultaneously by means of computer tailored interventions. They divided their sample of participants into a simultaneous, a no-treatment and two sequential groups. The simultaneous group was exposed to tailored messages focusing both on fat intake as well as exercise behavior. Sequential groups received computer tailored information on fat intake and physical exercise in a sequential manner in two different orders. Results showed that tailored materials outperformed the control group in all cases; however there was no significant difference between the sequential and the simultaneous treatment for the full participant sample. Nonetheless, reported effect sizes were larger for ‘unhealthy’ participants compared to those that were considered ‘healthy’. In other words, people with a higher need for a lifestyle intervention benefitted more when multiple behaviors were targeted at the same time.

The literature above shows that serious efforts are being undertaken to find appropriate tools and theories that can help to improve general health. Although the most recent studies increasingly utilize interactive technologies to deliver tailored health interventions to its users, none so far has focused on the information delivery issue. Strategies of past and current day interventions mainly target motivational aspects in the people under study and focus less on the actual behavior that has to be performed in order to get to a better health. Some intervention strategies do focus on helping people in making an exercise plan, but do this by setting large abstract goals. Other intervention strategies offer its user small and clear goals or assignments but do not tailor these to its user, thereby decreasing their effectiveness. Current day strategies have so far not bothered with using the enormous amount of information that is already out there. Our idea is to focus less on the motivational aspect of behavioral change but instead bring large and abstract goals within reach by dividing them into small and clear cut pieces. These pieces are based on the large amount of information available on healthy living. By structuring this information, we are able to present the right piece to the right person. Handling such large quantities of data and finding items that are personally relevant is something that recommender systems are good at. Therefore we aim to investigate the applicability of recommender systems in healthcare.

### **1.3 Recommender systems**

Recommender systems are interactive software applications that support its user in making decisions from a large set of items. Large quantities of data are usually very hard for individuals to comprehend or handle when it comes to decision-making. Therefore recommenders filter this data for personally relevant items first, helping the user in the decision making process. The input for these filters is always supplied by the user, recommender system elicit this input either implicitly or explicitly. Implicit user input mechanisms do not bother the user with questions but rather analyze browsing, buying or clicking behavior whereas explicit systems ask users to rank, score or select items or characteristics they prefer. Filtering in a recommender system is usually done in one of three



ways, either by collaborative filtering, content-based filtering or a hybrid form that combines the first two (Adomavicius & Tuzhilin, 2005).

Collaborative filtering mechanisms learn from previous user behavior and do not need to know the content of the items they are actually filtering. By analyzing (choice) behavior such as clicks, choices and ratings of previous users they are able to predict what user is most similar to you and thereby what your next choice might be. People that frequently visit web stores have most likely been prompted at one time or another with the text: “Users that chose this item also chose this”. Common examples of websites employing collaborative filtering mechanisms are Last.fm, Amazon and bol.com. Not having to understand the complex nature of the product they are providing recommendations on is a major advantage of collaborative recommender systems. On the other hand, these systems do require a large amount of user behavior first in order to be able to make sensible recommendations. Furthermore, since collaborative filtering mechanisms often need to handle huge amounts of data they need a lot of computing power, however this can be costly financially as well as in processing power.

Content-based filtering on the other hand does rely on knowing and understanding the items it is providing recommendations on. An implicit content recommender does not look for similarities in users, instead it analyses the characteristics of an item and checks other products that might be similar or related to it. An explicit recommender on the other hand lets its user indicate their preference following which it matches this to the characteristics of the items to be chosen by means of a particular decision making algorithm. The fact that content-based recommenders require knowledge of the items they are recommending on is at the same time a disadvantage since this is in some cases very difficult and rather abstract. (i.e. what are the underlying properties of healthy lifestyle advice?) Additionally, a content-based recommender requires more work to include new items in the dataset. The advantage of content-based recommenders lies in the possibility to mimic human decision making strategies more closely and thereby facilitating an easier decision making process.

Content-based and collaborative systems can be combined into a hybrid system in one of the following ways: (1) combining the predictions from both models, (2) importing content-based characteristics into a collaborative system, (3) importing collaborative characteristics into a content-based system or (4) generate a general unified model that incorporates characteristics of both models. (Adomavicius & Tuzhilin, 2005)

Even though none of the three types of recommender systems has a clear a priori edge over the other in the healthcare domain, we chose to use a content-based recommender system in this research. Lifestyle changing decisions typically involve making trade-offs between competing characteristics of these suggested lifestyle changes such as: *How much does that cost me? How long will it take me to do that? How many calories will I burn doing that?* Since content-based recommenders rely on ‘product’ characteristics to base their filtering on, we anticipate a content-based recommender to be most effective to facilitate this decision making process. Furthermore, since no database with rated items exists yet, building a collaborative filtering algorithm would take too much time to develop and train for a master thesis.

Xiao & Benbasat (2007) suggest that a one size fits all strategy for recommender system algorithms might be sub-optimal, instead the algorithm at hand should match the user’s

individual decision strategy. Research by Knijnenburg, Willemsen, & Reijmer (2011) confirmed the suggestion that different types of decision makers call for different recommender systems in order to optimize user satisfaction. In their research, they investigated the effects of five different types of recommender systems on both novices as well as expert users in the domain of energy saving measures. Results showed that novice users benefitted most from a system that presented them with the most popular recommendations whereas expert users achieved higher levels of satisfaction by using a system that incorporated a model from decision-making theory called the multi-attribute utility model.

The multi-attribute utility theory (MAUT) is a theory on how to determine the utility of the outcome of a decision when multiple trade-offs between attributes have to be made. The MAUT theory uses different weights for the different characteristics of a certain item or product to calculate an overall utility score. When implemented in a recommender system, weights ( $w$ ) can be set according to personal preference following which the utility ( $U$ ) of every item or product can simply be calculated by using different attributes<sup>3</sup> ( $a$ ) of this item or product according to the equation below:

$$U = \sum w \cdot a$$

The MAUT model can for instance be applied to the process of deciding what meal to eat for dinner tonight. Table 1 shows three different meals including their scores on different attributes: preparation time, healthiness of a meal (defined on a 1-10 scale) and price.

**Table 1 - Values for different attributes for 3 different dinner meals.**

Meal	Prep. time (min)	Healthiness	Price (€)
Pasta bolognese	7	6	5
Fries with ketchup	5	3	3
Rice, fish & steamed vegetable	10	8	7

Applying the MAUT model means that we indicate our preference (for instance on a 5-point scale) by setting weights for all of the attributes of the meals. Following this the MAUT model will indicate the utility of each item and order them accordingly.

A low interest in preparation time ( $w_{\text{PrepTime}}=1$ ), a high preference for healthiness ( $w_{\text{Healthiness}}=5$ ) and a medium interest in price ( $w_{\text{Price}}=3$ ) will according to the MAUT model result in the Rice as the top choice<sup>4</sup>, followed by the pasta and the fries in last place.

### **1.4 Healthy decision making**

Parallel to the energy savings domain, people interested in improving their personal health also come in different sorts (ranging from experts to novices) and have different needs

<sup>3</sup> We simplified the model by omitting utility values for separate attributes.

<sup>4</sup>  $U_{\text{Pasta}} = 1 \times 7 + 5 \times 6 + 3 \times 5 = 52$

$U_{\text{Fries}} = 1 \times 5 + 5 \times 3 + 3 \times 3 = 29$

$U_{\text{Rice}} = 1 \times 10 + 8 \times 8 + 3 \times 7 = 95$

(maintaining a good health or lose weight). Decisions regarding exercising behavior and nutritional intake are being made every day, we could therefore benefit from tools that help us in the process. Not surprisingly we see that here too that a firm body of literature concerning decision-making is dedicated to this field.

A large survey study by Glanz, Basil, Maibach, Goldberg, & Snyder (1998) focused on identifying membership of certain lifestyle clusters and determinants of dietary habits. They found several determinants for decision-making regarding nutritional choices. According to their study, elderly people, women and certain ethnic groups value the importance of nutrition higher than others. Young people on the other hand are more focused on the cost and convenience of food. An overall analysis of the most important factors that influence food choice revealed taste to be the most important one, followed by cost, nutrition, convenience and weight control.

A later study by Neumark-Sztainer, Story, Perry, & Casey (1999) also studied determinants of food choice, but focused specifically on adolescents. This study however was performed in focus groups rather than by means of a survey. Adolescents in the ages 12-13 and 15-16 were questioned during these focus groups regarding their reasons for choosing particular foods and their opinion on how to change this for the good. Factors of primary importance in adolescents' decisions regarding food were reported to be hunger/food craving, taste, time and convenience. Factors of secondary importance were found to be availability, parent influence, perceived benefits and situational (social) aspects.

In order to be able to thoroughly comprehend people's decision-making process when it comes to food and nutrition, it is necessary to look beyond choice determinants and also focus on possible barriers that people face while making these decisions. Stevenson, Doherty, Muldoon, Barnett, & Trew (2007) did just that by means of a set of focus groups. Similar to the work by Neumark-Sztainer et al. (1999) they also targeted adolescents in order to find out what exactly prevented them from making healthy decisions. One of the most interesting findings of their work however was that healthy eating as a goal on its own is completely absent in young adolescents. Four key factors that underlie this phenomenon and compose the main barriers to healthy eating were found to be: physical and psychological rewards of food, perceptions of food and eating behavior, perceptions of contradictory messages and issues with the concept of healthy eating and dieting. In short, adolescents often have or use inaccurate information to base their decisions regarding healthy eating on. On top of this, they may have built up false associations between emotion and food. Finally, there often exist misconceptions about short-, and long-term effects of different foods.

The adolescents in the study above could therefore benefit from more accurate information on food and more insight in long-term consequences of an unhealthy lifestyle. Field experiments by Downs, Loewenstein, & Wisdom (2009) showed that providing nutritional information can result in positive effects in the consumer's health. They argue that supplying information on food to the consumers facilitates the ease with which healthy choices can be made. It is even possible to help people move in the right (healthy) direction by framing this information in such a way that does not limit the decision of the user but rather 'nudges' them towards a healthy choice. Their work does put a warning however on using caloric information to influence peoples dietary behavior, since this particular piece of information

can also lead to perverse effects. Caloric information can – especially by dieters - sometimes be misinterpreted and even lead to an increase of calorie intake.

Above work shows that information on food and nutrition should be dealt with carefully. When delivered to the right people and combined with the decisional determinants discussed above on the other hand it can result in better and healthier decision-making.

### **1.5 This research**

As the literature discussion on current health interventions in paragraph 1.2 also shows, existing work solely focuses on methods and theories that change people’s motivation or attitude. Intervention studies change people’s attitude, help them set goals and tell them to maintain these goals. What this approach often lacks however is to explain how exactly to reach these goals. In this study we will therefore take the next step by investigating how participants that are motivated at the start of our intervention can be helped with more information. Contrary to earlier studies, we aim to help users to reach their goals by allowing them to choose from small clear-cut assignments that tell them exactly what they need to do. By implicitly dividing the end-goal in smaller sub-goals and allowing for active participation in selecting personally relevant items by the user we aim to increase the effectiveness of our intervention.

Delivering the right information to the right people thus will be key in improving health of our participants. Our discussion of health interventions in paragraph 1.2 also showed that effectiveness of these interventions is largely determined by the accuracy and personal relevance of the given information.

With regard to the existing information on healthy living the goal of this research is twofold. Firstly, we aim to collect as much information on healthy eating and physical exercise as possible from both experts and available literature and formulate this in small and clear-cut assignments.

Secondly, we will structure this information based on a set of generic attributes. Such a set of generic attributes can help the user comparing potential health measures. It will allow them to tradeoff aspects, just like product features can be traded off in consumer decision-making. Structuring of the data will allow us to apply a recommender system that can make use of these attributes and help the user in making decisions that will possibly lead to a healthier lifestyle.

Since the field of computer tailored interventions in the health domain is relatively young, third generation interactive interventions have not been explored extensively. While most of these experiments have prominently relied on behavioral theories, none so far have incorporated recommender systems to deliver health improvement measures to the public. However, especially in third generation systems, recommender systems have a large potential value; as such systems typically have limited screen sizes and thus cannot provide users with a lot of information at once, so it is even more important to tailor the advice to the user. We

will use the most important decision making determinants that were discussed in paragraph 1.4 as the basis for this tailoring.

## **1.6 Hypotheses**

Studies into the effectiveness of computer tailored interventions in the field of nutrition (Brug et al., 1999) and in the field of physical activity promotion (Kreuter et al., 2000b) confirm that people can indeed benefit from digital tools that help them in finding their way in the densely populated area of health recommendations and tips.

This research extends this knowledge by combining and structuring information on both areas of health improvement - both nutrition and physical activity – and help people in deciding which of these measures are most appropriate for them by allowing them to compare and trade-off measures on their relevant attributes (e.g. time needed, calories burned). For example: when planning your daily workout on a busy day the system might suggest you to take your bike to work and eat a light meal instead of attending a spinning class in the evening.

Knijnenburg et al. (2011) show that a similar system for energy saving measures was able to maximize experiential aspects such as choice satisfaction and trust in the system by accounting for personal characteristics. Providing tailored recommendations is expected to increase experience and might subsequently lead to more behavioral change.

In order to validate above-mentioned reasoning, we have formulated the main research question of this work as follows:

*Will a tailored lifestyle recommender system result in improved user experience and increased healthy behaviour compared to a non-tailored recommender?*

In order to answer this question we have formulated 5 hypotheses that are based on literature and will serve as the main guidelines for the set-up of this research.

Work by Knijnenburg et al. (2011) showed that accounting for personal characteristics can be used to maximize user experience. This research will therefore compare a content-based recommender system - more specifically an attribute-based system – that generates tailored recommendations versus a non-tailored system. Normally the best (non-tailored) baseline would be a system that simply presents the user with the most popular measures, however since this is not available yet we use a non-tailored recommender system that simply generates a list of random items as the baseline for this experiment. Using a random set of recommendations is also considered to be closely related to the way in which this information is generally found on the Internet in real life; unstructured and in no particular order. Since this research will compare an attribute-based recommender system with a randomly generated set of recommendations, we hypothesize that:

**H1:** Tailored recommendations will lead to increased user experience, such as a higher system satisfaction and more choice satisfaction with suggested measures compared to a set of non-tailored recommendations.

Since tailored information is considered to be more personally relevant and therefore read more and processed more consciously (Brug et al., 1999) we expect similar effects of the recommender system on the amount of selected measures as was found in work by Reijmer (2011b). We therefore hypothesize that:

**H2:** Tailored recommendations will directly, or mediated by (choice) satisfaction, increase the amount of selected measures compared to non-tailored recommendations.

As was identified by Norman et al. (2007) frequency of use and goal setting attribute to a greater likelihood of behavioral change. Selecting healthy measures from the system under study is considered a form of goal setting. Dividing end-goals into smaller sub-goals was also identified to be a particular strength of eHealth interventions and contributed to more behavioral change. Selecting a list of healthy measures can be considered constructing a list of sub-goals that lead to a common end goal: becoming healthier. The latter leads us to hypothesize that:

**H3:** Higher choice satisfaction and increase in the number of selected measures will lead to more healthy behavior compared to lower evaluations of choice satisfaction and number of selected items.

Although earlier research has shown some effects of tailored health materials on behavior, these effects differed in magnitude across personal characteristics such as age, gender and physical condition. Because a study by Vandelanotte et al. (2005) already showed that unhealthy participants were more likely to benefit from tailored health materials than healthy individuals, we hypothesize that:

**H4:** The effectiveness of (tailored) recommendations on user experience and healthy behavior will be different for different user groups.

The schematic overview in Figure 4 illustrates the hypothesized relations between the different concepts.

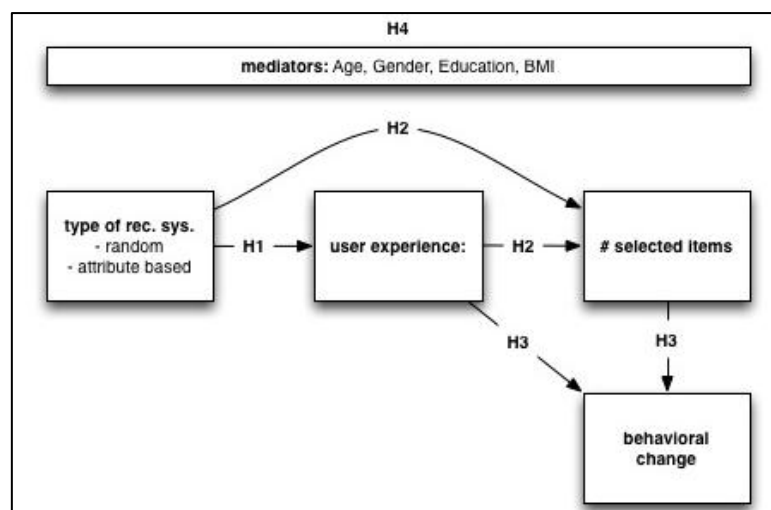


Figure 4 - Schematic overview of hypotheses stated above.

