

**Ideation Phase**  
**Literature Survey on Selected Project and Information Gathering**

Date	19 September 2022
Team ID	PNT2022TMID24963
Project Name	Project - Nutrition assistant Application
Maximum Marks	4 Marks

Healthy nutrition contributes to preventing non-communicable and diet-related diseases. Recommender systems, as an integral part of mHealth technologies, address this task by supporting users with healthy food recommendations. However, knowledge about the effects of the long-term provision of health-aware recommendations in real-life situations is limited. This study investigates the impact of a mobile, personalized recommender system named Nutrilize. Our system offers automated personalized visual feedback and recommendations based on individual dietary behaviour, phenotype, and preferences. By using quantitative and qualitative measures of 34 participants during a study of 2–3 months, we provide a deeper understanding of how our nutrition application affects the users' physique, nutrition behaviour, system interactions and system perception. Our results show that Nutrilize positively affects nutritional behaviour (conditional  $R^2=.342$ ,  $R^2=.342$ ) measured by the optimal intake of each nutrient. The analysis of different application features shows that reflective visual feedback has a more substantial impact on healthy behaviour than the recommender (conditional  $R^2=.354$ ,  $R^2=.354$ ). We further identify system limitations influencing this result, such as a lack of diversity, mistrust in healthiness and personalization, real-life contexts, and personal user characteristics with a qualitative analysis of semi-structured in-depth interviews. Finally, we discuss general knowledge acquired on the design of personalized mobile nutrition recommendations by identifying important factors, such as the users' acceptance of the recommender's taste, health, and personalization.

## Motivation and objective

The research in recommender systems has been recently interested in food recommender systems addressing, among others, nutritional health with different approaches (Trattner and Elsweiler [2018](#)). These systems have the potential to help users navigate the growing amount of multimedia food content (Min et al. [2019a](#)) while fostering healthy eating patterns. Conventional recommender systems learn the users' preferences and try to cater to them, which might enforce recommendations for unhealthy food as well (Elsweiler and Harvey [2015](#); Schäfer et al. [2017b](#)). Thus, health-aware recommender systems need to also incorporate different parameters related to taste and health into their systems (Elsweiler et al. [2015a](#), [2016](#); Ge et al. [2015a](#)). The use of nutrition assistance systems is promising since previous studies have shown that persuasive technologies can help people to eat healthier (Orji and Moffatt [2016](#)).

Existing national and international dietary guidelines constitute important informational sources for nutrition (Painter et al. [2002](#)) but are often based on guidelines for a whole population instead of dealing with individual requirements. Yet, efforts towards personalised nutrition recommendations have been made by Zeevi et al. ([2015](#)) and within the Food4Me study (Celis-Morales et al. [2015](#)). Zeevi et al. ([2015](#)) incorporated different individual aspects into their algorithm based on machine-learning techniques like dietary behaviour, anthropometrics, blood biomarkers and the gut microbiome. Based on this, they could successfully predict the post-prandial glucose response (PPGR) that varies greatly between different individuals for the same meal. Celis-Morales et al. ([2015](#)) examined in their 6-month study the effectiveness of personalised nutrition advice, which was based on dietary, phenotypic and genotypic information. Their results showed higher effectiveness in changing nutrition habits through personalised dietary advice than conventional dietary advice.

Evaluations of recommender systems often focused on measuring algorithmic accuracy, which is insufficient in explaining user experience (Knijnenburg et al. [2012](#)). Further, studies on recommendations and mobile applications frequently cover short-term usage, i.e. think-aloud lab studies of multiple hours or surveys comparing recommender algorithms on a quantitative level. Especially in the context of health recommender systems, it is important to not only evaluate recommendation accuracy. Since healthy recommendations might contrast user preferences, it is crucial to evaluate user satisfaction and changes in behaviour and health over a longer period of time, i.e. multiple weeks, (Schäfer et al. [2017b](#)). Traditional nutrition interventions require months to show lasting effects on nutritional behaviour and physique. Therefore, we evaluate our proposed system, *Nutrilize*, based on a 2–3 months study using a mixed-methods evaluation of the system effects and the user experience. It is our goal to show how long-term usage of a nutrition assistance system influences the users' (a) physique, (b) nutrition behaviour, (c) system interaction and (d) system perception. Furthermore, we want to gain insights into the reasons for observations appearing in long-term but not in short-term usage by analysing semi-structured in-depth interviews.

## Background

---

This section discusses related work in four different areas relevant to the application and study design of the presented work: Personalized Nutrition, Food Recommender Systems, Behaviour Change Interfaces, and Evaluation of Recommender Systems. Additionally, we present how our prior work on this system fulfils different prerequisites of the layered evaluation framework of Paramythis et al. ([2010](#)).

## Personalized nutrition

Food recommendations for nutrition personalized according to individual health requirements are a major research gap identified by several food recommender systems surveys (Trattner and Elsweiler [2017a](#); Mauch et al. [2018](#); Min et al. [2019a, b](#); Theodoridis et al. [2019](#)). One of the most prominent studies on personalized nutrition is the Food4Me study (Celis-Morales et al. [2015](#)). Over an intervention time of six months, 1607 participants across multiple European study-centres received four types of advice via e-mail: 1) a control group receiving conventional advice, 2) a group receiving personal advice based on dietary intake, 3) a group receiving personal advice based on dietary intake and phenotype, 4) a group receiving personal advice based on dietary intake, phenotype, and genotype. Regarding the impact on dietary behaviour, the study shows higher scores according to the MedDiet Mediterranean diet (MedDiet) (Davis et al. [2015](#)) for the personalized advice group than for the control group (Food4Me Study [2016](#)). The phenotypic and genotypic groups showed no significant difference to the group with personalized feedback based on dietary intake (Food4Me Study [2016](#)). While the personalization was more effective, the overall improvement in MedDiet scores was only modest (Food4Me Study [2016](#)). The authors suggest continuous internet-based delivery of advice to increase the impact (Food4Me Study [2016](#)). Similarly, a second evaluation focused on the improvement of the Healthy-Eating-Index shows that participants receiving personalized feedback consumed less red meat, less salt, less saturated fat, more folate, and had higher Healthy-Eating-Index scores at month six than the control arm (Celis-Morales et al. [2016](#)), with no significant difference between the personalization branches. In line with these insights, the *Nutrilize* system is integrating personalization according to intake history, phenotype, and blood measures.

## Food recommender systems

Food recommender systems have been implemented and evaluated using many different algorithms and evaluation methods. Most algorithmic solutions are based on standard content-based or collaborative filtering methods, as shown in the overview by Trattner and Elsweiler ([2017a](#)). While work with smaller user samples concludes that content-based methods are of superior performance (Freyne and Berkovsky [2010a](#); Harvey et al. [2013](#)), experiments on larger samples show a higher performance of collaborative approaches (Trattner and Elsweiler [2019](#)). In general, different food recommender systems tested by Trattner and Elsweiler ([2019](#)) show lower performance, i.e. area under the ROC curve, than similar systems addressing movies

or e-commerce tested by Rendle et al. (2012). Approaches to health-focused food recommender systems have used energy balancing (Ge et al. 2015b), distance from an estimated nutritional requirements (Elsweiler et al. 2015b), or re-weighting according to health metrics (Trattner and Elsweiler 2017b). Beyond prediction of ratings or ranking based on an existing set of recipes, other efforts have been focused on suggesting healthier ingredient substitutes (Achananuparp and Weber 2016) or on generating healthier pseudo recipes (Chen et al. 2019). The implemented algorithm in the *Nutrilize* system is a content-based approach for both health and taste estimation that integrates both energy balancing and nutritional requirements.