There are many studies that have focused on behavioral change in the health domain, those studies that did find a positive effect often also found this effect to persist over time. However, the effect size of most behavioral interventions is dependent on the frequency of use and commitment to the chosen items. Although we anticipate the effect size of our intervention to decrease over time, we hypothesize that:

**H5:** Changes in the use, evaluation and compliance with the tailored recommender over time will be smaller than those in a non-tailored recommender.

## **2. Method**

### **2.1 Design**

We tested the above hypotheses in two studies using the same recommender system, the first study tested the system in a single session while the second study tested the system in a longitudinal way over the course of 3 weeks. The first study addressed hypotheses 1 to 4 whereas study two was based on multi-session interaction and therefore also covered the temporal aspect (H5) of this research. Additionally, the second study put more focus on actually measuring behavioral change (H3). Both of the studies had a between subjects design and were performed online.

### **2.2 Participants**

The people that were targeted for participation in both studies had to meet only a few requirements. First of all, participants needed to have a basic understanding of how to use a computer and how to browse the Internet. This knowledge was required to be able to use the Lifestyle Coach website without difficulties. A second requirement followed from the assumption we made earlier in paragraph 1.5, namely that participants had to be intrinsically motivated to at least try to improve their general health. In order to assure this, financial rewards were kept low and emphasis in recruiting participants was put on possible positive health effects resulting from use of the lifestyle coach.

Participants in the first study were recruited by means of three different channels. First of all from a database containing people interested in online research. Through an e-mail request participants were invited to enroll for the experiment. Secondly, employees from the NIVEL institute were also invited for participation by means of an e-mail invitation. Finally social media were employed to find people interested in participation in this experiment.

All of the participants above were promised a €2 reward upon completion of the experiment. Participants were not allowed participation in both studies; therefore recruitment for the second study was done in a different manner. Newspaper articles in the ‘Roosendaalse Bode’, ‘BN De Stem (Roosendaal edition)’, ‘BN De Stem (Bergen op Zoom edition)’ and TU/e’s ‘Cursor’ invited people interested in participation in a “healthy” experiment to subscribe. A €20 reward was raffled per 5 participants upon completion of the experiment.

### **2.3 Procedure**

Below, we present the full experimental procedure per study. There is a large overlap between the set-up of the first and second study, however differences and communalities will be clearly outlined in the following paragraph. In both studies, participants were randomly selected to use either the attribute-based recommender system or the random recommender system.



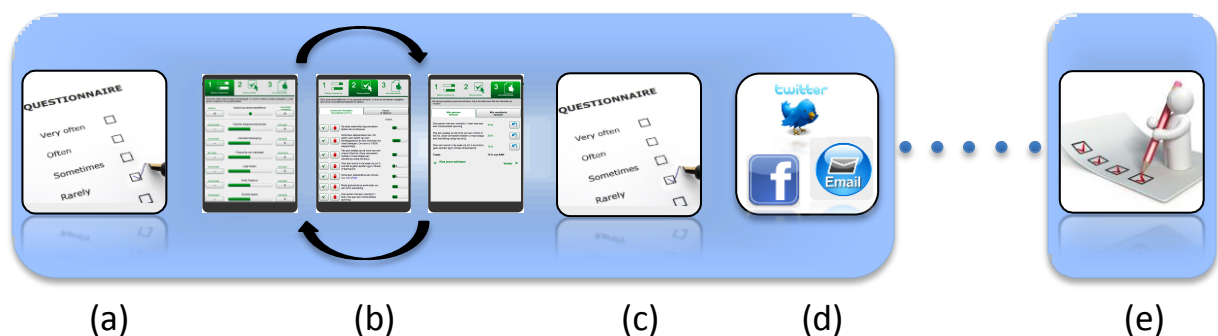
### 2.3.1 Study 1

Prior to participation in this online experiment, participants were briefly informed of the purpose of this study. They were told that they were about to use the Lifestyle Coach website to make a selection of healthy measures. They were also told that a week after participation they would receive an email, inquiring them about the status of their chosen measures.

On loading the lifestyle coach website, a message was presented to the user assuring discretion and anonymity in processing of the experimental data. Once enrolled in the Lifestyle Coach experiment, participants first filled in the demographic and personal health information survey. Following this, participants started with the actual Lifestyle Coach system as described in paragraph 2.4.2. Included in the lifestyle coach was a 5 step introductory text (See Appendix 7.4) that explained all of the following functionality and terminology of the system.

During interaction with the Lifestyle Coach, participants composed a personal list of healthy measures. There was only one requirement to the number of measures to be chosen, participants had to be able to execute them over the course of one week. Apart from that participants were free to choose any amount of healthy measures they desired. After iterative interaction with the Lifestyle Coach participants were asked to fill in the user experience questionnaire that was described in paragraph 2.5. Before finalizing the session, participants had the opportunity to share their list of chosen healthy measures through different social media (Twitter, Facebook, LinkedIn or Google+). While this last page was being loaded, an automatic email message containing the participant's personal list of healthy measures was sent to the participant.

A week after the participant composed his or her above mentioned personal list, he or she again received his or her personal list by e-mail followed by a request to indicate a status for all of the selected measures and to reply the email to sender. Figure 5 shows a schematic overview of the procedure for study 1.

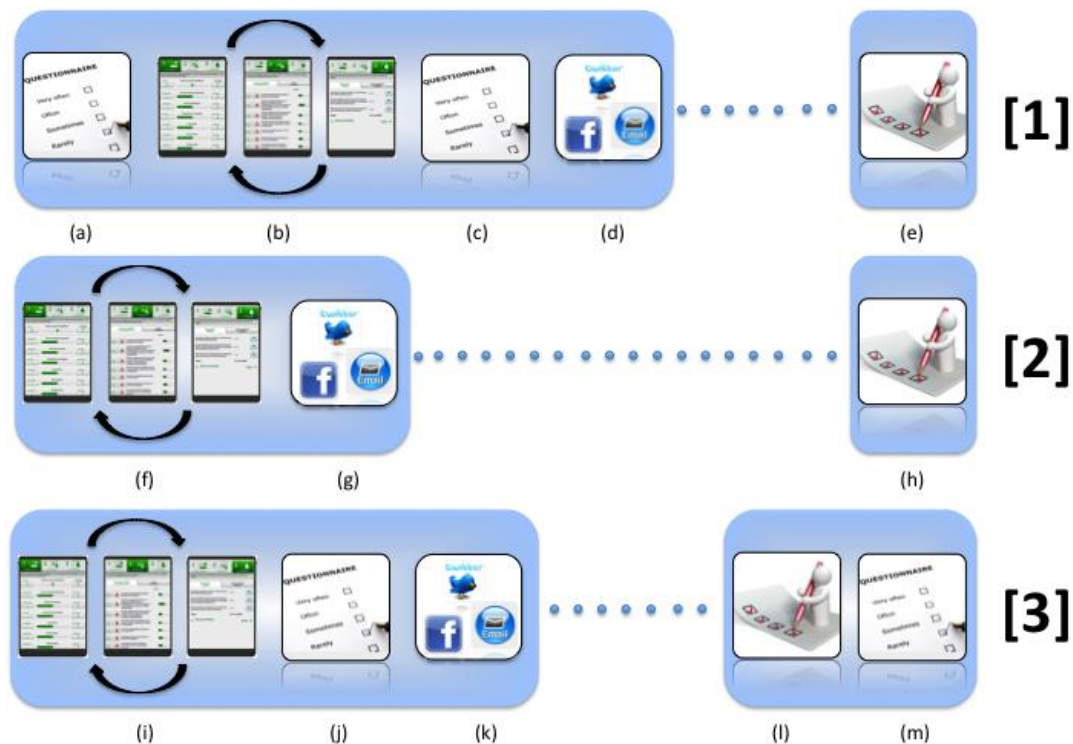


**Figure 5 - Schematic presentation of the procedure for study 1. (a) Demographic questionnaire, (b) interaction with the lifestyle coach, (c) user experience questionnaire, (d) automatic email reminder with personal list and possibility to share chosen measures & (e) user fills in status for all of the chosen measures on his list.**

### 2.3.2 Study 2

The second study employed a similar design as the first study when it came to interaction with the Lifestyle Coach; the temporal aspect however resulted in a slightly different experimental design as can be seen from Figure 6. In this study, participants worked with the lifestyle coach system for 3 weeks, starting on the 9<sup>th</sup> of June and ending on the 1<sup>st</sup> of July. During the weekends, participants used the lifestyle coach to select healthy measures and composed a personal to-do list. During weekdays, participants had the opportunity to follow up on their personally selected list.

During the first weekend of this experiment, participants first filled in a demographic questionnaire, identical to the one used in the first study. Following this, participants were prompted with the two questionnaires concerning fruit, vegetable and fat intake and physical exercise habits as will be discussed in paragraph 2.5. After completion of the three before mentioned questionnaires, participants received simple feedback on their BMI and their current habits concerning food intake and physical exercise. The feedback served as an extra motivation for using the lifestyle coach.



**Figure 6 - Schematic representation of study 2. The numbers in square brackets indicates week numbers. (a) Demographics, food and exercise questionnaire, (b) User interaction with the recommender system, (c) User experience questionnaire, (d) Possibility to e-mail or share selected measures by the user, (e) User fills in status for all of his or her selected items, (f) User interaction with the recommender system, (g) Possibility to e-mail or share selected measures by the user, (h) User fills in status for all of his or her selected items, (i) User interaction with the recommender system, (j) User experience questionnaire, (k) Possibility to e-mail or share selected measures by the user, (l) User fills in status for all of his or her selected items & (m) Food and exercise questionnaire.**

Having successfully completed the questionnaires and received feedback, participants were allowed to interact with the lifestyle coach in a similar way as in study 1. After this

interaction participants were prompted with the user experience questionnaire that will be discussed in paragraph 2.5, the session ended after completion of the questionnaire.

During the weekdays that followed, participants could at any time log into the system to view their personal list of selected healthy measures. While logged in, they were able to indicate whether they had already completed a given measure or not.

In the next weekend, participants first had to set a status (“Yes I did this”, “I Partly did this” or “No I did not do this”) for all of the healthy measures that they had chosen the week before. After doing so, they started a new session with the lifestyle coach during which they composed a new personal list of healthy recommendations that they then attempted to carry out during the week that followed.

This cycle repeated itself after the second weekend and after the third and last weekend of the experiment. During the last weekend, participants only had to indicate whether they had completed the items on their personal list, no interaction with the Lifestyle Coach followed. Before finalization of the experiment, participants once more filled in the food and physical exercise questionnaire they had filled in at the start of this research. User experience was also measured a second time; this however took place during the second last weekend of research, since that was the last time they interacted with the Lifestyle Coach for selecting healthy measures.

## **2.4 Materials**

### **2.4.1 Database with healthy measures**

Large amounts of data are the key ingredient for any recommender system, ours is no exception to this. In order to use these huge piles of data, it needs to be structured first. The type of recommender system that we are using was content-based, made use of products’ attributes and was built on the MAUT decision model discussed in paragraph 1.3. Multi attribute utility theory assumes several key attributes that together define the characteristics of a certain product that is evaluated in the choice for this product. Structuring of our data was therefore done based on these attributes. In the case of this research, these products were healthy recommendations that focused on both nutrition and physical exercise. These healthy recommendations were mostly formulated as small but clear assignments. The recommender system, titled *Lifestyle Coach*, contained 140 of these recommendations. Examples of recommendations are for instance: ‘Brush your teeth while standing on one leg this week’ or ‘Go for a 30 minute lunch walk twice this week’.

All of the 140 recommendations were scored on different attributes (1-5 scale) in order to allow for applying the MAUT to the decision process. Based on expert interviews and the literature reviewed in paragraph 1.4, 7 we chose the seven different attributes that are listed below:

- *Food/Physical exercise*

The first attribute was binary in nature, recommendations were scored either positive when referring to physical exercise or negative when referring to food and nutrition.

- *Calories burned/saved*

Even though a healthy lifestyle comprises more than just a reduction of calorie intake and an increase in calorie expenditure, research has shown that it is still one of the more prevalent determinants in decisions regarding health. Because of the possible perverse effects mentioned in the literature discussion, calorie expenditure was used in an abstract way; users could only indicate if they wanted to burn/save many calories on a 5 point scale without seeing actual calorie values.

The underlying algorithm of the Lifestyle Coach does however not use calories as a direct measure to calculate effect size of a certain recommendation; instead we used so-called MET<sup>5</sup> values. Since MET values are directly applicable to physical exercise recommendations but not to nutrition and food related recommendations we tackled this issue in a different manner. By using the reduction of calorie intake that follows from following up on a certain recommendation we calculated an equivalent MET score that could be used to compare recommendations on both physical exercise and nutrition to each other. Finally, all recommendations were categorized in groups ranging from 0(no applicable MET score) to 5 (MET score > 8) in steps of 2 MET.

- *Intensity*

The intensity attribute was only applicable for the physical exercise recommendations. Scores again ranged from 0 (passive action e.g. buy this or monitor that) to 5 (high intensity). All nutrition related recommendations were scored 0 on the intensity attribute.

- *Frequency*

The frequency attribute indicated how often the user had to follow-up on the given recommendation, scores ranged from 0 (do something once) to 5 (do something continuously). Examples are for instance ‘prepare today’s healthy dinner suggestion from [www.voedingscentrum.nl](http://www.voedingscentrum.nl)’ or ‘make sure you do not drink any soda but rather plain water this week’.

- *Little extra time*

Closely related to the frequency of occurrence of a certain recommendation is the time it takes to carry it out. Since some recommendations can be performed while doing some other activity, the time attribute refers to the extra time consumed by the user when following up on a recommendation. Scores range from 0 (0% of the time it takes to follow up on this recommendation can be used to spend on other activities, e.g. take part in a spinning class) to 5 (100% of the time it takes to follow up on this recommendation can be used to spend on other activities, e.g. refrain from eating chocolate or candy for a day).

---

<sup>5</sup> The *Metabolic Equivalent of a Task (MET)* is a physiological measure that expresses the intensity of physical exercise in multiples of one’s resting metabolic rate.(Ainsworth et al., 2000)

$$1 \text{ MET} = 1 \frac{\text{kCal}}{\text{kg} \cdot \text{h}} = 4.184 \frac{\text{kJ}}{\text{kg} \cdot \text{h}}$$

- *Low cost*  
Although health can be regarded as a greater good than financial wealth, money has also proven to be a key ingredient for decisions regarding personal health. Therefore the cost attribute represented the amount of money needed for a particular recommendation. Scores ranged from 0 (no costs or even profit, e.g. get off the buss one stop too early and walk the last bit) to 5 (> €80, e.g. buy a ball-chair for your office work-place) in steps of €20.
- *Social*  
The social attribute depicts the amount of social influence that is present in following up on a given recommendation. Since humans are social beings and are therefore highly susceptible to social influence, the social attribute allows the user to indicate his preference for the amount of social influence in the recommendations. Scores again range from 0 (recommendation only focuses on user without social interaction, e.g. stand on one leg while brushing your teeth) to 5 (recommendation involves other people or is performed in public with high social impact, e.g. agree with your colleagues to take turns in providing fruit for the whole office).

Experts from the field of physiotherapy, nutrition and human movement sciences composed the database containing the 140 healthy measures. In order to get a diverse set of products for our database, we designed different scenarios for which to construct recommendations. (e.g. Bram is a busy work-a-holic with little time to spare, please write down 10 recommendations he can use to improve his health) Scenario's typically combined 2 or 3 attributes in order to give experts some direction in designing the recommendations. Scenarios were varied on these attributes in order to get an evenly distributed set of recommendations. Following this, experts scored the recommendations on the attributes we discussed above. In most cases this was done by two experts, if scores didn't match we used the mean score of the two. Finally we checked if all attributes were valued equally important, regardless of personal preferences, in making decisions regarding personal health. Latter information was used to set the initial weight for the attributes. A small survey was sent out to 25 people in order to determine these initial weights. Initial values for the weights can be found in Table 2. The initial weights were multiplied with the weights set by the participants and then used in the calculation for utility scores. Participants were still allowed to indicate specific personal preferences.

**Table 2 - Initial weights for the different attributes in the lifestyle coach system.**

Attribute	Initial weight multiplication factors
Intensity	2.5
Calories burned/saved	2.5
Frequency	1.5
Cost	1
Time	1.5
Social aspect	1

## 2.4.2 Software

The recommender system was built in two different versions (see Figure 7), addressing the two different conditions comprising this work. Recalling the research question that is central to this research, the main focus was on finding evidence for the possible added value of tailored recommendations compared to non-tailored recommendations. Therefore one version of the recommender system provided the user with tailored recommendations while the other version generated random recommendations. Random recommendations were in this case regarded as being non-tailored and served as the baseline in this experiment.

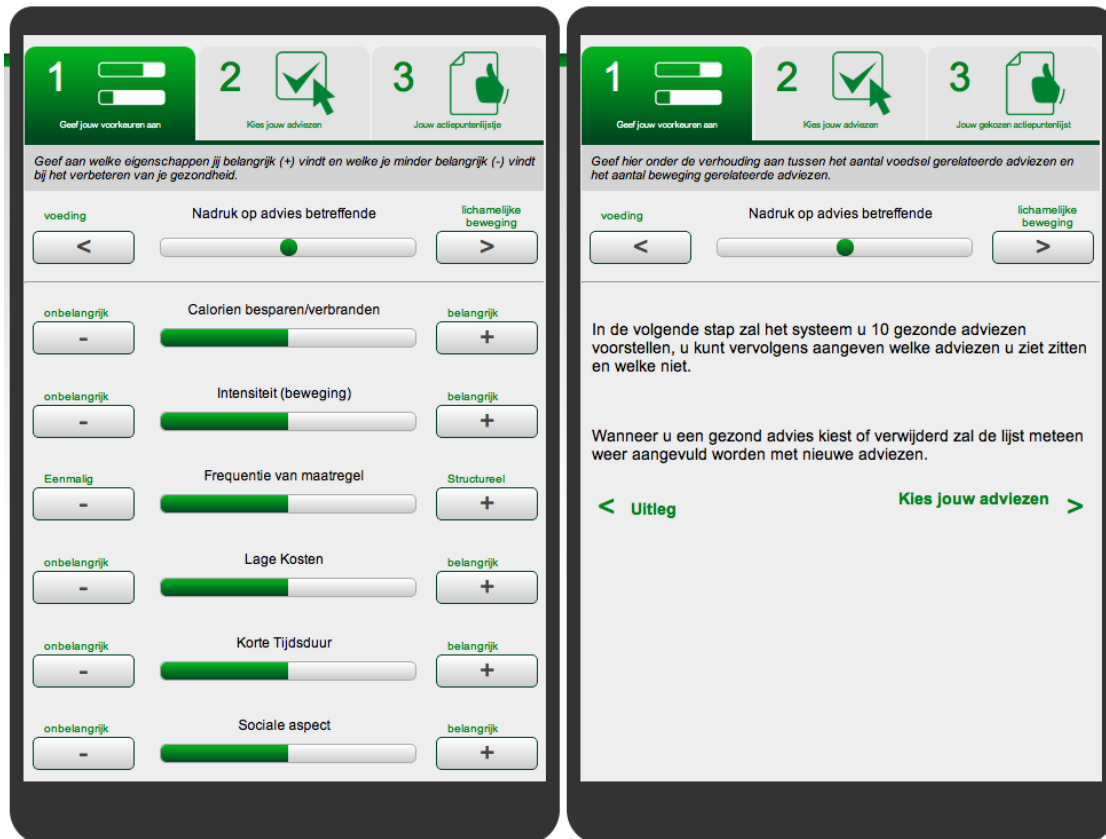


Figure 7 - Two different versions of the lifestyle coach, attribute-based (left) and random (right).

Altogether, the user – when in the attribute-based condition – could indicate his or her preferences by setting weights for the attributes discussed in the previous paragraph. Following this, the lifestyle coach calculated a utility score for each of the recommendations and sorted them accordingly. The food/physical exercise attribute was present in both the attribute-based condition as well as in the random condition.

To illustrate how attributing weights to the above-mentioned attributes lead to a personalized list of recommendations we describe two persona's below and provide examples of a possible set of recommendations that the lifestyle coach could give:

- ❖ Bram has a busy office job with little time to spare, money is not really an issue and he doesn't care much what other people think. So he will set the time attribute to 3, the cost attribute to 5 and the social attribute to 4, recommendations could be:
  - Replace your desk chair by an ergonomic ball-chair
  - Prepare your lunch and fruit the night before you leave to work



- Reduce your amount of internal e-mail; just drop by your colleague's office to deliver the message in person.
- Buy a headset and take a walk through your office or company while calling.
- ❖ Paul is a poor student that is rich on free time, does not want the whole world to know he's trying to improve his health and doesn't mind breaking a sweat. He will therefore set the time attribute to 0, the social aspect also to 0 and the frequency, calorie and intensity attribute to 5 in order to speed up the process, recommendations could be:
  - Consult the Albert Heijn recipe website at least twice a week and choose one of the budget and healthy recipes for dinner.
  - Go for a walk just before you go to bed, its healthy and helps to relax before sleeping.
  - Babysit a friend's dog for the weekend; he needs a walk 3 times a day and so do you!
  - Do you like wine? Drink red wine instead of white. This is slightly healthier.

The Lifestyle Coach consisted of three steps. In the first step the user indicated his or her personal preferences. (see Figure 8A) In the attribute-based condition this meant setting weights for the different attributes while in the random condition the user only determined whether the lifestyle coach should focus more on food or physical exercise recommendations. In the second step the Lifestyle Coach presented 10 recommendations to the user; this was equal for both conditions. (see Figure 8B) The Lifestyle Coach presented the recommendations together with details on the financials and time they would cost the user. Additionally, every recommendation was expressed in percentage of weekly-recommended amount (in Dutch: Aanbevolen Wekelijkse Hoeveelheid – A.W.H.) to allow the user to compare the impact of the recommendations that were presented. In this stage of the Lifestyle Coach, the user could either choose or delete items from the list. The presented list of recommendations was immediately complemented to 10 recommendations after each removal or choice.

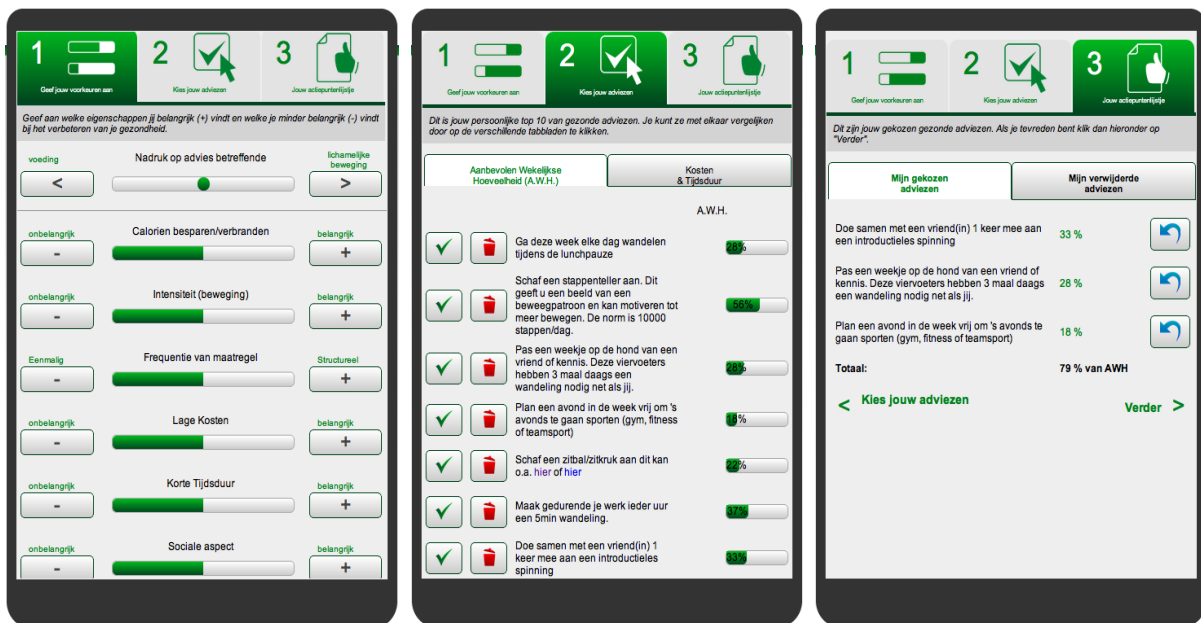


Figure 8 - The three stages in the lifestyle coach: A. setting weights for the attributes (left), B. selecting and deleting items (middle) and C. reviewing the list with selected and deleted items (right).

Step 3 presented the user with his or her personal list of chosen and deleted items. (see Figure 8C) The Lifestyle Coach was designed to allow the user to go through different steps iteratively until a satisfactory list was composed.

Furthermore, the Lifestyle Coach was designed to work on all platforms and browsers, meaning smartphones, tablets and desktop computers. The latter was done to increase accessibility of information in the system.

## **2.5 Measures**

Although measures overlap between our two studies, we will discuss them separately and relate back to our hypotheses when necessary. In order to measure user experience, we prompted users with a 25-item questionnaire (Appendix in paragraph 7.1) following interaction with the lifestyle coach. The questionnaire consisted of 5 items from the QUIS (Chin, Diehl, & Norman, 1988) questionnaire, measuring *general user experience*. The remaining 20 items are dedicated to measuring 4 specific factors: *system satisfaction, choice satisfaction, perceived system effectiveness and perceived control* (“Table\_of\_Contents.”, 2011). To allow for comparison of user experience measures across studies, we performed a factor analysis on the full dataset of the two studies combined. (The full dataset included only values from the first week of study 2)

We report result of the factor analysis in this paragraph rather than in the results section since it was not the objective of this research to verify the questionnaire but merely to use it as a measure.

Prior to analysis, scores on the variables UX\_SysSat\_4, UX\_SysEff\_4, UX\_ChoSat\_4, UX\_ChoSat\_6 and UX\_PercContr\_1 to UX\_PercContr\_4 were mirrored in order to align the direction of all of the questions.

Reliability analysis on the 5 QUIS items showed that the initial value for Cronbach’s alpha was 0.672, however after removing question 4 this increased to 0.774. Therefore we did not include QUIS question number 4 in calculating the overall score for general user experience.

The remaining 20 items of the user experience questionnaire were subjected to a factor analysis in order to get one single score per factor measured. Principal axis factoring was used to determine which questions captured the above-mentioned constructs best. Furthermore we applied an oblimin rotation to the factor analysis because of the interdependencies between factors. After removing 4 items (UX\_ChoSat\_6, UX\_PercContr\_5, UX\_SysSat\_4 and UX\_SysEff\_4)<sup>6</sup> based on their low communalities and excluding 3 items (UX\_ChoSat\_2, UX\_ChoSat\_3 and UX\_ChoSat\_4) based on high cross-loadings, we find the questions displayed in the table below to best represent the earlier mentioned user experience constructs. Together these 4 factors explain 72.7% of the variance in the data. Factor loadings are shown in Table 3 below.

---

<sup>6</sup> See appendix 7.1 for full questionnaire.



**Table 3 - Factor loadings for the relevant items from the 25-item user experience questionnaire (loading scores <= 0.300 are suppressed).**

	Factor 1	Factor 2	Factor 3	Factor 4
[SysSat_1] I liked the items that were recommended by the Lifestyle Coach.	0.747			
[SysSat_2] The items that the Lifestyle Coach recommended fitted my preferences.	0.809			
[SysSat_3] The items that the Lifestyle Coach recommended were relevant.	0.540			
[SysSat_5] The items that the Lifestyle Coach recommended were well chosen.	0.734			
[ChoSat_1] I liked the items that I selected.		0.688		
[ChoSat_5] The items that I selected fitted my preferences.		0.869		
[SysEff_1] I enjoyed myself while working with the lifestyle coach.			-0.638	
[SysEff_2] I would recommend the Lifestyle Coach to other people.		0.348	-0.615	
[SysEff_3] Working with the Lifestyle Coach was a pleasant experience.			-0.789	
[PercContr_1] I had limited control over the way in which the Lifestyle Coach recommended items to me.				0.615
[PercContr_2] The way in which the Lifestyle Coach recommended items to me was very strict.				0.386
[PercContr_3] The Lifestyle Coach was very limited compared to the way in which I usually make healthy decision regarding my lifestyle.				0.688
[PercContr_4] I would have liked to have more control over the Lifestyle Coach.				0.802

The mean scores for relevant items per factor were calculated and used in further analysis. We chose means instead of regression scores since the means tell us more regarding true evaluation of our system instead of the regression scores that express the factors as linear combinations of each other. The same was done for the four relevant items making up the QUIS measure.

### 2.5.1 Study 1

The focus of this study was mainly on testing hypotheses 1, 2 and 4 and to a lesser degree hypothesis 3. Prior to interaction with the lifestyle coach, participants were presented with a general questionnaire focusing on demographics and on several questions regarding personal health. Demographics included *age, gender, ethnicity, educational level, length and weight*. Health related questions were concerned with people's current *fruit and vegetable consumption and amount of daily exercise*. (See Appendix 7.2 and 7.3)

Hypothesis 1 focused on the effect of type of recommender system on the user experience, measured by the mean score of the above mentioned user experience items from the 25-item questionnaire measuring: *system satisfaction, choice satisfaction, perceived system effectiveness, perceived control and general user experience (QUIS)*.

Hypothesis 2 focused on the number of chosen items. This was addressed by storing the amount of chosen and deleted healthy measures in our online database.

One of the consequences of performing an online study is that this research had to rely on self-reports to measure behavioral effects. Measuring the amount of healthy behavior needed to test hypothesis 3 was in the case of study 1 done by means of a follow up email one week after interaction with the lifestyle coach. Participants received an e-mail including their personally chosen list of healthy measures and a request to reply and indicate whether they had actually performed their chosen measures. Allowed answers were ‘Yes, I did this’, ‘I partially did this’ or ‘No I did not do this’.

Finally, the demographic information measured allowed for comparison of the effect sizes of the Lifestyle Coach on different groups of users (hypothesis 4). Possible mediating variables that could have led to between group differences were: *age, gender, educational level* and *BMI*<sup>7</sup>.

### 2.5.2 Study 2

Study two was similar to the first study; however it had a more elaborate design in regard to the temporal aspect and the measures of behavioral change. As mentioned earlier, study 2 involved multiple session of user interaction with the lifestyle coach. Participants now had 3 sessions during which they used the lifestyle coach to choose healthy measures. This multi-session aspect addressed the final hypothesis (H5) and thereby allowed for investigating the temporal effects of system interaction on user behavior, user experience and amount of behavioral change.

Building on the first study, study 2 investigated the amount of behavioral change as a result of user-system interaction more thoroughly. On top of participants’ self-report from study one, this study also prompted the users with an elaborate questionnaire regarding food consumption habits and amount of physical exercise. Food consumption was in this case measures by a 10-item questionnaire focusing on fruit and vegetable intake (van Assema, Brug, Ronda, Steenhuis, & Oenema, 2002) and a 35-item questionnaire focusing on fat intake (van Assema, Brug, Ronda, & Steenhuis, 2001)(see Appendix 7.2). Both questionnaires have been tested and verified and have been used in earlier work that analyzed the effectiveness of tailored health interventions (Oenema et al., 2005). Daily exercise was measures by means of the SQUASH questionnaire (Appendix 7.3).

---

<sup>7</sup> Body Mass Index (BMI) is a generalized measure for indicating one’s health and is calculated by the equation below:

$$BMI = \frac{m}{l^2}$$

, where *m* is the weight of a person in kilogram and *l* is his length in meters.

## **2.6 Analysis**

Before starting any analysis on the data from this work we removed several outliers in order to increase generalizability of our results.

The lifestyle coach recorded all of the process data on the browsing behavior of our participants, this included amongst other the time they spent on each page of the Lifestyle Coach. We removed values for the browsing time variable for all participants that were inactive (no clicking behavior) for more than 10 minutes. Participants were not excluded completely from the data; we merely excluded the value of this particular variable. Further analysis of the data revealed several outliers based on their z-scores. We deleted values that were located more than 3 standard deviations away from the mean; this did not influence results significantly. In one case we deleted the particular participant completely since multiple outliers were found in his/her data.

Following this, all variables involved in analysis were checked for normality and equality of variance in order to determine which statistical tests to apply. Variables that were not normally distributed were transformed by using a log transform; this solved the normality issue in most cases. Analysis for differences between experimental treatments were performed using an independent samples t-test, when data transformation was not able to solve normality issues in the data we applied a non-parametric Wilcoxon-rank test.

Correlation analysis was used to find possible interesting effects between variables, following which we performed a regression analysis to build our full model on.

### 3. Results

We will discuss results of both studies in this work separately because of the involvement of the temporal aspect in the second. Moreover, different samples of participants have been used for both studies. Demographics of both samples are reported prior to reporting the experimental results.

#### 3.1 Study 1

The participant sample consisted of 90 participants, divided over the two experimental treatments. Demographic data on participants in Table 4 show that both groups were equal on most variables except B.M.I. The attribute group contained a higher number of overweight (B.M.I. > 25) participants while at the same time the average B.M.I. was marginally significantly higher ( $t(87)=1.968$ ,  $p=0.052$ ) in the attribute condition compared to the random condition. The education variable appeared to be highly skewed towards higher educated people.