### Behaviour change interfaces

Beyond the accurate ranking of personalized healthy and tasty recipes, the design of the user interface is a crucial element determining the acceptance of recommendations and the change in behaviour. According to Chen et al. (2017), it is beneficial from a nutritional care perspective to provide information on nutrients and energy in apps. Current commercial nutrition apps mostly focus on the provision of calorie counting and macronutrient distributions such as MyFitnessPal (Google LLC 2019a), Yazio (Google LLC 2019f), MyNetDiary (Google LLC 2019b) or MyPlate (Google LLC 2019e). Only a few use colour- and category-based information to ease the food item choices of users and provide educational content on nutrients such as Fooducate (Google LLC 2019c) or LifeSum (Google LLC 2019d). *Nutrilize* goes beyond common mHealth systems for balanced nutritional advice because of its level of personalization included in the underlying algorithm for generating nutritional advice and its design of feedback. Our concept for developing feedback strategies in *Nutrilize* relies on several practical as well as theoretical considerations. We draw on scientific experience from Front-of-pack nutrition labelling (FoPL) schemes using traffic light schemes that proved to be effective for food item choice support in shopping environments (Malam et al. 2009; Dunford et al. 2014; Koenigstorfer et al. 2014; Julia and Herberg 2017). From a theoretical perspective, our implementation of (visual) feedback

bases on the principles of reflective practice as proposed by Schön (1983) in terms of feedback offering reflection on ongoing food choices as well as past nutrition behaviour to optimally support dietary choices. The offered types of (visual) feedback in *Nutrilize* according to Schön (1983) are reflection-in-action implemented in the home screen through colour-coded advice symbols and the food details screen through simulative/predictive feedback on how selected food intake affects individual critical nutrients. The concept of reflection-on-action is offered by different retrospective views on single individual nutrient levels and overall energy in a daily, weekly, and monthly view. Additionally, the system design is oriented by Persuasive System Design as proposed by Oinas-Kukkonen and Harjuma (2009). Based on their model of persuasive system design (PSD model), *Nutrilize* implements the following PSD elements in its system components: personalization, self-monitoring, suggestion, and simulation (Terzimehić et al. 2016). We further introduce our considerations for feedback implementation in the context of recommender systems by the taxonomy of Nunes and Jannach (2017). They conducted a review on explanations in recommender and decision support systems and ranked visualization of input parameters after natural language-based texts to the second-most used explanation. Further, they rank different purposes of using explanations in studies according to use frequency. The most frequent purpose is transparency, followed by effectiveness, trust, persuasiveness, satisfaction, and education. By integrating textual and visual explanations into our system *Nutrilize*, we tried to increase healthier dietary decisions (efficiency) as well as to make the parameters for calculating recipe suggestions visible (transparency), which again should result in higher levels of trust towards the automated suggestions of the system. Since *Nutrilize* offers nutrient information on a very fine-grained level, with textual and visual feedback, it provides educational potential towards an improved understanding of foods and relating nutrients.

## Evaluation of recommender systems

Food recommender systems are most frequently evaluated in offline comparisons of algorithms on a benchmark dataset (e.g. Trattner and Elsweiler [2019](#)). Such offline evaluations often assess a combination of different evaluation metrics for recommender systems (Vargas and Castells [2011](#)) such as accuracy, diversity, and novelty. For recommendations in the food domain, diversity is of special importance for user satisfaction since food choices are recurrent and frequent. *Nutrilize* generates new recommendations each day for all meals. These recommendations should provide diversity on two different levels: within a recommendation and between recommendations over time. A similar evaluation of diversity within and between recommendations has been conducted by van Schaik ([2019](#)), who proposed package recommendations for healthy meal plans. Intra-List Diversity (ILD) (Vargas and Castells [2011](#)) is a fitting metric to measure diversity in the context of meal recommendations. The diversity of recommendations over time is relevant in the food context since food decisions are recurrent. To measure diversity over time, we use the Self-System Diversity (SSD) (Vargas and Castells [2011](#)). The SSD considers two subsequent recommendations, which corresponds to two subsequent days in the case of *Nutrilize* recommendations. We additionally propose the Weighted Self-System Diversity (WSSD). It considers more subsequent recommendations and employs the weighting approach proposed by Ding and Li ([2005](#)) to model the gradual forgetting of past experiences, a notion presented by Koychev and Schwab ([2000](#)). While offline evaluation is frequent and important, online evaluation such as surveys (Musto et al. [2020](#)) and user studies (Massimo et al. [2017](#)) have become more common in the area of food recommender systems. While many user studies on food recommender systems are conducted in shorter sessions (Massimo et al. [2017](#)), the health-context often requires longer durations to show the behavioural and physical impact. One limitation of studying the behaviour within such a setting is that many users behave differently when they are part of a study (known as the Hawthorne effect (McCarney et al. [2007](#))). Two similar approaches to *Nutrilize* are provided by Alrige and Chatterjee ([2018](#)) and Fallaize et al. ([2019](#)). To date, Alrige and Chatterjee ([2018](#)) have conducted a quality assessment of the mobile app without measuring the health-impact. The

eNutri app from Fallaize et al. (2019) which addresses the level of personalization in nutrition-related apps as proposed by Chen et al. (2018) and Franco et al. (2016) is close to ours in using the validated Food4Me Food Frequency Questionnaire (FFQ) (Food4Me 2016) as a basis for offering automated personalized nutrition advice. However, the eNutri app does not offer an automated personalized recipe or food item recommendations but only a personalized report. The eNutri app (Fallaize et al. 2019) has been evaluated by asking users for input and feedback as well as comparing the output to nutrition expert advice. We evaluated *Nutrilize* in several focus groups, simulations, user tests, and a pilot study to verify its validity before showing its impact on perception and interaction in a short-term study. The work continues this line of evaluation by focussing on behavioural and physical changes in a long-term study.