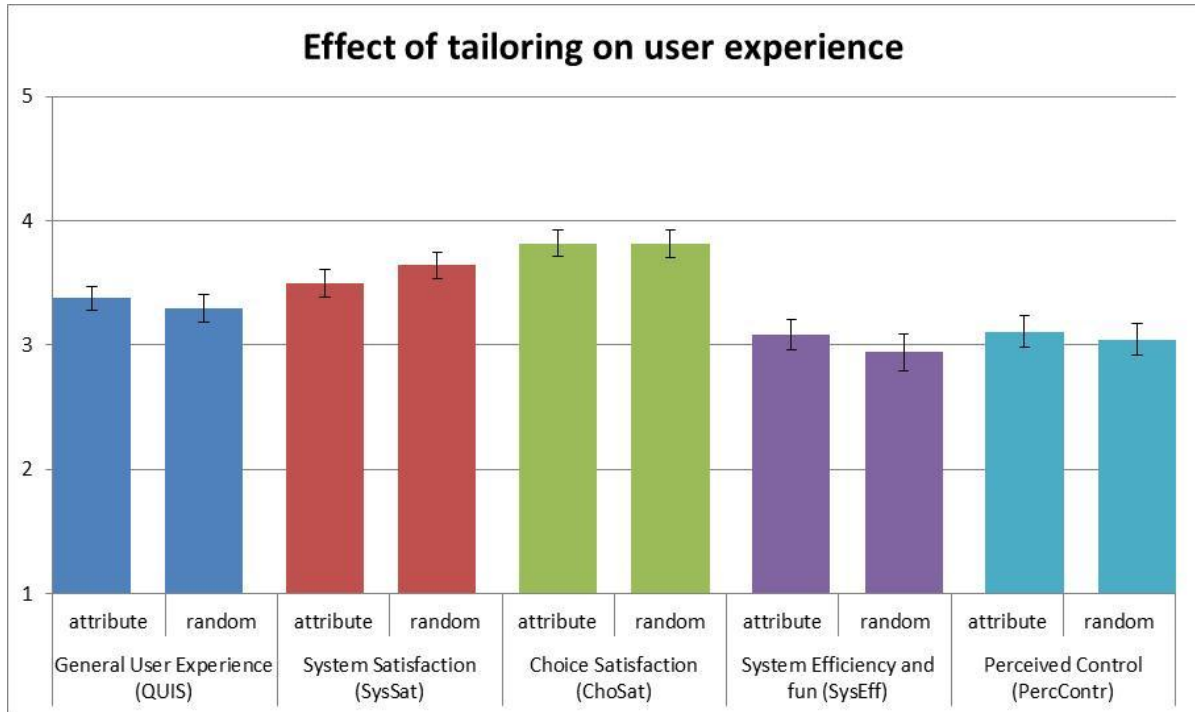
**Table 4 - Demographic data of full participant sample.**

Variable		Total		Attribute condition		Random condition	
		Statistic	s.e.	Statistic	s.e.	Statistic	s.e.
Participants	N	90		44		46	
Gender	Male	42 (47%)		21 (48%)		21 (46%)	
	Female	48 (53 %)		23 (52%)		25 (54%)	
Age		30.66	1.028	31.49	1.567	29.89	1.353
BMI	Mean	23.49	0.310	24.11	0.514	22.91	0.339
	> 25	23 (26%)		14 (33%)		9 (20%)	
Education	None	0		0		0	
	VMBO	3		2		1	
	HAVO	1		1		0	
	VWO	4		1		3	
	MBO	1		1		0	
	HBO	22		12		10	
	WO	59		27		32	

#### 3.1.1 The effect of tailoring

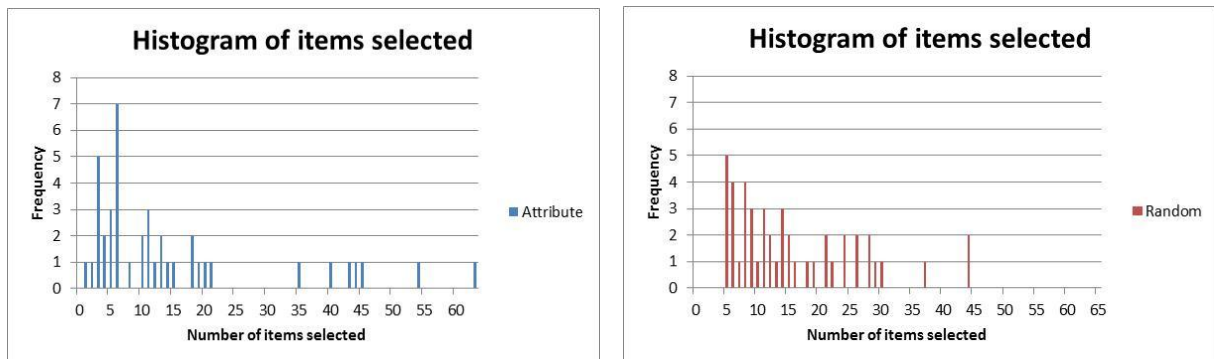
The participant sample valued both systems significantly above average on general user experience ( $t(39)=3.202$ ,  $p=0.003$ ), system satisfaction ( $t(39)=4.776$ ,  $p=0.000$ ) and choice satisfaction ( $t(39)=8.961$ ,  $p=0.000$ ), and slightly below or around average on system efficiency and perceived control. User experience scores are displayed graphically in Figure 9.

An analysis of the means on the different user experience items showed no significant differences between experimental treatments and therefore no proof for validation of hypothesis 1: there was no increase of user experience in the tailored treatment compared to the non-tailored treatment.



**Figure 9 - The effect of a tailored (attribute-based) recommender system compared to non-tailored (random) recommender system, measured by means of 5 user experience factors. None of the observed differences in means are significant. (error bars are one standard error of the mean)**

We first inspected results concerning the direct effect of tailoring on the number of selected and removed items. Participants selected on average 15.0 (se=2.39) items in the attribute treatment compared to 15.7 (se=1.51) items in the random treatment. Both numbers seem to be rather high given the task to make a list that could be executed in one week. Results are displayed graphically in Figure 10.



**Figure 10 – Number of items selected in both experimental conditions.**

The random treatment allowed for only one method of browsing through the recommendations: deleting items that were not relevant. The attribute condition on the other

hand offered a second way of browsing; by changing the attribute weights participants were able to find new recommendations that were therefore more personally relevant. We anticipated that increased personal relevance would lead participants in the attribute condition to remove fewer items compared to the random condition. Results however showed that participants in the attribute treatment removed on average 28.0 (se=4.88) items compared to 27.3 (se=4.48) items in the random treatment, this difference was however not significant.

As described in the previous chapter, each item was given an A.W.H. (percent of recommended weakly amount) score, indicating the amount of the item's contribution to a better health. Analysis of each participant's sum of A.W.H. scores showed that participants in the attribute treatment selected on average 167% A.W.H. (se=19.6) compared to 124% A.W.H. (se=16.3) in the random treatment, again this difference did not prove to be significant. The total percentage of selected A.W.H. per participant correlated marginally significant ( $r=0.238$ ,  $p=0.075$ ) with participants' B.M.I. indicating that people with a higher need for interventions set their personal goals higher than people with a lower need for this. Participant with high B.M.I. scores were however not the cause for the high numbers of selected items since there was no correlation between A.W.H. and number of selected items ( $r=0.023$ ,  $p=0.834$ ). However, in order to truly compare the choices between participants we focus on the A.W.H. value per selected item. There was no correlation between B.M.I. and number of selected items, the average amount of A.W.H. per selected items on the other hand is significantly higher ( $t(55)=2.038$ ,  $p=0.046$ ) for participants with a B.M.I. above 25 ( $M=15.45$ ,  $se=1.859$ ) compared to participants with a B.M.I. of 25 and lower ( $M=17.71$ ,  $se=1.177$ ).

Finally, we checked for a possible direct effect of tailoring on the percentage of executed items by the participant. The response rate for the percentage of executed items was 57%, meaning that 57% of the participants indicated the status of their selection of chosen items. Participants in the attribute treatment on average fully executed 46.4% (se=4.64) of their chosen items compared to 48.1% (se=3.02) of their chosen items in the random treatment. Similar to the analyses above, this difference was not significant. Figure 11 shows a graphical representation of the relation between number of selected items and the percentage of items that is executed. The trend line in Figure 11 indicates that selecting more items will lead to less items being executed.

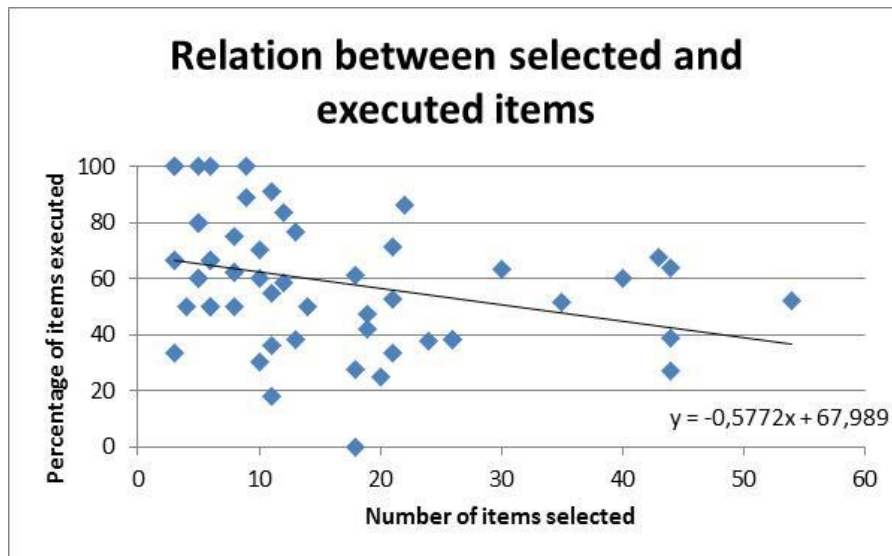


Figure 11 – Relation between number of selected items and percentage of items executed and partially executed by participants. Trend line equation indicates a negative effect of number of selected items on the amount of items executed. Regression equation is significant ( $p=0.014$ ).

The lack of differences between the attribute and random condition at first seemed surprising. However, our results suggested that a large part of the participants in both conditions did not set realistic goals: they were asked to compose a list for one week, but many selected a high number of items (some over 40). Also the data suggested that many participants in the attribute condition were using the system in the same way as in the random condition: finding measures by removing unwanted items from the list (rather than by changing the attribute weights). Although this phenomenon is interesting in itself, we are more interested in user experience and choice behavior for participants that set realistic goals and followed the experimental task. In order to determine the effect that unrealistic goal setting by a subset of our participant sample has had on our results, we created a post hoc condition that divided the initial two participant groups in two smaller groups based on the number of items they selected resulting in a 2x2 between-subjects design.

If we take into account the average number of selected items ( $m=15.4$ ,  $se=1.38$ ) and the average percentage of A.W.H. ( $m=147$ ,  $se=13.2$ ) selected by the full participant sample, a selection of about 10 items seemed to reflect about 100% of recommended weekly amount of healthy measures. Participants were therefore assigned to either the realistic goal group (selected 10 items or less) or the unrealistic goal group (selected >10 items).

An analysis of the choice behavior of participants that only selected 10 items or less now did show significant differences for number of selected items as can be seen from Figure 12. Participants that set a realistic goal (selected  $\leq 10$  items) selected on average 5.04 ( $se=0.494$ ) items while participants in the random condition selected 6.94 ( $se=0.399$ ) items, this difference was significant ( $t(38)=-2.895$ ,  $p=0.006$ ).

We also found a significant main effect for set goal (realistic or unrealistic) on the number of removed items ( $F(1,77)=54.640$ ,  $p=0.000$ ). Participants in the attribute condition removed on average 6.33 items ( $se=1.183$ ) when they had set a realistic goal versus 32.46 items ( $se=1.155$ ) when they had set an unrealistic goal. A significant interaction effect of goal set

and experimental treatment ( $F(1,84)=6.449$ ,  $p=0.013$ ) on the number of selected items was also found. In case of a realistically set goal, participants selected more items in the random condition ( $M=6.75$ ,  $se=1.117$ ) than in the attribute condition ( $M=4.48$ ,  $se=1.105$ ).

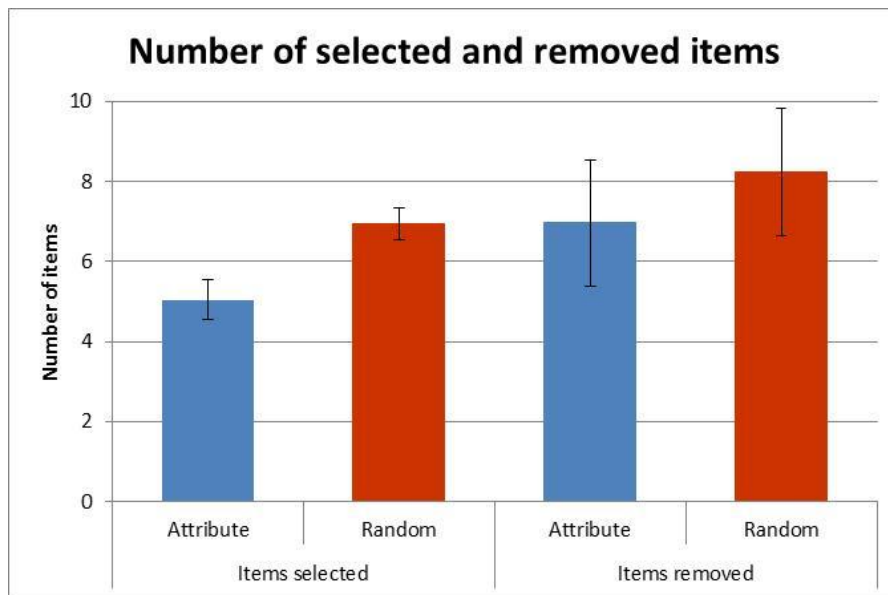


Figure 12 – Number of items selected and removed by participants in the attribute and random condition. Data is taken from participants that selected only 10 items or less. (error bars are one standard error of the mean)

Analysis of the user experience measures, including the new post hoc conditions now revealed significant interaction effects between type of recommender (attribute vs. random) and goal (realistic vs. unrealistic) for two of the user experience measures: general user experience ( $F(1,84)=4.157$ ,  $p=0.045$ ) and system effectiveness ( $F(1,84)=6.476$ ,  $p=0.13$ ). General user experience was rated higher ( $M=3.58$ ,  $se=0.151$ ) in the attribute condition, compared to the random condition ( $M=3.18$ ,  $se=0.167$ ) when realistic goals are set. However, when participants set unrealistic goals, this effect is absent or perhaps even reverses as can be seen from Figure 13 (left).

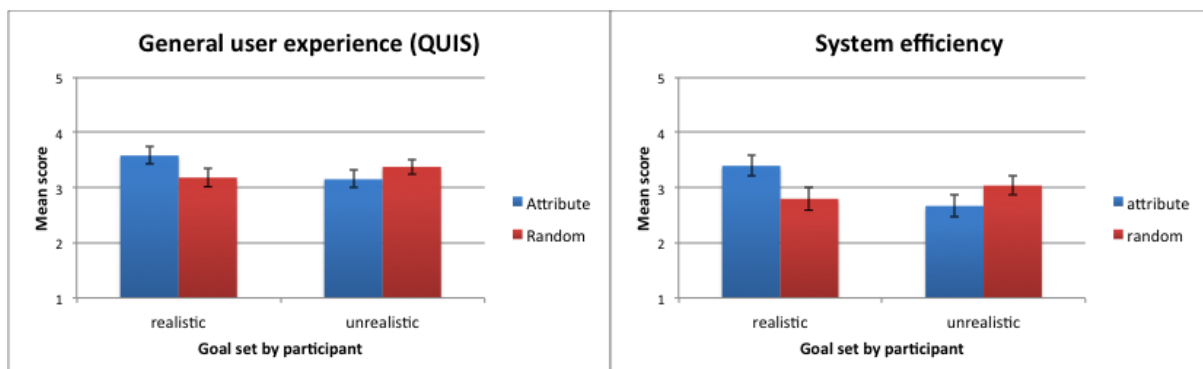


Figure 13 – Interaction effects between type of recommender system (attribute vs. random) and goal set by participant (realistic vs. unrealistic).

System effectiveness also showed to be higher when realistic goals were set in the attribute condition ( $M=3.39$ ,  $se=0.187$ ) compared to the random condition ( $M=2.79$ ,  $se=0.207$ ) when accounting for the interaction effect of type of recommender system and goal set by



participant (Figure 13, right). In case of unrealistic goals the before mentioned effect is absent or perhaps reverses.

The evidence presented above showed that some participants selected an extremely high number of items, resulting in an unrealistically high end-goal. The evidence also showed that differently set goals lead to significantly different evaluations of user experience and number of selected items (Figure 12 and Figure 13). Since we are only interested in user experience and behavior that follows from realistically set goals, we exclude participants with unrealistically set goals (i.e. participants that selected > 10 items) from further analysis.

### 3.1.2 The analysis of realistic choice behavior

The subset of our original participant sample now only consisted of 40 participants; demographic data are displayed below in Table 5. The data showed no major differences in age and education between groups while males and overweight participants are slightly underrepresented in the random condition compared to the attribute condition.

**Table 5 - Demographic data of the subset of participants that selected 10 items or less (realistic goal set).**

Variable		Total		Attribute condition		Random condition	
		Statistic	s.e.	Statistic	s.e.	Statistic	s.e.
Participants	N	40		22		19	
Gender	Male	18 (45%)		11 (50%)		7 (37%)	
	Female	22 (55%)		11 (50%)		11 (63%)	
Age		31.63	1.711	32.59	2.622	30.44	2.099
BMI		23.66	0.531	24.41	0.813	22.74	0.589
Education	> 25	12 (30%)		8 (36%)		4 (21%)	
	None	0		0		0	
	VMBO	1		1		0	
	HAVO	1		1		0	
	VWO	2		1		1	
	MBO	1		1		0	
	HBO	13		8		5	
	WO	22		10		12	

Analysis of this subset of our data with regard to our hypotheses that we stated at the beginning of this work in paragraph 1.6 resulted in the model that is displayed in Figure 14. The model in Figure 14 shows all of the significant and marginally significant relations between different variables in this model, relations are reported by their standardized regression coefficients.

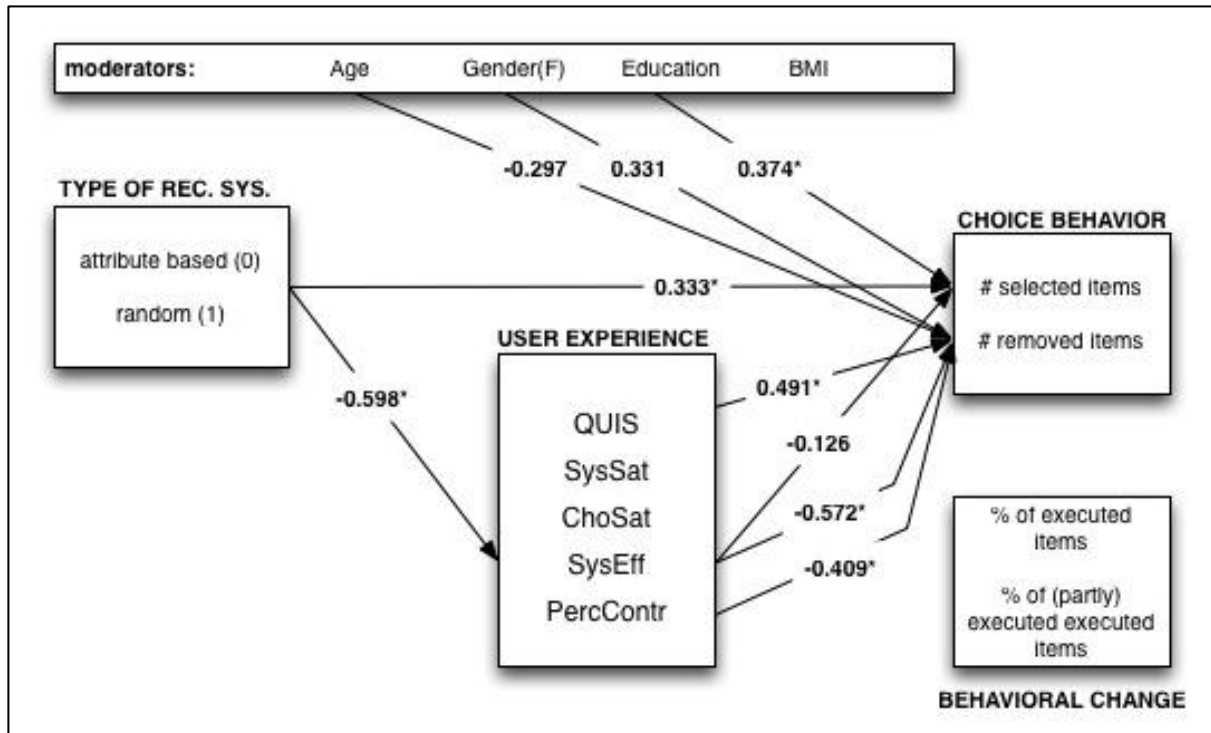


Figure 14 - Graphical overview of the results of study 1 including only participants that selected 10 items or less. Results are standardized regression coefficients. (Significance indicated by \*\* for  $p < 0.01$ , \* for  $p < 0.05$  and no subscript for  $0.05 < p < 0.10$ )

Following the order of hypotheses, we first discuss the regression performed on user experience. Five separate regression analyses were performed on each of the user experience measures. Variables were entered stepwise with the first block containing only the type of recommender and the second block all of the possible moderating variables (Age, B.M.I., Gender and Education) resulting in the regression coefficients noted in Table 6 below. Variables were excluded from the model in a stepwise fashion when significance exceeded 0.100. Following inspection of the significance of variables involved in this analysis, type of recommender (attribute = 0, random = 1) was the only significant predictor left in the regression equation for system effectiveness. This means that system effectiveness was the only user experience measure that was directly affected by the type of recommender system that was being used by the participants.

Table 6 - Predicting system efficiency. Table shows unstandardized regression coefficients. ( $R^2=0.161$ )

	Unstandardized		Standardized		p
	estimate	std. err.	estimate	t	
(constant)	3.394	0.148		22.880	0.000
Type of rec. sys	-0.598	0.221	-0.402	-2.703	0.010

In the following regression analysis we regressed the number of selected items. This variable was first log transformed in order to comply with the normality assumptions of a regression analysis. Similar to the analysis above, variables were entered stepwise at first and later removed if their significance value exceeded 0.100. The first block contained only the type of recommender system, the second block added the five user experience items and the third and

final block contained the possible moderating variables. Table 7 below shows the unstandardized regression variables used to predict the log transform of the number of selected items by our participants. Higher education as well as use of a random (non-tailored) recommender system will lead to higher numbers of selected items in users of the lifestyle coach system.

**Table 7 - Predicting the log transform of number of selected items. Table shows unstandardized regression coefficients. ( $R^2=0.323$ )**

	Unstandardized		Standardized		p
	estimate	std. err.	estimate	t	
(constant)	0.747	0.293		2.551	0.015
Type of rec. sys	0.325	0.130	0.348	2.495	0.017
Education	0.151	0.056	0.374	2.686	0.011

A separate analysis on the effect of evaluation of system efficiency on the number of selected items showed a significant relation. Since this opened the door to a possible (partial) mediation effect we performed a Sobel test to verify this. The Sobel test determined whether system efficiency acted as a mediator in the relation between type of recommender and the number of selected items. The Sobel test statistic was 1.73 ( $p=0.042$ ), indicating at least partial mediation. We re-calculated the true regression coefficients by using the Preacher & Hayes (2008) mediation analysis script resulting in the regression coefficients displayed in Table 8.

**Table 8 - Re-calculated regression coefficients by using the Preacher & Hayes (2008) script. ( $R^2=0.2248$ )**

	Standardized			
	estimate	std. err.	t	p
Expcond --> SysEff	-0.5976	0.2211	-2.7028	0.0102
SysEff --> ItemsSelected	-0.1264	0.0994	2.9949	0.2114
ExpCond --> ItemsSelected	0.3334	0.1479	2.2544	0.0302

As can be seen from Table 8, the mediation effect of system efficiency on the relation between type of recommender and number of selected items is fairly small indicating that most of the variance in the number of selected items can be explained by the type of recommender and only a little part by evaluation of system efficiency as a result of type of recommender.

Finally, we performed the regression analysis on the number of removed items in a similar way as with the number of selected items. Results for the regression analysis are displayed in Table 9. The data shows that females will remove more items during use of the lifestyle coach. A positive general user experience (QUIS) as well as more perceived control also contribute to a higher number of removed items. A high evaluation of system efficiency and higher age on the other hand result in lower numbers of removed items.

**Table 9 - Predicting the log transform of number of removed items. Table shows unstandardized regression coefficients. ( $R^2=0.418$ )**

	Unstandardized		Standardized		p
	estimate	std. err.	estimate	t	
(constant)	0.390	1.185		0.329	0.744
QUIS	0.684	0.279	0.588	2.451	0.021
SysEff	-0.722	0.268	-0.592	-2.696	0.012
PercContr	0.602	0.205	0.501	2.933	0.007
Age	-0.025	0.013	-0.297	-1.931	0.064
Gender (0=male)	0.625	0.307	0.331	2.034	0.051

We also performed a fourth regression analysis on the percentage of executed and partly executed items, no predictors were found significant. Altogether the three regression analyses reported above composed the model that is displayed in Figure 14. In order to allow for the combination of the different regression analyses into one model we used standardized regression coefficients in the model instead of the unstandardized values reported in the separate tables.

### 3.1.3 Discussion

We see that the type of recommender system that was being used by our participants resulted in different user experiences. It must be noted however that we only focused on participants that selected 10 items or less.

Although our first hypothesis reasoned that mainly choice satisfaction and system satisfaction would be affected by the type of recommender system that was used, we only found evidence for differences in system efficiency. Our participants experienced the attribute-based system to be more efficient than the random system.

Increased user experience would in turn lead to an increase in the number of selected items according to our second hypothesis. We found however that the opposite was true, participants in the random treatment chose significantly more items than participants from the attribute-based treatment. We did find evidence that system efficiency operated as a mediator in the effect of the type of recommender system on the number of selected items. However, this effect was very small since most of the variance in the number of selected items could be explained by the different types of recommender systems that were used.

Perceived control, system efficiency and general user experience did affect the number of removed items, however they were not affected by the type of recommender system being used and could therefore not be regarded as moderators.

As we did not find any variables that were able to predict the amount of actual behavior (items executed), we also have to discard our third hypothesis. Nonetheless, experimental data showed that participants fully executed 52.17% ( $se=7.579$ ) of the items they chose in the attribute condition compared to 53.77% ( $se=3.585$ ) in the random condition. Furthermore when we include also the partially executed items, we see that participants in the attribute condition executed 68.83% ( $se=6.026$ ) versus 70.25% ( $se=4.394$ ) in the random condition. Latter data suggest that although we do not find a difference between experimental

treatments, the overall effect of the intervention is reasonably good.

An exploratory analysis of the full dataset showed that the number of selected items was negatively correlated with the percentage of executed items. Since we excluded participants with large numbers of selected items based on their unrealistic goal setting, it is difficult to say whether this finding has any external validity.

Whether this effect is similar to all types of user groups (hypothesis 4) is hard to tell by the data from this study. We found no (moderating) variables that directly influenced the amount of behavioral change. However we did find marginal proof for the fact that people with a higher need for interventions (e.g. people with a higher B.M.I.) use these interventions more rigorous (i.e. select higher percentage of A.W.H.). Although we cannot show indisputable proof of differences between groups based on gender, age or B.M.I. we do have some evidence to support the fact that it is likely that at least B.M.I. results in different choice behavior. The latter therefore partially confirms our fourth hypothesis, however much more research has to be done to allow for claims concerning effect sizes.

### 3.2 Study 2

In total 111 participants registered for this experiment, they were distributed turn by turn in one of the two experimental treatments. Unfortunately, only 66 participants completed the first session with the Lifestyle Coach, leading to an unevenly distributed number of participants in both conditions. The difference between the 111 initial registered participants and the 66 that completed the first session was caused by 17 participants quitting during the first questionnaires and 28 more that dropped out during interaction with the Lifestyle Coach. Dropout numbers during the first interaction with the Lifestyle Coach were not found to be significantly different between conditions. Demographics of the participants that completed the first session of the experiment are displayed below in Table 10.

**Table 10 – Demographics of participants that started with study 2.**

Variable		Total		Attribute condition		Random condition	
		Statistic	s.e.	Statistic	s.e.	Statistic	s.e.
Participants	N	66		39		27	
Gender	Male	21(35%)		10 (26%)		11 (41%)	
	Female	45(65%)		29 (74%)		16 (59%)	
Age		34.94	1.667	34.44	2.063	35.67	2.823
BMI		25.18	0.469	24.64	0.437	26.04	0.979
Education	> 25	25(39%)		15 (39%)		10 (40%)	
	None	0		0		0	
	VMBO	3		2		1	
	HAVO	8		6		2	
	VWO	7		6		1	
	MBO	10		5		5	
	HBO	22		12		10	
	WO	16		8		8	

Also the remainder of this study suffered from an extremely high dropout rate as can be seen from Figure 15. Since quantitative analysis of the pre and post questionnaires regarding exercise behavior and fruit, vegetable and fat intake severely lost on external validity as a result we did not report them in this work. Results of the user experience questionnaire on the other hand were analyzed and reported since they comparison with results from the first study could lead to interesting findings.

Analyzing participants' comments, the high dropout rate of this study could in part be caused by the high pace of the experiment. The Lifestyle Coach was designed rather strict considering its frequency of interaction. The weekly cycle employed in this research might therefore have been too short to allow participants to execute the items they put on their to-do list.

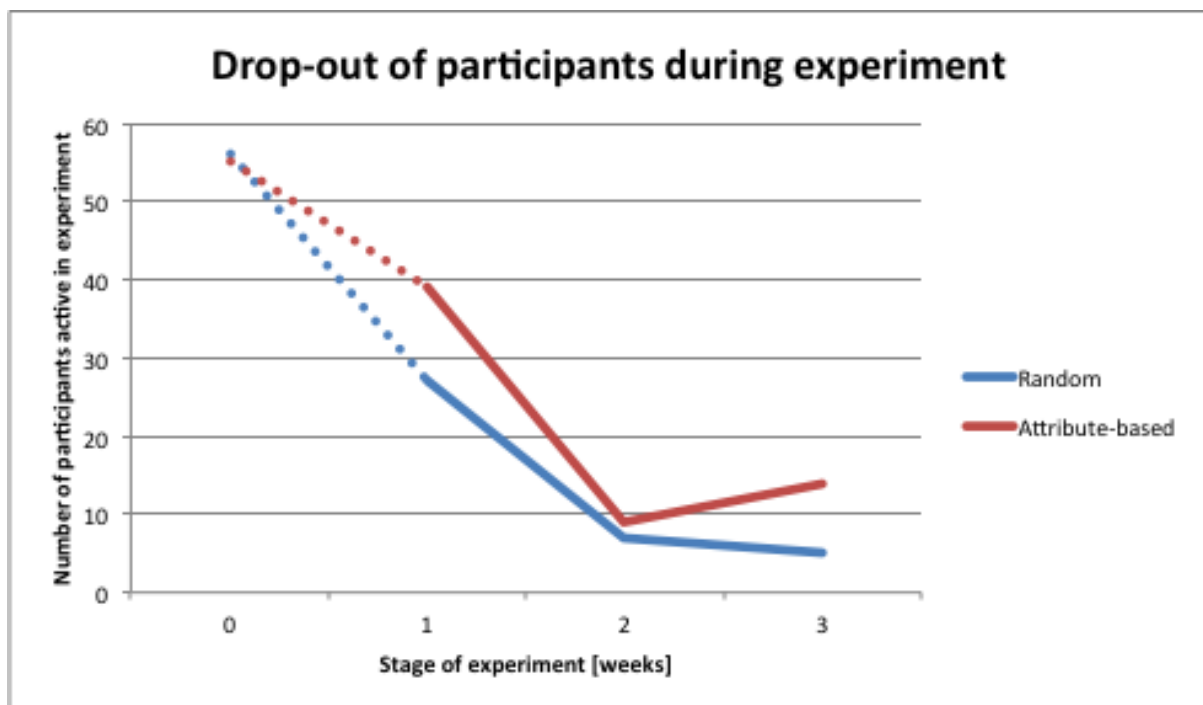


Figure 15 - Overview of participant dropout during experiment. (Dashed lines indicates the pre-experimental period during which about 40 participants already decided not to participate in the experiment, reasons are unknown)

The following paragraph will first discuss the results of the first week of this study, since the first week still contained enough participants to allow for quantitative analysis. Paragraph 3.2.2 will report on trends that follow from the results from the second and third week from this study.

### 3.2.1 The effect of tailoring

Figure 16 shows the number of selected items per experimental condition for study 2; we see that - similar to study 1 - again most participants selected an unrealistically large set of items.

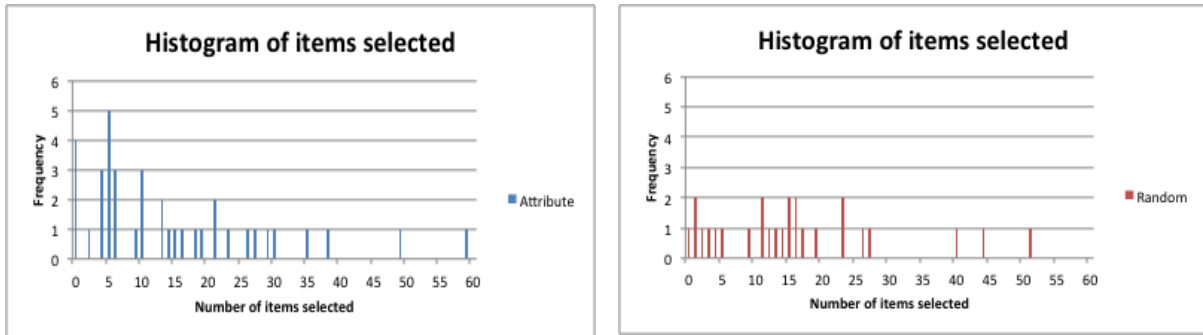


Figure 16 – Number of selected items for the two different experimental treatments: attribute (left) and random (right).

Applying the same exclusion principle as we did in study 1, we are left with only 15 participants in the attribute condition and 7 in the random condition. We now combine data from study 1 and from the first week of study 2 thereby increasing the external validity of our findings. Figure 17 shows the unified model, built from data from both study 1 and study 2 (first week only).