#### Prior work

The components of the *Nutrilize* system have been evaluated in several unpublished pre-studies, a published design concept (Terzimehić et al. 2016), a published pilot study (Leipold et al. 2018) and a published short-term study on ability-based personalization (Schäfer and Willemsen 2019). This incremental validation of *Nutrilize* can be presented in terms of the framework of layered evaluation by Paramythis et al. (2010). This framework consists of five layers: (a) Collection of input data (CID), (b) Interpretation of the collected data (ID), (c) Modelling of the current state of the “world” (MW), (d) Deciding upon adaptation (DA), and (e) applying (or instantiating) adaptation (AA). Beyond these five layers, Paramythis et al. (2010) present suggestions for evaluating adaptation as a whole. Regarding the first layer, the *Nutrilize* application collection input data on the food behaviour of users and on the interactions with the different screens. Regarding the quality of food items, we compared three different nutritional databases regarding the accuracy and variety of nutritional information. Regarding the nutritional values of recipes, we tested our matching of recipes to the BLS database with a test set of ground truth samples provided by nutrition scientists. For the tracking of user interactions, we verified the completeness and validity of our tracking system during our pilot



study (Leipold et al. [2018](#)). Regarding the second layer of interpretation of the collected data, we simulated both the health recommendations as well as the taste recommendations for prototypical user inputs and let the results be verified by a potential user, in the taste case, and nutrition scientists, in the health case. For the third layer of accurately modelling the world, we used focus groups, interviews, and user tests using the think-aloud method to verify the appropriateness and understandability of our system. On the fourth layer of choosing the fitting form of adaptation, we performed to date only one comparative study that tests the effectiveness and perception of two different levels of personalization (Schäfer and Willemsen [2019](#)). Finally, for the fifth layer, we conducted a full pilot study over three weeks to measure the systems robustness as well as the user's interactions and perceptions (Leipold et al. [2018](#)). The long-term study presented in this paper can either be described as an evaluation of the system as a whole instead of a layer-wise or piece-wise evaluation of specific factors. Many of the components of our system are interconnected and cannot be measured holistically in separate evaluations. For example, the quality of real user input influences the predictive quality, which in turn changes the impact of the persuasive and personalized feedback that the system provides. Thus, during our long-term study we want to collect a variety of in-detail assessments of a real-life usage scenario.

### The *Nutrilize* application

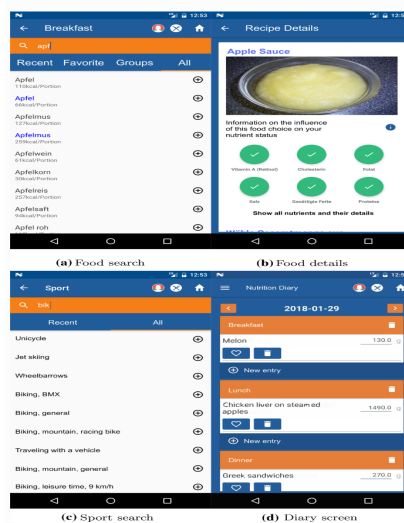
We examined our research questions using the nutrition assistance system *Nutrilize*. The early design and background of *Nutrilize* is described by Terzimehić et al. ([2016](#)) and a pilot study on the general usage of *Nutrilize* is reported in Leipold et al. ([2018](#)). Schäfer and Willemsen ([2019](#)) show the most current version of the system, but with a Rasch-based tailoring and a Dutch target group. This section will shortly describe all the features of the current system version used during our long-term study. First, we describe all features required for tracking the daily dietary intake of the participants, namely the food-search, food-details, sports-search, and diary. Second, we describe the recommendation features. Third, we describe all visual feedback screens, namely the statistics screen, nutrition status screen, home screen, and energy

overview. Finally, we show all the administrative features such as the preference screen, the profile screen, the login screen, and the settings screen.

## Tracking features

The diet tracking of each user is done using a search interface (Fig. 1a). The user can either perform a free text search, select the food item from a tree structure, or select one of his/her recent or favourite items. When choosing a food item, the user can either directly add the default amount by clicking on the plus button, or first, click on the food item to receive some more detailed information on the food's nutrients and choose a custom portion size to add to the diary (Fig. 1b). The same mechanism is given for entering physical activity (Fig. 1c). Instead of portion sizes, here the user should choose the amount of time for the performed physical activity. Finally, the user can review and update all his entered food items in the diary. Here he can also enter food for past days or delete wrong entries (Fig. 1d).

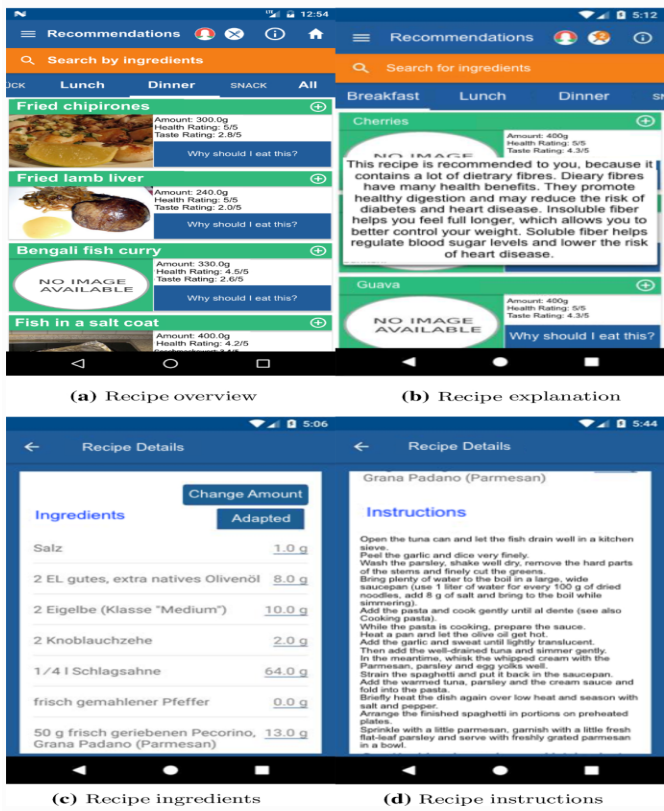
Fig. 1



Tracking screens

Full size image

Fig. 2



Recommendation screens

## Full size image

### Recommendation features

The recommendation feature shows a list of recommended recipes split by meal type to the user (Fig. 2a). The ranking of the recipes is based on a weighted average rating based on 75% of the predicted health rating and 25% of the predicted taste rating of the user, as described in more detail in Sect. 4. The health rating prediction process is described in our previous work (Leipold et al. 2018). The taste rating is based on the relative amount of food types in the user's dietary history. The user can scroll through the recommendations and get an explanation of why the recipe was recommended when clicking the respective button (Fig. 2b). The user can furthermore open the recipe and see related information on ingredients (Fig. 2c), instructions (Fig. 2d), and the nutritional content of the recipe (Fig. 1b).

### Visualization features

The user can receive visual feedback on his nutritional behaviour in different ways. The most prominent feature is the home screen, which is shown every time the user opens the application (Fig. 3a). Here the user sees the six most critical nutrients and how he should improve their balance in his diet. It also shows a message about how many of the tracked nutrients are currently in an optimal intake state. The user then has two options for getting more detailed insights. He either clicks on one nutrient to access the nutrient-details screen (Fig. 3b). Here he can see his progress over the last three days and receive information on the nutrient itself, such as the food it is prominent in, and the effects it has on the body. Additionally, the user can click on the middle bubble of the home screen to get the nutrition status of all nutrients (Fig. 3c). All three views are based on the average intake of the previous

three days, excluding the current day. In addition to the immediate feedback, the user can see more detailed information on his history in the energy screen and the statistics screen. The energy screen offers a combined overview of the energy intake and the bonus calories gained through physical activity over the previous week (Fig. 3d). The statistics screen offers a visualization of the nutritional history of the day, week, and month (Fig. 3e). Additionally, it shows the history of profile changes such as weight, physical activity, or waist-to-hip ratio.