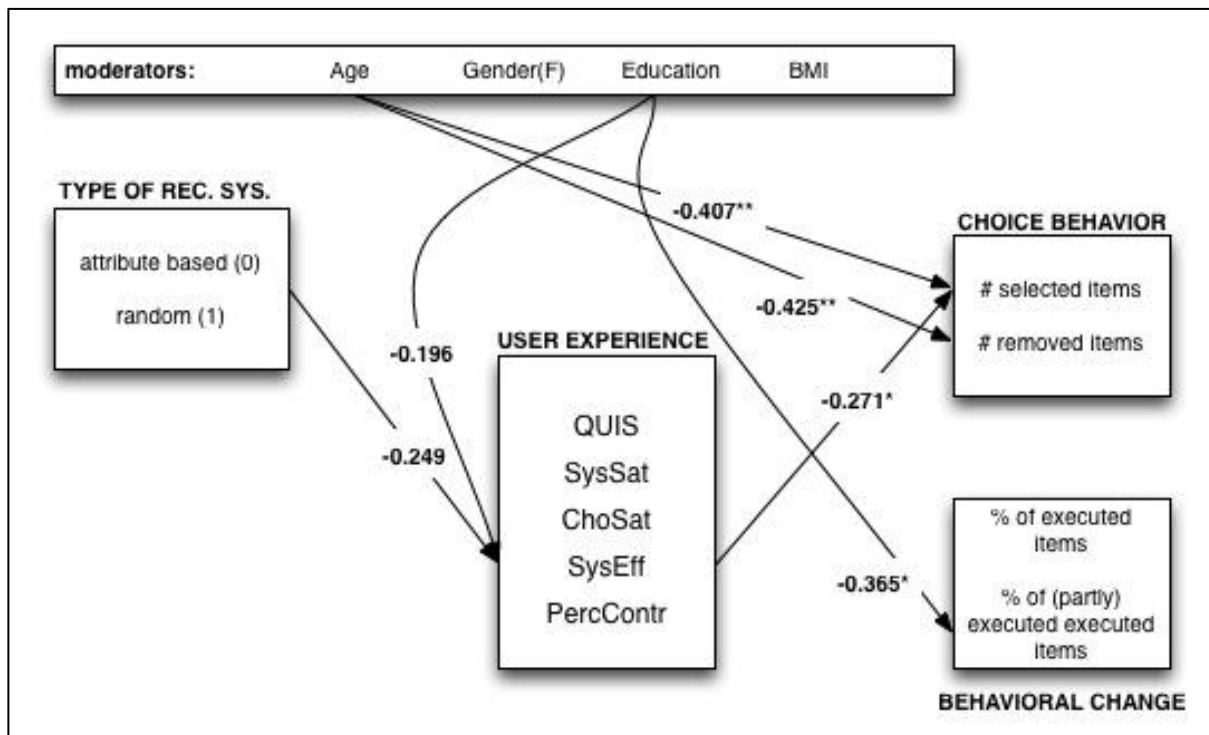


Figure 17 - Graphical overview of the results of study 1 and 2 combined including only participants that selected 10 items or less. Results are standardized regression coefficients. (Significance indicated by \*\* for  $p < 0.01$ , \* for  $p < 0.05$  and no subscript for  $0.05 < p < 0.10$ )

We performed separate regression analysis in order to build the model displayed above. Variables were inputted in the same way as in study 1. The analysis on the combined data from study 1 and 2 showed that system efficiency is dependent on the type of recommender system being used and the education of our participants as can be seen from Table 11.

**Table 11 - Predicting the system efficiency. Table shows unstandardized regression coefficients. ( $R^2=0.075$ )**

	Unstandardized		Standardized		
	estimate	std. err.	Estimate	t	p
(constant)	3.925	0.391		10.034	0.000
Type of rec. sys	-0.270	0.176	-0.196	-1.535	0.130
Education	-0.137	0.07	-0.249	-1.946	0.057

Regression analysis on the number of selected and removed items was again performed similarly to that in study 1. Results show that only system efficiency and age were significant predictors.

Table 12 shows the regression coefficients for the relevant predictors.

**Table 12 – Predicting the log transform of number of selected items. Table shows unstandardized coefficients. ( $R^2=0.260$ )**

	Unstandardized		Standardized		
	estimate	std. err.	estimate	t	P
(constant)	2.865	0.294		9.731	0.000
SysEff	-0.205	0.088	-0.271	-2.321	0.024
Age	-0.019	0.005	-0.407	-3.489	0.001

Higher system efficiency and higher age lead to a lower number of selected items. As Figure 17 also showed, system efficiency acts as a mediator between the experimental treatment and the number of selected items.

A regression analysis on the number of removed items showed that only age had any predictive value on the log transform of number of removed items as can be seen from Table 13. Participants with a higher age removed slightly less items than those that were younger.

**Table 13 - Predicting the log transform of number of removed items. Table shows unstandardized coefficients. ( $R^2=0.164$ )**

	Unstandardized		Standardized		
	estimate	std. err.	estimate	t	P
(constant)	3.061	0.405		7.562	0.000
Age	-0.039	0.012	-0.425	-3.257	0.002

Finally, regression analysis showed that only education was able to predict the percentage of items that was executed and partly executed by our participants. Higher education thus has a negative influence on the percentage of executed items as can be seen from the coefficients in Table 14

**Table 14 – Predicting the amount of executed and partly executed items. Table shows unstandardized coefficients. ( $R^2=0.104$ )**

	Unstandardized		Standardized		
	estimate	std. err.	estimate	t	P
(constant)	107.485	20.834		5.159	0.000
Education	-8.132	3.786	-0.365	-2.148	0.040



### 3.2.2 Discussion

The results above showed that adding the 22 participants from study 2 to our participant sample of the first study did not strengthen the model from the first study. Instead we found that the effect of experimental condition on system efficiency became very marginally significant and that effects of perceived control and general user experience disappeared completely.

These results thus suggest that both groups of participants were not as similar as we anticipated. Participants' choice behavior in Figure 16 shows that the distribution of number of selected items is fairly different than that over the first study indicating that the slight differences in experimental design have had their effect on user behavior.

Additionally, the means by which participants were recruited for both studies differed slightly and could therefore also have introduced noise to the combined dataset. In the first study, participants were recruited mainly from an online database for Internet research, whereas study 2 relied much more on newspaper advertisements appealing to people interested in improving their health.

These results therefore suggest that it would be better to redo the first study with a larger and more homogeneous participant sample instead of combining them with a subsample of the participants from our second study.

### 3.2.3 Temporal effects of the Lifestyle Coach

We will now look beyond the first stage (week) of this study and discuss the results from, and differences between the second and third week of this study. However since the participant numbers were too low (see Figure 15) to produce quantitative results that also had external validity, we will not go into detail and mostly describe trends that could help in designing future research.

Figure 18 shows how user experience of the Lifestyle Coach system changed over time. Measures were taken after the first and third week, following an equal number of interaction sessions with the system. A paired t-test showed that only changes in choice satisfaction ( $t(11)=-3.563$ ,  $p=0.004$ ), system satisfaction ( $t(11)=2.803$ ,  $p=0.017$ ) and general user experience ( $t(11)=1.836$ ,  $p=0.094$ ) appeared to be significant and marginally significant.

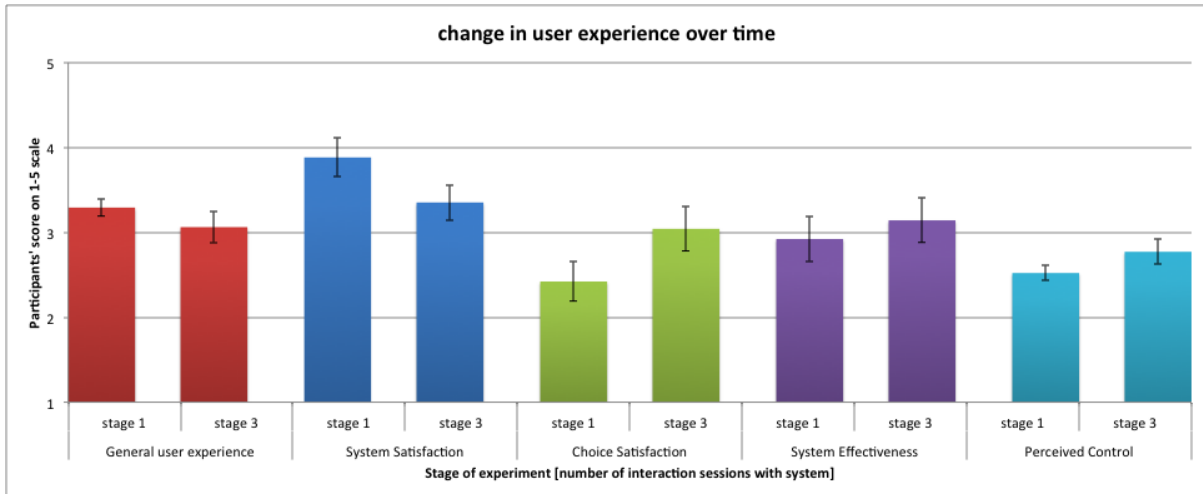


Figure 18 - Graphical display of the change of the five difference user experience measures over time. (N=66 at stage 1 and N=19 at stage 3)

The decrease in system satisfaction indicated that participants grew less fond of the recommendations they received from the Lifestyle Coach as the experiment progressed, possibly originating from the fact that their style of decision-making did not match that of the recommender system. The increase of choice satisfaction over the course of the experiment showed that either participants gained more experience in finding relevant items (i.e. they knew better how to use the system to find what they wanted) or grew fonder of the items they (possibly) repeatedly chose.

Looking at Figure 19 we see an unexpectedly high number of removed items in the attribute condition for week 2 and 3. Following our theory on the decision making strategies being employed by our participants we expected this effect to reverse; increase in removed items for the random condition instead of the attribute one. It has to be noted however that participant numbers for weeks 1, 2 and 3 were 64, 16 and 18 respectively, making this data extremely vulnerable to random noise.

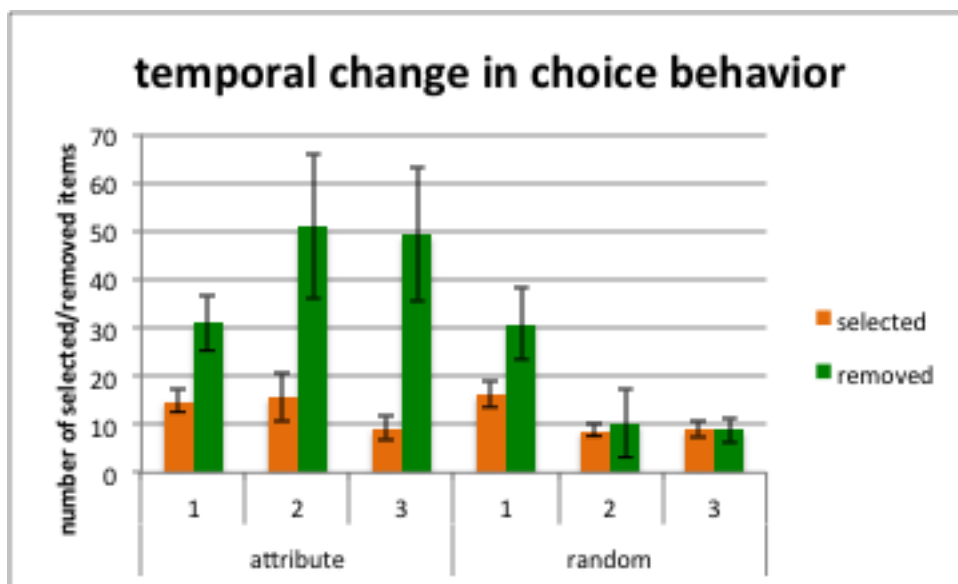


Figure 19 - change of choice behavior over time; 1,2 & 3 indicate week numbers of the experiment. (N=66 at stage 1, N=16 at stage 2 and N=19 at stage 3)

If we finally check the number of executed items over time (Figure 20), we see that the percentage of fully as well as fully and partly executed items seemed to drop faster over time in the random condition compared to the attribute-based condition. This might be due to a higher personal relevance of the chosen items in the attribute condition as can be seen by the choice satisfaction measure in Figure 18. However, also the results in Figure 20 were based on very low participant numbers and therefore not reliable and in dire need of more research.

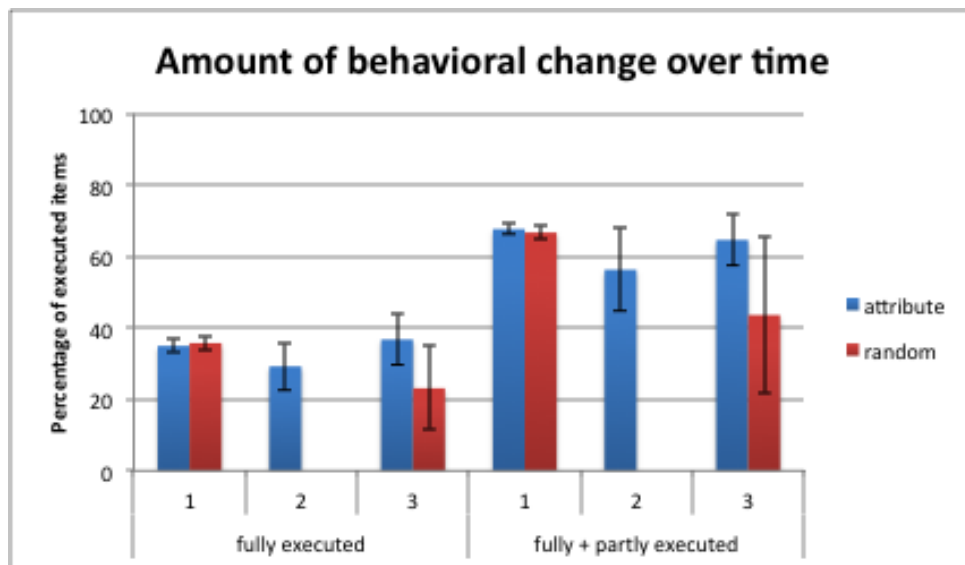


Figure 20 - Graphical representation of the difference in behavioral change per week. (N=66 at stage 1, N=16 at stage 2 and N=19 at stage 3)

Finally we present data on the participants that completed all three stage of the experiment in Table 15 below. We see that predominantly women completed the experiment, additionally also fewer participants in the attribute condition dropped out.

Table 15 - Demographics of the participants that completed all three stages of study 2.

Variable		Total		Attribute condition		Random condition	
		Statistic	s.e.	Statistic	s.e.	Statistic	s.e.
Participants	N	19		14		5	
Gender	Male	2 (11%)		1 (7%)		1 (20%)	
	Female	17 (89%)		13 (93%)		4 (80%)	
Age		35.78	3.025	37.86	3.775	30.00	3.924
BMI		24.27	0.695	24.26	0.436	24.29	2.557
Education	> 25	5 (26%)		4 (29%)		1 (20%)	
	None	0		0		0	
	VMBO	1		1		0	
	HAVO	2		2		0	
	VWO	0		0		0	
	MBO	2		1		1	
	HBO	8		7		1	
	WO	6		3		3	

## **4. General discussion**

We will now try to put our findings in perspective by going back to our original hypotheses and discuss our results in logical order to get to a final conclusion on our main research question, namely: did tailoring really help our participants in their decision-making process on their journey to better health?

### **4.1 Main findings**

Following up on the work of Knijnenburg et al. (2011) – that showed that accounting for personal characteristics results in increased user experience – we reasoned that comparing our tailored recommender system against a non-tailored recommender system would favor the former. We found that using the random recommender system resulted in lower system efficiency in both the first and second study compared to the attribute system. Results were stronger in our first study since our participant sample was more homogeneous. Although we anticipated choice satisfaction to be influenced by the type of recommender being used, we found that it is actually the system efficiency that is different between our systems. We can therefore partly confirm our first hypothesis: use of a random recommender system will lead to a lower perceived system efficiency compared to an attribute-based recommender system. In other words, tailoring of the information lead participants to evaluate the system as being more efficient.

As Brug et al. (1999) showed, personally relevant information is read more and mentally processed more thoroughly. Our second hypothesis therefore stated that tailoring recommendations would lead to increased (choice) satisfaction and in turn would lead our users to select more items from the lifestyle coach. We also anticipated the effect of tailoring to work directly on the number of selected items without the mediation of satisfaction in our second hypothesis.

As we found with the first hypothesis, there was no effect of the type of recommender being used on the (choice) satisfaction of our participants. However, we did find an effect on the system efficiency. We also found a direct effect of the type of recommender system on the number of selected items in our results from the first study. This effect was however opposite to what we hypothesized. Participants in the random condition chose more items compared to those in the attribute condition; tailoring did not result in an increase of selected items. We therefore discard our second hypothesis.

Besides the direct effect of the type of recommender system on number of selected items, we also found a mediated effect through the system efficiency measure. This was however found to be very small since most of the variance of the number of selected items was already explained by the before mentioned direct effect.

Contrary to the number of selected items, we find that number of removed items is influenced by multiple user experience items. General user experience, system efficiency and perceived control all have an effect on the number of removed items.