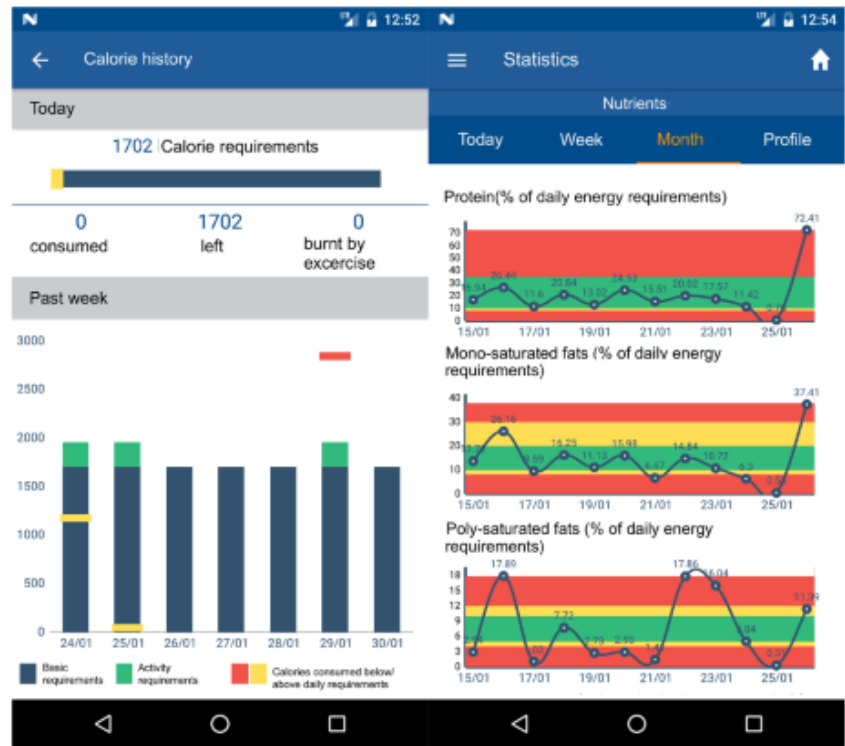
Fig.3



(a) Home screen

(b) Nutrient details

(c) Nutrition status



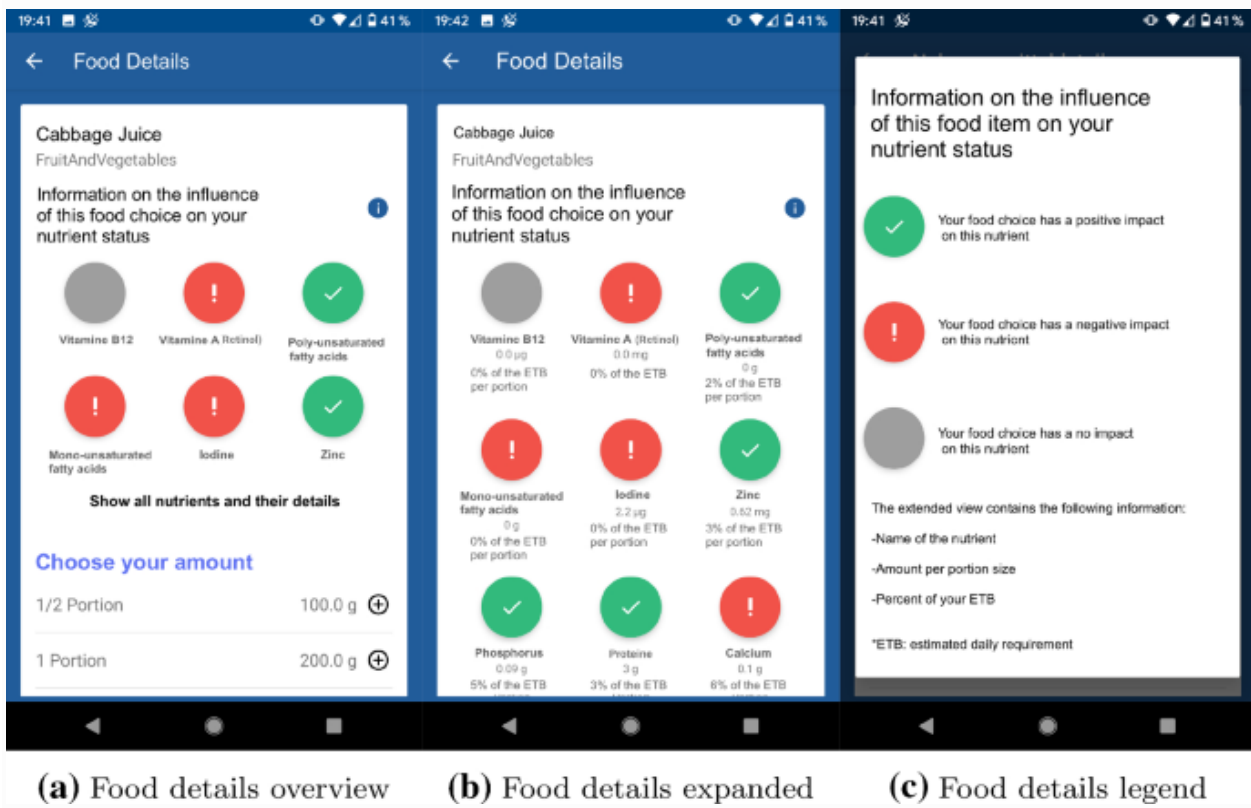
(d) Energy overview

(e) Nutrient statistics

Visualization screens

Full size image

Fig. 4



Visualization feature offering decision support by simulation

## Full size image

The visual displays of the home, food detail and statistics screen in *Nutrilize* are influenced by the theoretical concept of reflective practice by Schön (1983). According to that, we distinguish between the two concepts of reflection-in-action and reflection-on-action, whereas we offer reflection-in-action in two different ways. First, when the user enters the screen, she is provided by the entry screen and a display of the six critical nutrients. This screen aims at triggering reflection-in-action on individual nutrient states according to colour codes and advice for action that should be set into focus when engaging with *Nutrilize* to plan or search future food items or recipes.

Furthermore, *Nutrilize* provides a visualization feature offering a simulation of future intake based on the current nutrient status of the user. This feature aims at supporting the user with decision help on his individual nutrition choices besides the recommendation items. The screen can be entered after using the search through clicking on search result items—either when searching a single food item through the search bar, or alternatively when clicking on a recipe offered in the search result list. The screen content presents the six critical nutrients that are also shown on the home screen. However, here the colouring of the nutrients indicates if the critical nutrients would be influenced by the selected food when intaken. Simulated information is shown by three different states of influence—red (containing exclamation mark for negative influence), green (containing checkmark for positive influence), and grey (indicating no influence) as shown in an explanatory legend (Fig. 4c). The simulation screen exists in two variants, a quick overview screen as described above (Fig. 4a) and an expanded variant (Fig. 4b). The expanded screen offers simulative information on all 27 nutrients that are considered within *Nutrilize*. While only showing simulative colour codes and symbols on the quick view, the expanded view presents further information for the user, such as quantity per portion and percentage of the recommended daily requirement.



## Administration features

The administrative screens enable the user to configure his experience with the application. First of all, the login screen is used to identify the user. After a successful login, the application should remember the current user when reopening the system. However, technical changes such as server maintenance might lead to the need of a new login after a while. Once logged in, the user can administer his application usage in the settings screen. Here he can change his username and password or switch his tracking of physical activity to a non-daily basis. In the profile screen, the user can also update his personal characteristics, such as weight, height, hip measure, waist measure, and his personal motto/goal. Finally, the preference screen enables the user to exclude certain food items or food groups from his recommendations, to facilitate special needs such as vegetarian, kosher, halal, or simply dislike of a specific ingredient.

## Algorithm for personalized recipe recommendations

The recommender algorithm was reviewed, tested, and refined in collaboration with nutrition science experts in multiple iterations. Significant extensions compared to our prior work are the addition of blood value personalization, taste preference, and explanations. These will be explained in more detail below.

## Health recommender system

Formal nutrition models need to cover the user profile, food information, and advice functions. The user profile variables that were considered are: (1) the BMI (Body-Mass-Index) calculated from the user's height and current weight, (2) the waist to hip ration calculated from the hip-measurement and waist measurement, (3) the user's basal metabolic rate calculated from age and gender, (4) the total energy expenditure calculated from the physical activity level and basal metabolic rate, and (5) the user's risk for diseases based on blood values. The variables

from the user profile are later used to weight different nutrients in a personalized way. While BMI and WHR (waist-to-hip ratio) influence the recommended portion sizes, gender and age are parameters used in the advice functions. The food diary builds a bridge between pure food and pure user-related information. It is a crucial part of the personalization to consider previous dietary intake. Regarding the information on food items, two types of input need to be distinguished. The most critical decision for accurate nutritional feedback is the choice of a nutritional database. Between the popular options BLS, FDDb, and FatSecret, only the BLS (Hartmann et al. [2005](#)) offered sufficient types of nutrients and adequate accuracy. Single ingredients such as an apple are retrieved from the BLS database (Hartmann et al. [2005](#)) where all nutritional values and information on the item's food group are stored. Items in the BLS database consist of multiple variants (e.g. cooked, baked, raw) of basic food items (e.g. apple), food products (e.g. milk, yoghurt), and even ready-made meals (e.g. spaghetti bolognese). Recipes can be crawled from different online sources but need to be matched to the BLS database to extract their nutritional values (Müller et al. [2012](#)). The final set of recipes is obtained from the KochWiki (Koch-Wiki [2019](#)) which is licensed under Creative Commons Attribution - ShareAlike 3.0 (Creative Commons [2020](#)). The KochWiki offers 11.360 recipes focused on everyday cooking, including base recipes for sauces, doughs, or mixtures for drinks. The recipes used in the application were matched to the BLS with an adapted version of the matching algorithm presented in Müller et al. ([2012](#)). The final selection of recipes in *Nutrilize* was screened for both correctness of matching and usefulness in everyday cooking.

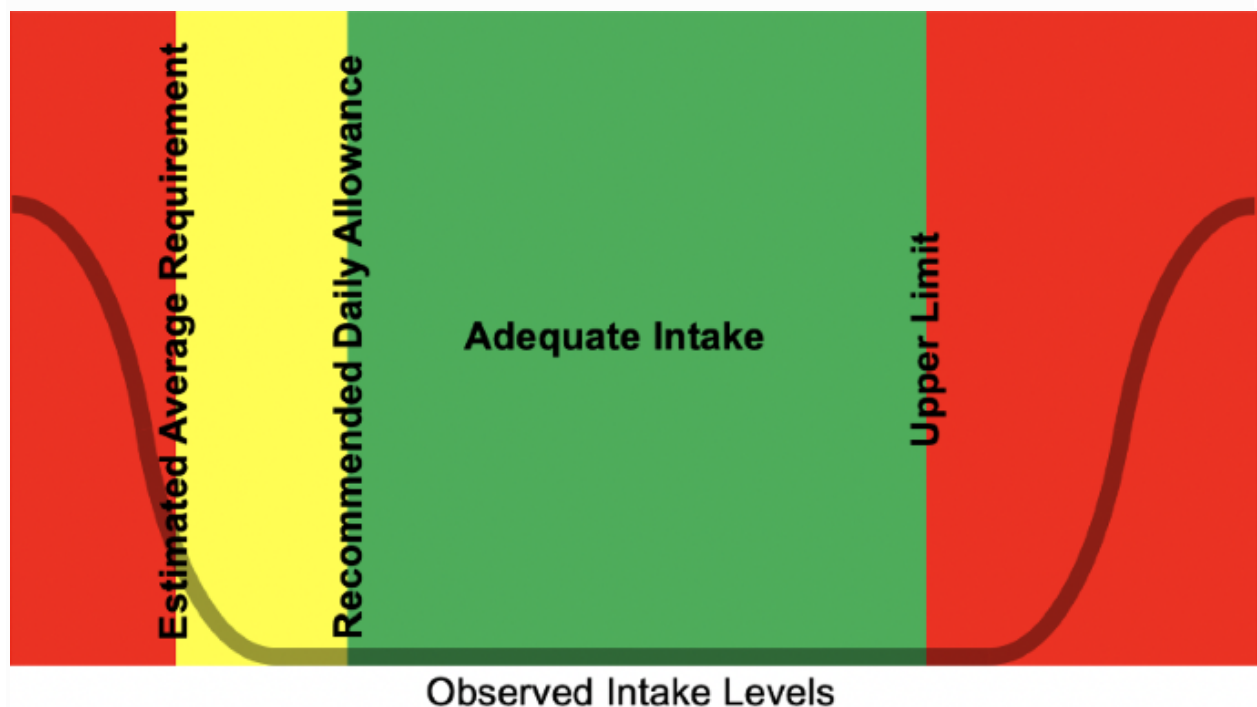
The advice functions used in the *Nutrilize* system are based on the dietary reference intake (DRI) from the Institute of Medicine (Institute of Medicine [2000](#)) and the D-A-CH reference values (D-A-CH [2008](#)).

The general idea of the advice functions is to map the previous intake of nutrients to a value range from strong increase via maintain to strong decrease. For this, we use the classification of the DRI curve into five areas shown in Fig. [5](#). In our algorithm, no intake, if below the EAR (estimated average requirement), is mapped to 1 for a strong increase. Intake above the UL

(upper limit) is mapped to values below  $-1$  for a strong decrease. All other values are mapped in a continuous way around the optimal intake (0 maintain) for values between EAR and RDA (recommended daily allowance). For example, iron has a value of 1 for high increase when no intake has happened yet. Once the intake reaches the EAR value, it is represented by a value of 0.5 for increase. If the intake is above the RDA value, but below the UL, the value is returned as 0 for maintain. If the intake surpasses the UL, the value will decrease into negative values.

The recipe value calculation comprises four core components: an intake vector, a weighted advice vector, a recipe efficacy vector, and the final ranking. The intake vector is calculated by averaging the consumed amount of each nutrient over a window of the past three days.