Through our third hypothesis we anticipated larger sets of selected items to result in more actual behavior. Neither our first study nor the results of our first and second study combined show evidence for this claim. Although participants executed on average more than half of the measures they chose, we have to conclude that no evidence can be shown to prove a possible relation between the number of items selected and the number of items executed. The main cause for this was the exclusion criterion of more than 10 selected items that we introduced in the first study. The range of selected items was too small to find significant effects here. An exploratory analysis of the full dataset did show a negative correlation of the number of selected items on the number of items that were executed. The validity of this finding is however debatable considering the fact that these participants set unrealistically large goals.

Altogether we cannot confirm our third hypothesis based on the findings from our studies.

Although many other studies find clear differences in effect size, caused by gender, education, ethnicity and other demographic variables we do not find unified evidence for this. However we found some demographic variables to correlate with our main dependent variables. The results from the first study show that age (positively) and being female (negatively) influenced the number of removed items. Education on the other hand was the only variable that positively correlated with number of selected items.

According to the combined results from the first and second study showed that a higher education caused the evaluation of system efficiency and the percentage of (partly) executed items to be lower. Furthermore, higher age also led to lower numbers of both selected and removed items.

The only effect we found, related to participants' B.M.I., was that higher B.M.I. led participants to select items that had on average a higher A.W.H. value. Effect sizes for age were found to be very small in all our regression analyses. Regarding our fourth hypothesis we can say that females tend to remove more items when working with the lifestyle coach while being higher educated leads to selecting more items. From the results of the first and second study combined we know that higher educated people execute fewer items on the other hand and also evaluate system efficiency lower.

We turn to our second study in order to discuss our fifth and final hypothesis that stated that the decrease in effect size of our intervention will be less for participants that received tailored recommendations compared to those that received non-tailored recommendations. The fact that dropout rates were so high for our second study decreases the generalizability of our results enormously. The data that we have on the other hand shows some interesting phenomena. When looking at the user experience evaluation and the choice behavior over time, we see the choice satisfaction increase while satisfaction with the system decreases. At the same time we also see an enormous increase in number of removed items in the attribute condition and a decrease in number of selected items. Since the majority of the participants

still enrolled in the experiment in stage 3 are from the attribute condition, we assume their evaluation to drive the observed significant effects.

Obviously there are many reasons for explaining the observed data, however it seems most likely that (some) participants simply changed their searching strategy from using the attribute-weights to simply removing unwanted items or combining these strategies<sup>8</sup>. The latter would off-course demand more effort, possibly causing system satisfaction to drop but will eventually also lead to finding appropriate items and thus maintaining or even increasing choice satisfaction.

Similar to findings regarding the other hypotheses in this work, it seems likely that our attribute system did not offer all of its participants the same amount of help in their decision-making process as we anticipated.

Given the results from this study it does seem likely that the attribute system performed better over time – dropout rates were lower and higher percentages of selected measures were executed – compared to the random system. Participant numbers were however too low to strengthen this claim with empirical findings. We can therefore only state that our fifth hypothesis is likely to be true, but we cannot confirm this.

The results presented above might suggest that the introduction of recommender systems into the healthcare domain was not particularly successful. Shifting focus from our hypotheses to the bigger picture on the other hand reveals a story that is more promising.

Working backwards through our model, we see that both the participants that worked with the random recommender as well as those that worked with the attribute-based recommender executed on average more than half of their chosen measures. If we add to this the amount of partially executed items this number rises even further. All of the executed items contributed to achieving a better health in our participants. The Lifestyle Coach thus proved to be a worthy competitor in the field of computerized intervention tools.

We also found that participants did not use the Lifestyle Coach as we expected them to, they set unrealistic goals by selecting very large sets of items. Our data on user experience showed that participants did not experience this as much of a problem themselves, they evaluated most user experience items significantly above average. We suspect that the large numbers of items that participants selected and removed was also due to their curiosity for the healthy measures in the Lifestyle Coach system. Especially the data on participants in the attribute condition showed that participants did not make as much use of the attribute buttons but rather removed items in order to browse through the list of healthy recommendations. The lack of a significant difference between the numbers of removed items in both types of recommenders suggests that regardless of the type of recommender, the most popular way of browsing through the list of healthy measures seemed to be simply by using the remove buttons.

---

<sup>8</sup> The weights of the attributes were reset every week. Participants therefore had to remember their preferences if they wanted the same recommendations as the week before. Results showed that the use of the attribute buttons did not decrease over time, suggesting that participants used both the attribute buttons and the ‘remove-strategy’ in order to find their desired set of items.

In general this suggests that participants did not like to make their choices for healthy lifestyle improvements by assigning weights to explicit attributes. It seems that simply clicking their way through the database of healthy measures – by means of selecting and removing items – was favored over the conscious and well-balanced decision making process that is facilitated by the attribute-based recommender. The fact that we did find significant relations between some of our experimental measures suggests that at least some participants did make their decisions in this well-balanced conscious way. Nonetheless, our data suggest that most participants could have benefitted more from a system that did not make the decision making process as explicit as the attribute system but rather implicitly learns from the participant's browsing behavior and implicitly recommends new items based on this. Recalling the introduction on recommender systems from paragraph 1.3, a collaborative-based recommender system would be the most logical next step in this process. A collaborative recommender learns from the browsing behavior of its user and recommends accordingly. Since our participants seemed to prefer a browsing strategy to an explicit decision making process it seems likely that a collaborative-based recommender would fit the needs of our participants in this particular domain better.

A final interesting finding of this study regards the percentage of executed items. We found this number to be equal for both types of recommenders, however the number of selected items was found to be significantly higher in the random condition. The latter implies that the random recommender was more successful in improving the health of our participants. It is hard to determine whether this difference is actually significant since we were not able to analyze the total percentage of A.W.H. at the items level. In other words, we cannot say that the random condition resulted in more healthy behavior compared to the attribute condition, even though more items were executed since we do not know the total A.W.H. value of these particular items.

What this information does tell us on the other hand is that recommendations in the attribute condition were perhaps not diverse enough, thereby preventing the participant from finding enough personally relevant items.

## **4.2 Implications**

If we recall the rationale for our first hypothesis – different users call for different interaction methods (Knijnenburg et al., 2011) – we might be able to explain our findings in part. The fact that we find evidence for the positive effect of tailoring on user experience only in a subset of our population might suggest that prompting people with the 'wrong' type of recommender system is just as good as giving them random recommendations. The attribute-based system that we designed for the purpose of generating tailored recommendations therefore probably did not address the needs of all of the participants in that treatment, hence the lack of significant findings. Possible use of other types of recommender systems that better suit the needs of our participants such as a collaborative recommender could thus result in more significant differences between our two treatments.

It can be debated however whether the needs of all of our participants were similar. The higher average number of A.W.H. per selected item for people with a higher B.M.I. shows

that at least the goals of our participants differed. Performing this study again on a sample of only overweight (B.M.I. > 25) people could therefore result in different results. The majority of participants in both our first and second study had a B.M.I. that was lower than 25. If needs did indeed differ between people with lower and higher scores on B.M.I. our participant sample was most likely not sufficiently diverse on this variable to reveal differences.

Another explanation for our lack of significant findings on participants' user experience could lie in the quality of our database of healthy recommendations and the attributes on which they were scored. Although only future research can verify this, we think that because of the effort and care that has been taken in constructing this dataset and the expert knowledge that has been put into it, that our first explanation is more likely.

Even if attributes are chosen perfectly, it is still possible that an attribute-based recommender system is not the best way to help every individual in his or her decision-making process. Future research should therefore also investigate the effects of collaborative or hybrid recommenders in order to judge the possible added value of recommender systems in healthcare interventions.

We anticipated the effect of tailoring to work directly or by mediation of satisfaction on the number of selected items in our second hypothesis. Our first study initially showed no such effect. However on inspection of the subset of the population that selected 10 items or less we did find a direct positive effect of tailoring on the number of selected items. As we discussed in the first hypothesis, this could be due to our tailored system that possibly did not address the entire group of participants in the tailored experimental condition. Again, future research with a different kind of tailored recommender could possibly provide more insight on this issue.

Through our third hypothesis we anticipated larger sets of chosen items to result in more behavioral change. Behavioral change was in this case only defined by the self-report of amount of executed measures. Although prior to this research we planned to use validated questionnaires to assess the change of health in our participants' over the course of the experiment (study 2), high dropout rates prevented the collection of adequate data.

The behavioral change that we measured through self-reports of amount of executed items, turned out to be negatively affected by the number of selected items in our first study. In the analysis of our smaller subset of the population of study one, we did not find this negative correlation. This might suggest that there is some kind of optimum in the amount of chosen measures per person. Since many of our participants in the first study chose too many items, they were not able to execute all of them, leading to this negative correlation between number of chosen items and amount of behavioral change. Since the set of selected measures all together symbolizes the total end-goal per participant, we conclude that the further away one puts his end goal, the less likely it is he will achieve it. Despite the fact that the end-goal was divided into smaller steps (the separate healthy measures that a participant selected), choosing an unrealistic end-goal actually seemed to work counterproductive in achieving a better health.



The reasoning above might also suggest that more guidance or coaching is needed to help participants decide what their end-goal should be. Apparently the weekly recommended amount (A.W.H.) measure that was offered in the lifestyle coach did not work as we anticipated. The latter is confirmed by the fact that average scores for total percentage of chosen A.W.H. were significantly higher than 100. Selecting more items is not necessarily better.

The high dropout rates observed in our second study in particular suggests that our software did not work optimally. A more clear description of, and more emphasis on the experimental task could possibly improve results of our study greatly. The bottom line of our findings is that we perhaps did not tailor our interventions enough to suit all of our participants. Tailoring can be done based on the information that our recommender system provides and also on the type of recommender system being used for the intervention. Additionally more specific targeting of groups with similar needs (perhaps based on B.M.I.) has the potential to reveal more consistent findings.

#### **4.3 Conclusion & future research**

This work shows that we cannot irrefutably claim that recommender systems will make it easier for people to decide how they can live a healthier life. What this work does show on the other hand is that not all hope is lost. By analyzing the way in which participants used our systems, we found that most participants do not prefer to make lifestyle decisions as consciously as we anticipated. Instead they prefer to browse through the database in search for those items that are relevant to them. There still is room for facilitating this process by means of recommender systems, however this has to be done less explicit, possibly by means of collaborative filtering.

Besides using different methods for generating recommendations, future research should also focus on decreasing the dropout rate of this type of interventions. In the case of this work, one of the core reasons for dropout (study 2) was the high pace and strict nature of our software. Because we wanted to keep the number of interactions and the periods in between constant for all of our participants, we chose a weekly cycle that only facilitated one set interact interaction with the recommender system during the weekend. From user feedback we learned however that a week was in some cases too short. Additionally, during weekends it was easy to forget to use the lifestyle coach since people were often occupied otherwise.

To summarize, future systems should be designed more flexible and allow the participant to use it at his own pace. Furthermore, more emphasis could be put on the fact that the system can also be used on a mobile device, allowing for more interaction opportunities.

Another interesting point of focus for future research is the goal setting part in the recommender system. In our research, the experimental task was apparently not crystal clear to our participants. Determining the earlier discussed optimum in amount of measures to be selected per person can greatly increase the amount of executed items.

Additionally, future research should also be designed in a more longitudinal way in order to allow for measuring actual behavioral changes. Since changes in behavior typically occur over a timespan of weeks or even months, this research was probably too short to be able to measure them anyway.

Although we might have lost this fight in our war against overweight and obesity, the battle is far from over.

## 5. Acknowledgements

I dedicate this chapter to all those that helped in realizing this project, without their help this project would most likely not have been possible.

First off all I would like to thank Roland Friele from the NIVEL institute for granting me the opportunity to earn a contract at NIVEL, allowing me to carry out this research. Since the subject of this study has both my personal as well as my professional interest, I am very grateful to have had the opportunity to dedicate my master thesis to this project and the fight against obesity in general.

Secondly, I would like to thank Petr Kosnar, who helped a very great deal in building the Lifestyle coach system. Without his knowledge of PHP and the cake framework I would have never been able to build the Lifestyle Coach in its current form.

Furthermore I thank Karin van Beek, Jacqueline Tol, Claire Verheul, Cindy Veenhof and Daniël Bossen for offering their expert knowledge in constructing the database with healthy measures that served as the basis for the Lifestyle Coach system.

Finally I thank Martijn Willemsen and Cindy Veenhof for not only supervising this project in a very enthusiastic, personal and fun way but also for believing in the goal and potential of this project.

## 6. References

- Adomavicius, G., & Tuzhilin, a. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749.
- Ainsworth, B. E., Haskell, W. L., Whitt, M. C., Irwin, M. L., Swartz, a M., Strath, S. J., O'Brien, W. L., et al. (2000). Compendium of physical activities: an update of activity codes and MET intensities. *Medicine and science in sports and exercise*, 32(9 Suppl), S498–504.
- Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, (50), 171–211.
- Bandura, A. (1998). Health promotion from the perspective of social cognitive theory. *Psychology & Health*, 13(4), 623–649.
- Berland, G. K., Elliott, M. N., Morales, L. S., Algazy, J. I., Kravitz, R. L., Broder, M. S., Kanouse, D. E., et al. (2001). Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish. *Journal for the American Medical Association*, 285(20), 2612–2621.
- Brug, J., Campbell, M., & van Assema, P. (1999). The application and impact of computer-generated personalized nutrition education: a review of the literature. *Patient education and counseling*, 36(2), 145–56.
- Brug, J., Oenema, A., & Campbell, M. (2003). Past, present, and future of computer-tailored nutrition education. *The American journal of clinical nutrition*, 77(4 Suppl), 1028S–1034S.
- Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. *CHI '88*, 213–218.
- Downs, J. S., Loewenstein, G., & Wisdom, J. (2009). Strategies for Promoting Healthier Food Choices. *American Economic Review*, 99(2), 159–164.
- Eysenbach, G. (2001). What is e-health? *Journal of medical Internet research*, 3(2), E20.
- Glanz, K., Basil, M., Maibach, E., Goldberg, J., & Snyder, D. (1998). Why Americans eat what they do: Taste, nutrition, cost, convenience, and weight control concerns as influence on food consumption. *Journal of the American Dietetic association*, 98(10), 1118–1126.
- Godin, G., & Kok, G. (1996). The theory of planned behavior: a review of its applications to health-related behaviors. *American Journal of Health Promotion*, 11(2), 87–98.
- Kaptein, M., De Ruyter, B., Markopoulos, P., & Aarts, E. (2012). Adaptive Persuasive Systems. *ACM Transactions on Interactive Intelligent Systems*, 2(2), 1–25.

- Knijnenburg, B.P. (2009). Adaptive advice: adapting a recommender system for energy saving behaviors to personal differences in decision making (2009).
- Knijnenburg, B. P., Willemsen, M. C., & Reijmer, N. J. M. (2011). Each to His Own : How Different Users Call for Different Interaction Methods in Recommender Systems. *Methods*.
- Kreuter, M. W., Oswald, D. L., Bull, F. C., & Clark, E. M. (2000a). Are tailored health education materials always more effective than non-tailored materials? *Health education research*, 15(3), 305–15.
- Kreuter, M. W., Oswald, D. L., Bull, F. C., & Clark, E. M. (2000b). Are tailored health education materials always more effective than non-tailored materials? *Health education research*, 15(3), 305–15.
- Miller, N. E., & Dollard, J. (1941). *Social learning and imitation* (p. 341). New Haven, CT, US: Yale University Press.
- Neumark-Sztainer, D., Story, M., Perry, C., & Casey, M. A. (1999). Factors influencing food choices of adolescents: findings from focus-group discussions with adolescents. *Journal of Nutrition Education and Behavior*, 99(8), 929–937.
- Norman, G. J., Zabinski, M. F., Adams, M. A., Rosenberg, D. E., Yaroch, A. L., & Atienza, A. A. (2007). A Review of eHealth Interventions for Physical Activity and Dietary Behavior Change. *American Journal of Preventive Medicine*, 33(4), 336–345.
- Oenema, A., Tan, F., & Brug, J. (2005). Short-term efficacy of a web-based computer-tailored nutrition intervention: main effects and mediators. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine*, 29(1), 54–63.
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3), 879–891.
- Prochaska, J. O., & DiClemente, C. C. (1982). Transtheoretical therapy: Towards a more integrative model of change. *Psychotherapy: Theory, Research & Practice.*, 19(3), 276–288.
- Reijmer, N. J. M. (2011). Improving energy saving decisions by matching recommender type with domain knowledge and mindset (November, 2011).
- Stevenson, C., Doherty, G., Muldoon, O. T., Barnett, J., & Trew, K. (2007). Barriers to healthy eating in adolescence, 1–37.
- Vandelanotte, C., De Bourdeaudhuij, I., Sallis, J. F., Spittaels, H., & Brug, J. (2005). Efficacy of sequential or simultaneous interactive computer-tailored interventions for increasing physical activity and decreasing fat intake. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine*, 29(2), 138–46.