Fig. 5



Nutrient response curve of the DRI concept by Otten et al. ([2006](#))

### Full size image

The above-calculated vectors are afterwards put together to get the intake history matrix. Based on the cumulated intake history matrix, we can then calculate the *average intake history vector*. The recipe efficacy vector summarizes the contribution of a recipe  $r$  to each nutrient advice. The recipes are scaled to the portion size according to the users personalized recommended energy intake before comparing their nutritional content. The recipe efficacy is calculated by multiplying (Hadamard) the recipe food properties with the user's weighted advice

Finally, the algorithm was used to generate explanations for each recommended recipe based on the most influential nutrient in this recipe's final rating. The explanation texts were handcrafted by collaborating nutrition experts from the *enable* project (enable cluster [2017](#)) and emphasized positive attributes of the respective nutrient.

### Personalization by blood parameters

To further personalize the dietary recommendations, we included dried blood spot parameters, such as whole blood levels of omega-3 fatty acids (Mashavave et al. [2016](#); Sarter et al. [2015](#)), cholesterol (Corso et al. [2016](#); Lakshmy et al. [2010](#), [2012](#)), triglyceride (Lakshmy et al. [2010](#), [2012](#)), and homocysteine (Gempel et al. [2000](#); McCann et al. [2003](#)). If levels were normal, dietary recommendations would not change. If levels were slightly suboptimal (orange) or critical (red), certain nutrients were weighted stronger in the recommendations. Fruit and vegetable (Aune et al. [2017](#)), as well as red meat (Wolk [2017](#)), were included in each food item

as an additional non-nutrient food property because they seem beneficial or harmful beyond their nutrient content. To integrate this into the algorithm, we created the following groups:

- Group 1: Omega-3 index between 4 and 8 (orange)
- Group 2: Omega-3 index below 4 (red)
- Group 3: Cholesterol levels between 200 and 240 mg/dl (orange)
- Group 4: Cholesterol levels above 240 mg/dl (red)
- Group 5: Triglyceride levels between 150 and 200 mg/dl (orange)
- Group 6: Triglyceride levels above 200 mg/dl (red)
- Group 7: Homocysteine levels between 15 and 30 micromol/l (orange)
- Group 8: Homocysteine levels above 30 micromol/l (red)

Suboptimal levels in the omega-3 index (eicosapentaenoic acid (EPA) and docosahexaenoic acids (DHA) expressed as a percentage of total fatty acids) would only affect recommendations on omega-3 fatty acid intake (Harris et al. [2009](#)). Group 1 having only a slightly suboptimal omega-3 index would get an increased weight on recipes with high omega-3 fatty acid content. Group 2 would get a higher weight than group 1 on recipes with a high omega-3 fatty acid content, as they have a critical omega-3 index. Suboptimal cholesterol levels are linked to energy intake, saturated fatty acid intake (Hooper et al. [2020](#)), cholesterol intake (Lin et al. [2018](#)) and intake of dietary fibres (Sima et al. [2018](#)), as well as fruit and vegetable intake. The total energy of the recommended dishes would be reduced, recipes with high saturated fatty acid and cholesterol content would be ranked lower, while recipes with high fibre and fruit and vegetable content would be ranked higher in the recommendations. Modifications in group 4 are similar to group 3 but more pronounced. Suboptimal levels of triglycerides (c.f. groups 5 and 6) would change the recommendations towards a reduction of saturated fat intake (Hooper et al. [2020](#)) and increases of omega-3 fatty acid, dietary fibre, and fruits and vegetables. Recipes with high saturated fat would be ranked lower in the recommendations, while recipes with a high content of omega-3 fatty acids, dietary fibre, or fruit and vegetable content would be ranked higher. Suboptimal homocysteine levels (c.f. groups 7 and 8) would change the recommendations to

advise increased intake of the vitamins folate, vitamin B6, and B12 (Clarke and Armitage [2000](#)) as well as increased intakes of fruits and vegetables (Martí-Carvajal et al. [2017](#)). Accordingly, recipes or foods rich in these nutrients or food groups would be ranked higher for participants in group 7 and even more pronounced in group 8.

### Taste recommender system

While the target of our system is clearly to provide healthy recipe recommendations and our target user group is also focused on this goal, we added a taste component after feedback from our pilot study (Leipold et al. [2018](#)). Since taste preferences often conflict with health recommendation, the final ranking is based on 75% of the predicted health rating and only 25% of the predicted taste rating. Additionally, both ratings are shown in the list of ranked recipes, and each recipe is colour-coded by the health rating. Besides the taste recommendation, any dietary restrictions can be modelled by filter preferences. The user can exclude individual food items or full food groups from all recommended recipes. Since the user already has a high task load, asking for ratings might further decrease engagement and adherence. We thus built an algorithm to derive preferences from past consumption patterns. The recommendation strategies are based on the user's intake tracking over the past seven days, separated by meals. For instance, if a user always eats strawberry yoghurt for breakfast, the algorithm would rank breakfast recipes with fruit and milk products higher. All food items in the BLS and all recipe ingredients have a BLS identifier. This identifier indicates the type of food items with 20 main categories, such as "bread", and ten subcategories such as "tropical fruits". The items in a user's food diary are aggregated by their category or subcategory and weighted according to their frequency over the past seven days. We chose a moving window of seven days to accommodate changed behaviour into the preference modelling. Otherwise, old habits would be incorporated into the recommended recipes and impede the acceptance of new healthier choices. The recipe rating from 1 to 5 is predicted according to its ingredients. Several pre-studies have been conducted to test the performance of our taste-recommender against different collaborative

filtering methods. An initial offline experiment was conducted with 38 overlapping ratings from 6 users on 22 recipes chosen at random from a total set of 171 recipes. A set of seven standard algorithms was evaluated with NDCG (Normalized Discounted Cumulative Gain) and MAP (Mean Average Precision) acting as comparative evaluation metrics according to Said and Bellogín ([2014](#)). The best performance in NDCG was achieved with an implicit matrix factorization method by Hu et al. ([2008](#)). In a follow-up user study, 13 users rated a subset of 50 recipes from a total set of 171 recipes. The participants' task was to simulate ratings from a 29-year old male, of which they had a 7-day diary. The participants were eight women and five men aged between 20 and 60. This evaluation aimed to see whether a diary-based rating prediction would show similar results compared to average user-generated ratings. The taste rating showed a high overlap between the top 6-8 recipes from the ratings and the diary-based list. Due to the artificial situation of simulating ratings from a diary, a second user study focused on diaries and ratings from the same person. The study collected more than 100 ratings from each of the three participants on 240 recipes. The same three participants were asked to fill in their food diaries for seven days. The resulting top three rankings for four meal types from both diary-based strategies and the best three standard rating-based algorithms were shown to the three users for evaluation. Overall, the users preferred the diary-based lists to the standard algorithms, with the implicit matrix factorization method by Hu et al. ([2008](#)) again being the best performing algorithm.

## Evaluation

This study uses both effect differences between group and effects within subjects. It is split into two waves of participants with slightly adapted conditions. The first wave consists of 11 application users (I) and 11 control group users (C) and has a duration of 3 months. The first wave is important to note the difference to a control group and to see the long-term effect of both groups. The application group of this wave could furthermore volunteer for an interview at the end of the study. The second wave consisted only of application users and had a duration of 2 months. Part of the participants had previously conducted the control group of the first wave and thus gave detailed insights into changes of users over time. The second wave of participants

had to additionally conduct an interview after four weeks and another one after the full eight weeks. We thus call this group the qualitative group (Q) or, if previously in the control group and the control qualitative group (CQ). We expect the application group to show higher improvements than the control group. At the same time, we monitor the system interactions, the nutritional behaviour, and several questionnaire-based outcome measures of all participants. We expect the interview data to give more background information and input on the rationale of the quantitative analysis of the data.

### Study procedure and participants

An overview of the study flow is given in Fig. 6. Before starting the study, we had the proposal reviewed by the data protection commission of TUM (accepted on 9th of May 2017) and the Ethics committee of the Klinikum Rechts der Isar (ID 477/16 S). The study is subdivided into 4 phases. The first phase is the screening process. Every participant had to complete a screening survey. If they fitted all the criteria, they had to send in their informed consent form. The second phase consisted of different surveys to collect the pre-study condition of the participants. The third phase consisted of 2–3 months of application usage or, in case of the control group, general nutritional advice via email. The third and final phase consisted of different surveys to collect the post-study condition of the participants. In the second wave, the participants additionally conducted interviews after four weeks and eight weeks. Figure 6 shows the number of participants completing the different study measurement points in each group. Between the initial recruitment and the final survey, we observed a dropout of 20–50%. The dropout in application usage might differ from the dropout in survey participation. The high dropout was expected for a realistic usage of this type of application since none of the participants was paid for conducting the study. As a small reward, any participant that conducted all measurements would receive a personal report on his nutrition during the study.

Fig. 6



First Wave (Controlled 3-month)	Measurement T0	Intervention Group (I)	N=11	Control Group (C)	N=11		
	Application Usage or Control Material 3 months						
	Measurement + End Interview T3/T0		N=4		N=8	Qualitative Group (Q)	N=12
Second Wave (Qualitative 2 month)				N=5			
	Application Usage 4 weeks						
	Mid-Time Interview			N=3	N=6		
	Application Usage 4 weeks						
	Measurement + End Interview T2/5			N=3	N=5		

Flowchart of the study procedure in each group. Overall measurements retrieved from 34 unique users in two different waves. Five users participated first in the control and then in the qualitative group leading to 39 participation datasets. Detailed measurements at each time-point are listed in Table 1

## Full size image

The participants were recruited using different articles and announcements in newspapers, such as the newspaper "Süddeutsche Zeitung" (Süddeutsche Zeitung GmbH [2019](#)), the local newspaper of Freising, an online article in the magazine "Ernährungsumschau" (Ernährungsumschau [2019](#)), and an article in the university's news media (TUM WZW [2019](#)). The targeted user group is slightly overweight and not yet obese, who want to improve their health and well-being without much emphasis on weight loss. However, participants with other BMI values were not excluded from participation. The screening was accessed by around 70 interested participants, of which around 40 finished the full screening survey. The screening excludes any participants that are minors, that have food allergies or intolerances, that suffer from illnesses which are influenced by diet (e.g. diabetes, liver, kidneys, lungs, heart disease, thyroid disease, anaemia), that are currently pregnant, that are following a strict diet, that have no Android device, and that have no internet access. Any participants that passed the screening received the informed consent form. As discussed above, the study was conducted by 34 of these participants who have each given their consent to the participation beforehand. As in previous studies (Leipold et al. [2018](#); Schäfer and Willemsen [2019](#)), we have a strong bias towards female participants (29/34). The average age is higher than in the previous studies with a maximum average of 41 in the intervention group and a minimum average of 33 in the qualitative group. The average BMI is very similar in the intervention and control group (around 29). The two qualitative groups (CQ/Q) have a slightly lower average BMI of 28 for the purely qualitative group and 26 for the participants that already did the control group. It is interesting to note that the control-qualitative group has the lowest BMI indicating either a bias of which people want to continue the study or the effects of the previous control intervention. Since the final questionnaire was only filled in by some participants, the data on user characteristics should be viewed with caution. While in the intervention group, three of the four final questionnaires indicate a university degree, less than half of the final questionnaires do so in

the other groups. The average number of interactions with the application range from 18342.64 per user in the 3-month group (I) to 14585.71 per user in the 2-month groups (CQ and Q).

## Data collection

An overview of the assessed measurements is given in Table 1. We had five different measurement types in this study. First, we measured the changes in physiology and eating behaviour using three standardized questionnaires (Norman (Norman et al. [2001](#)), Baecke (Baecke et al. [1982](#)), FFQ (Food4Me [2016](#))). Second, we measured the users eating habits throughout the study using the app's dietary diary. Third, we measured the users' interaction with the system using a tracking Matomo, formerly named Piwik (Matomo [2020](#)). Fourth, we measured the users' subjective perception of the system [SUS (Brooke [1996](#)), questionnaire from Schäfer and Willemsen ([2019](#))], their dietary change [based on Shannon et al. ([1997](#))], and their personality (Kaptein et al. [2009](#); Rammstedt et al. [2013](#)). Finally, we interviewed the participants on their interaction, their system perception, their dietary change, their motives and the effects of the application.

Table 1 Overview of variables assessed and questionnaires used in this study

## Full size table

### Standardized surveys on nutrition and physiology

All participants had to fill out three standardized surveys at the beginning and the end of the study to be able to compare the changes. The first survey is a food frequency questionnaire

(FFQ) (Food4Me [2016](#)) that covers the approximate intake of 150 typical food items over the past months. The items are connected to their nutritional information in the BLS, and each portion size (chosen by image) is recorded in grams. To ease the usability of this survey, the standard portion size is used for items that are only rarely eaten. The second survey is a Norman questionnaire (Norman et al. [2001](#)) that investigates the physical activity level of a person. In combination with the normal questionnaire, we also asked the user for a repeated entry of his weight, height, waist circumference and hip circumference. Finally, the Baecke survey (Baecke et al. [1982](#)) gives a more detailed insight into the physical activity a user is doing during work, leisure time and physical activity.

### **Daily eating habits**

All the non-control group users were tracking their nutrition using our application. Therefore, we have measures of any food item being consumed, their portion size, and their nutritional values. Furthermore, we track which of the nutrients were focused in the home-screen on the day an item was consumed. The six nutrients shown on the current day are called “focused” nutrients, while all others are “unfocused” for this day.

### **User interaction**

Using the Matomo tool (Matomo [2020](#)), we tracked the timing and duration of each interaction with the screens described in Sect. [3](#). Furthermore, we can extrapolate the number of interactions with each screen on a daily basis.