- Xiao, B., & Benbasat, I. (2007). E-Commerce product recommendation agents: use, characteristics, and impact. *MIS Quarterly*, *31*(1), 137–209.
- van Assema, P., Brug, J., Ronda, G., & Steenhuis, I. (2001). The relative validity of a short Dutch questionnaire as a means to categorize adults and adolescents to total and saturated fat intake. *Journal of human nutrition and dietetics : the official journal of the British Dietetic Association*, *14*(5), 377–90.
- van Assema, P., Brug, J., Ronda, G., Steenhuis, I., & Oenema, A. (2002). A short dutch questionnaire to measure fruit and vegetable intake: relavite validity among adults and adolescents. *Nutrition and Health*, *16*(2), 85–106.

## 7. Appendices

### 7.1 UX questionnaire

VARIABLE NAME	QUESTION
<i>general user experience</i>	Wat vond je van het systeem dat je zojuist hebt gebruikt? [1-5]
UX_QUIS_1	Verschrikkelijk --- Geweldig
UX_QUIS_2	Ingewikkeld --- Gemakkelijk
UX_QUIS_3	Frustrerend --- Bevredigend
UX_QUIS_4	Saai --- Opwekkend
UX_QUIS_5	Star --- Flexibel
<i>system satisfaction</i>	[oneens --- eens]
UX_SysSat_1	De adviezen die de lifestyle coach voorstelde/aanbeveelde bevielen me.
UX_SysSat_2	De adviezen die de lifestyle coach voorstelde/aanbeveelde pasten bij mijn voorkeuren.
UX_SysSat_3	De adviezen die de lifestyle coach voorstelde/aanbeveelde waren relevant.
UX_SysSat_4	De adviezen die de lifestyle coach voorstelde/aanbeveelde bevatte te veel slechte adviezen.
UX_SysSat_5	De adviezen die de lifestyle coach voorstelde/aanbeveelde waren goed gekozen.
<i>choice satisfaction</i>	[oneens --- eens]
UX_ChoSat_1	De adviezen die ik gekozen heb bevielen me.
UX_ChoSat_2	Ik ben blij met de door mij gekozen adviezen.
UX_ChoSat_3	Het kijken naar de door mij gekozen adviezen beviel mij.
UX_ChoSat_4	Ik vond het tijdverspilling om alle adviezen te bekijken.
UX_ChoSat_5	De door mij gekozen adviezen paste bij mijn voorkeuren.
UX_ChoSat_6	Ik zou zelf enkele adviezen kunnen bedenken die beter zijn dan de adviezen die ik gekozen heb.
<i>system effectiveness and fun</i>	[oneens --- eens]
UX_SysEff_1	Ik had plezier tijdens het werken met de lifestyle coach.
UX_SysEff_2	Ik zou de lifestyle coach ook aan andere mensen aanbevelen.
UX_SysEff_3	Het werken met de lifestyle coach was een fijne ervaring.
UX_SysEff_4	De lifestyle coach is waardeloos.
<i>perceived control</i>	[oneens --- eens]
UX_PercContr_1	Ik had beperkte controle over de manier waarop de lifestyle coach adviezen voorstelde.
UX_PercContr_2	De manier waarop de lifestyle coach adviezen voorstelde was erg strikt.
UX_PercContr_3	Vergeleken met de manier waarop ik normaal gezonde veranderingen in mijn levensstijl kies was de lifestyle coach erg beperkt.
UX_PercContr_4	Ik zou graag meer controle hebben gehad in de lifestyle coach.
UX_PercContr_5	De manier waarop de lifestyle coach gezonde adviezen voorstelt is te vergelijken met hoe ik het zelf zou doen.

## **7.2 Fruit, vegetable and fat intake questionnaire**

### **GROENTEN**

1. Hoeveel dagen per week eet u gewoonlijk gekookte, gebakken, gestoomde of anders verhitte groenten?
2. Op een dag dat u groenten eet, hoeveel opscheplepels eet u dan gewoonlijk van gekookte, gebakken, gestoomde of anders verhitte groenten?
3. Hoeveel dagen per week eet u gewoonlijk sla of rauwkost?
4. Op een dag dat u groenten eet, hoeveel opscheplepels eet u dan gewoonlijk van sla of rauwkost?

### **FRUIT**

5. Hoeveel dagen per week eet u gewoonlijk citrusfruit (sinaasappel, grapefruit, mandarijn)?
6. Op een dag dat u fruit eet, hoeveel stuks eet u dan gewoonlijk van citrusfruit?
7. Hoeveel dagen per week eet u gewoonlijk ander fruit?
8. Op een dag dat u fruit eet, hoeveel stuks eet u dan gewoonlijk van ander fruit?

### **VRUCHTENSAP**

9. Hoeveel dagen per week drinkt u gewoonlijk ongezoet vruchtensap, bijvoorbeeld sinaasappelsap of druivensap, vers of uit pak?
10. Op een dag dat u ongezoet vruchtensap drinkt, hoeveel glazen drinkt u dan gewoonlijk?

## **VET-VRAGENLIJST**

### **ZUIVEL**

1. Hoeveel dagen per week drinkt u meestal melk of karnemelk?
2. Op een dag dat u melk of karnemelk drinkt, hoeveel glazen drinkt u meestal per dag?
3. Welke soort melk drinkt u meestal?
4. Hoeveel dagen per week drinkt u meestal chocolademelk?
5. Op een dag dat u chocolademelk drinkt, hoeveel glazen drinkt u meestal per dag?
6. Welke soort chocolademelk drinkt u meestal?
7. Hoeveel dagen per week eet u meestal yoghurt, vruchtenyoghurt of biogarde?
8. Op een dag dat u yoghurt, vruchtenyoghurt of biogarde eet, hoeveel schaaltes eet u meestal per dag?
9. Welke soort yoghurt, vruchtenyoghurt of biogarde eet u meestal?
10. Hoeveel dagen per week eet u meestal vla, pudding of pap?
11. Op een dag dat u vla, pudding of pap eet, hoeveel schaaltes eet u meestal per dag?



## BROOD EN BROODBELEG

12. Hoeveel sneetjes brood, broodjes, crackers of beschuiten eet u meestal per dag?
13. Hoeveel sneetjes belegt u meestal met kaas?
14. Welke soort kaas doet u meestal op uw brood?
15. Hoeveel sneetjes belegt u meestal met vleeswaren?
16. Welke soort vleeswaren doet u meestal op uw brood (zo precies mogelijk opschrijven)?
17. Hoeveel sneetjes belegt u meestal met chocoladehagelslag, chocoladevlokken of chocoladepasta?
18. Welke soort boter doet u meestal op uw brood?

## WARME MAALTIJD

19. Hoe vaak eet u rookworst, verse worst of saucijsjes bij de warme maaltijd?
20. Hoe vaak eet u half om half gehakt bij de warme maaltijd?
21. Hoe vaak eet u speklappen bij de warme maaltijd?
22. Hoe vaak eet u kaas (geraspt, blokjes in de maaltijd of als kaassoufflé) bij de warme maaltijd?
23. Hoe vaak gebruikt u jus bij de warme maaltijd?
24. Hoeveel lepels neemt u als u jus gebruikt?
25. Wat voor soort lepel gebruikt u om de jus op te scheppen?

## TUSSENDOORTJES

26. Hoe vaak eet u naast de maaltijden snacks, zoals patat, kroketten, slaatjes enzovoort?
27. Hoe vaak eet u tussendoor pinda's of nootjes?
28. Hoe vaak eet u tussendoor chips, blokjes kaas, buitenlandse kaas of worst?
29. Hoe vaak eet u tussendoor gebak, cake of grote koeken?
30. Hoe vaak eet u tussendoor candy-bars, zoals Nuts, Mars, Snickers, enzovoort?
31. Hoe vaak eet u tussendoor chocolade?
32. Hoe vaak eet u tussendoor biscuit of rozijnenbiscuit?
33. Op een dag dat u biscuitjes eet, hoeveel eet u dan meestal?
34. Hoe vaak eet u tussendoor andere koekjes?
35. Op een dag dat u andere koekjes eet, hoeveel eet u dan meestal?

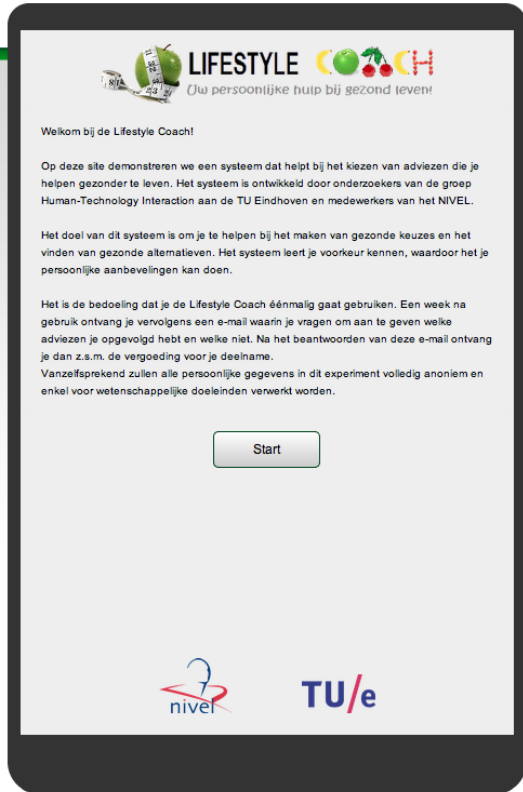
### 7.3 SQUASH questionnaire

Neem in uw gedachten een normale week in de afgelopen maanden. Wilt u aangeven **hoeveel dagen per week** u de onderstaande activiteiten verrichtte, hoeveel minuten u daar dan **gemiddeld** op zo'n dag mee bezig was en hoe inspannend deze activiteit was?

<b>WOON-WERK/SCHOOL VERKEER (heen en terug)</b>	<b>aantal dagen per week</b>	<b>gemiddelde tijd per dag</b>	<b>inspanning (omcirkelen a.u.b.)</b>
Lopen van/naar werk of school	<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten	langzaam/gemiddeld/snel
Fietsen van/naar werk of school	<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten	langzaam/gemiddeld/snel
Niet van toepassing	<input type="text"/>		
<b>LICHAMELIJKE ACTIVITEIT OP WERK EN SCHOOL</b>			<b>gemiddelde tijd per week</b>
Licht en matig inspannend werk (zittend/staand werk, met af en toe lopen, zoals bureauwerk of lopend werk met lichte lasten)			<input type="text"/> <input type="text"/> uur <input type="text"/> <input type="text"/> minuten
Zwaar inspannend werk (lopend werk, waarbij regelmatig zware dingen moeten worden opgetild)			<input type="text"/> <input type="text"/> uur <input type="text"/> <input type="text"/> minuten
Niet van toepassing			<input type="text"/>
<b>HUISHOUDELIJKE ACTIVITEITEN</b>		<b>aantal dagen per week</b>	<b>gemiddelde tijd per dag</b>
Licht en matig inspannend huishoudelijk werk (staand werk, zoals koken, afwassen, strijken, kind eten geven/in bad doen en lopend werk, zoals stofzuigen, boodschappen doen)		<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten
Zwaar inspannend huishoudelijk werk (vloer schrobben, tapijt uitkloppen, met zware boodschappen lopen)		<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten
Niet van toepassing		<input type="text"/>	
<b>VRIJE TIJD</b>	<b>aantal dagen per week</b>	<b>gemiddelde tijd per dag</b>	<b>inspanning (omcirkel a.u.b.)</b>
Wandelen	<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten	Langzaam/gemiddeld/snel
Fietsen	<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten	Langzaam/gemiddeld/snel
Tuinieren	<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten	Licht/gemiddeld/zwaar
Klussen/Doe-het-zelven	<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten	Licht/gemiddeld/zwaar
Sporten (Hier maximaal 4 opschrijven) <i>bijv.: tennis, handbal, gymnastiek, fitness, schaatsen, zwemmen</i>			
1. ....	<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten	Licht/gemiddeld/zwaar
2. ....	<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten	Licht/gemiddeld/zwaar
3. ....	<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten	licht/gemiddeld/zwaar
4. ....	<input type="text"/> dagen	<input type="text"/> uur <input type="text"/> <input type="text"/> minuten	licht/gemiddeld/zwaar
<b>TOTAAL</b>			
Op gemiddeld hoeveel dagen per week bent u, alles bijelkaar opgeteld, tenminste een half uur bezig met fietsen, klussen, tuinieren of sporten?			<input type="text"/> dagen per week

## 7.4 Screenshots of the Lifestyle Coach software

Below we demonstrate the interface of the lifestyle coach in logical order. Descriptions and/or translations are printed alongside the screenshots.

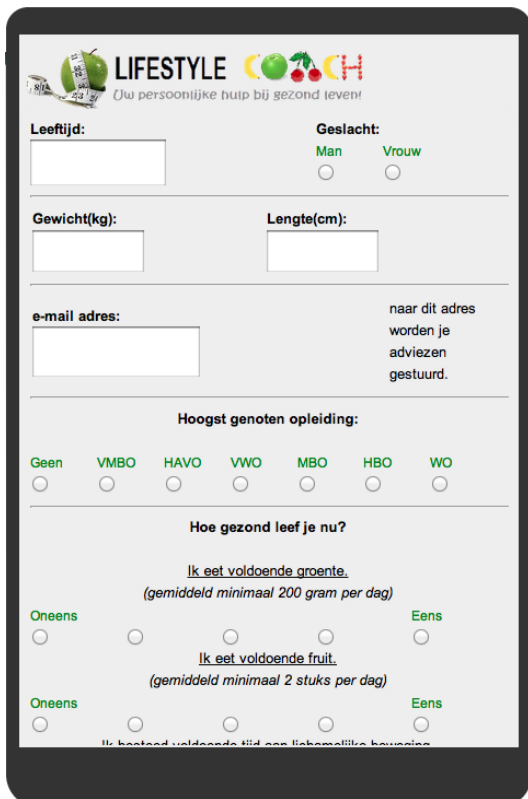


### *Welcome screen*

Participants are explained that the Lifestyle Coach software will help them in selecting measures that can improve their health.

Furthermore they are informed of the one-week cycle of the system and the fact that they will only receive their monetary reward if they responded to the automatically generated email by the lifestyle coach one week later.

Finally the text emphasizes that all personal data gathered in this experiment will be used exclusively for this experiment and will be processed anonymously and treated with great care.



### *Demographic questionnaire*

The email address that participants fill in here is used solely for e-mailing them their personal list of selected items after interaction with the system.



*Instructions 1/5*

“Lifestyle coach will help you in choosing healthy recommendations that fit your personal lifestyle.”



*Instructions 2/5*

“The first tab allows you to input your personal preferences by indicating what you value important and what you value less important.”



### Instructions 3/5

“In the second tab in the Lifestyle Coach will present you with 10 healthy recommendations according to the preferences you set on the previous tab (1).

At that point you will be able to indicate which recommendations you would like to add to your to-do list and which you would like to remove.

In case you are not satisfied with the presented recommendations you can always return to the first tab to modify your preferences.”



### Instructions 4/5

”The third tab allows you to inspect your personally composed to-do list and will indicate the healthiness score for all items combined.

Healthiness of each item is expressed as a percentage of the total recommended weekly amount (A.W.H.) of healthy recommendations. A.W.H.-scores are merely indicators of each items contribution to a better health and not strict guidelines. Since activities that are already in your daily routine also contribute to a better health it is not necessary to reach 100% on the A.W.H. score.”



### Instructions 4/5

“At any point during interaction with the lifestyle coach you can decide to go either one tab back to adjust your preferences or one tab forward to inspect your to-do list so far.

It’s up to you to decide how many items you select, however try not to select more items than you can execute in one week.”



### First tab of the recommender system

Users can indicate which attributes they value more important or less important by clicking on ‘+’ or ‘-’ respectively.

Additionally they can indicate whether they prefer to receive more recommendations regarding food, exercise or both by clicking the ‘<’ and ‘>’ buttons of the attribute at the top of the page.



*Second tab of the recommender system*

This tab displays 10 personally generated recommendations based on the user's preferences. Items can be compared on A.W.H. values, cost and time by clicking on the two sub-tabs.

Users can either select or remove items by clicking the green checkmark or the red trash bin respectively.



*Third tab of the recommender system*

The final and third tab shows the user his personally composed to-do list. By clicking the blue curved arrows users can still remove items from their to-do list.

After clicking the 'Verder' link at bottom right of the screen the to-do list is finalized and emailed to the user.

### *User experience questionnaire (25-items)*

In this screen, the user is asked to indicate his or her preferences concerning the use of the lifestyle coach and the recommendations they selected.

### *Finalization and Sharing*

Users are told that their personal to-do list of selected items has been sent to their e-mail address. If they wish to, they also have the opportunity to share their to-do list through Facebook, Twitter, LinkedIn or Google+.

Users are informed that in a week from now they will receive an e-mail reminder asking them to sent back their to-do list and indicate which items they actually executed and which ones they did not. If they do so, they will receive the monetary reward of €2.

In case users would like to continue to use the lifestyle coach they can use the personal link at the bottom of the page. The personal link leads to a version of the lifestyle coach that is stripped of questionnaires.