## Self-reported measures

In the final survey, after using our application, we ask the users several questions on their subjective perception. First, we ask about the perception of the system. Besides a rough estimate the usability of the system with a SUS questionnaire (Brooke [1996](#)), we ask feedback on each application feature about their helpfulness and likability. We furthermore ask four questions on the impression of the overall perceived usefulness of the system (Venkatesh and Bala [2008](#)), and three questions about the intention of use in the future. Finally, we surveyed our users on nine subjective constructs. Five of these constructs focused on the perceptions of the recommendations: perceived healthiness (5 items, i.e. "the recommended recipes were relevant for my health"), perceived tastiness (4 items, i.e. "the recommended recipes fitted my taste preferences", perceived personalization (4 items, i.e. "the recommended recipes were well adapted to my eating habits"), perceived diversity (4 items, i.e. "the recommended recipes were very diverse"), the perceived effort of preparing (4 items, i.e. "the recommended recipes were hard to prepare"). Four constructs focused on the interaction with the system: system effort (5 items, i.e. "the *Nutrilize* app worked very easily"), system effectiveness (4 items, i.e. "With the app, I make better dietary decisions"), choice difficulty (4 items, i.e. "Comparing the recommended recipes took a lot of effort") and choice satisfaction (5 items, i.e. "I am happy with the recipes I have chosen"). For all survey items, users indicated on 5-point Likert scales to what extent they agreed with them. Furthermore, we analyse the usefulness of each visualization using 10 items (i.e. "The xx screen helps me understand my personal nutrient status."). Second, we ask the users about their perceived dietary changes using a standardized dietary change questionnaire (based on Shannon et al. ([1997](#))). Third, we ask the user about their personal characteristics (e.g. job, education), their cooking habits (e.g. cooking type, cooking frequency), and their personality using a 10-item big five questionnaire (Rammstedt et al. [2013](#)).

## Interview data

We constructed interview guidelines for different semi-structured in-depth interviews in the middle of the study and the end of the study. These guidelines were used by the interviewer to structure the conversation while leaving the user the freedom to move back and forth between topics. For participants from the intervention study group, we only conducted one optional interview at the end of the study, where we used most of the mid-study interview guidelines plus some questions on future system usage. The first interview guideline is split into four areas of interest. The first area of interest covers changes in nutrition, with one block focusing on the actual changes and a second block focusing on difficulties and relapse patterns with these changes. With these questions, we want to verify the differences between observable nutrient intake changes and conscious change efforts by the user. The second area of interest covers the interaction with the application with one block focusing on interaction habits, one on the effect of the system components, one on difficulties with the system, one on the visualizations, and one on the recommendations. With these questions, we want to determine why some features are considered more helpful or more popular by users. The final area covers either previous experience with nutrition application (mid-study of the qualitative group) or future usage of nutrition application (end-study of intervention group). Both variants focus on the comparison of *Nutrilize* with other applications and the pro and cons of either system. With these questions, we want to determine the limitations of our system for future work. The second interview guideline is again split into four areas of interest. The first area covers changes in eating motives and habits compared to the first interview with one block focusing on changes in eating perception after using the application and a second block focusing on difficulties and relapse patterns with the eating changes. We hypothesize that more relapses and difficulties have occurred compared to the middle of the study. The second area of interest covers the interaction changes compared to the first interview with one block focusing on interaction habits, one on the interaction motives. We hypothesize that interactions have declined and the satisfaction with some features over others has increased. The third area of interest covers the perceived effects of the study on nutrition knowledge, physical attributes, and psychological

attributes. With these questions, we want to determine whether measured changes are also perceived by the user. The final area covers the future usage of nutrition application with one block focusing on the *Nutrilize* system and one on other possible replacement systems. All interviews were fully transcribed. The first run of coding the interview transcription was done using a deductive method with motives and categories derived from the literature or the final survey data. The second run of more detailed coding was done using the inductive methodology by going through the text and refining the deductive motives/categories with the existent transcription.

## Results

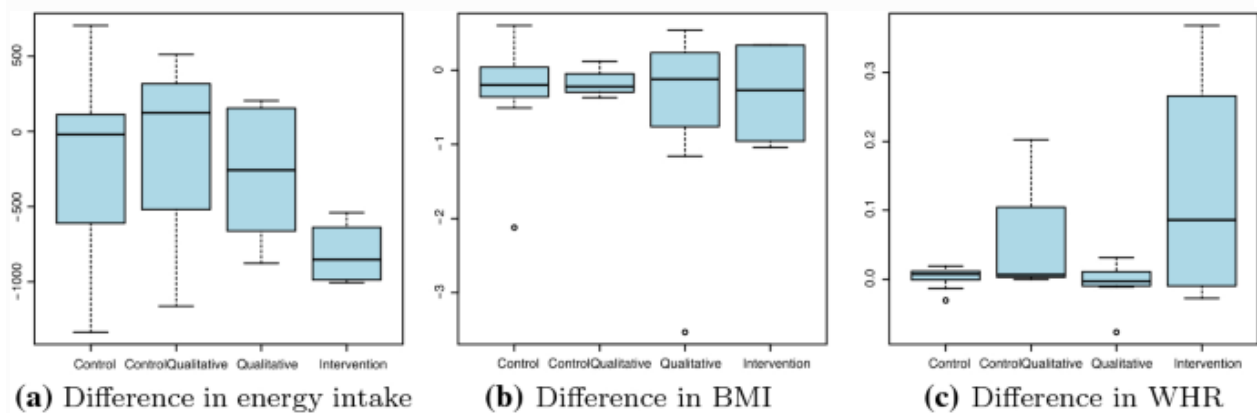
The following result section is structured in line with our research questions on the influence of the system on the users' (a) physique, (b) nutrition behaviour, (c) system interaction, and (d) system perception. The final section gives insights into the context of these results extracted from semi-structured interviews. We hypothesize the highest changes in physique and behaviour in the intervention group (I) and the lowest in the control group (C). We further expect higher dropout in the group with a longer interaction duration (I) than in the group with a shorter interaction duration (Q/CQ). Regarding the perception, we expect user to develop a preference for one type of persuasive feature (feedback, instruction, planning).

### **Standardized pre-/post-study measurements**

This section discusses the change in measurements pre- and post-study. We only included participants that have all pre-/post-measures covered. This restriction may bias the analysis toward motivated participants.

The pre-/post-study analysis of dried blood spot measurements was not possible due to challenges in data retrieval. The participants received a tutorial video and instruction manual to guide them through the process of taking the dried blood spot samples at home. Additionally, they received all the equipment for this process via mail. The experience from the Food4Me study about the quality, compliance, and feasibility of taking DBS (dried blood spot) samples at home was very positive. At the starting point of our study, we received DBS samples from 20 participants out of 22 (I/C) participants. After three months, we only received DBS samples from 12 participants out of 22 (I/C) participants. The quality of the DBS spots was often questionable. Future studies should develop a clear protocol of DBS quality evaluation to determine, which ones can be measured and which ones have to be excluded due to lack of quality. Alternatively, DBS sampling could be done by professionals.

Fig. 7



Difference in energy intake (in calories) and physical markers after the intervention. No significant change in BMI (I:  $t = 0.12$ ,  $p = 0.91$  Q:  $t = 0.28$ ,  $p = 0.78$  CQ:  $t = 0.09$ ,  $p = 0.91$  C:  $t =$



0.93,  $p = 0.85$ ). No significant change in WHR (I:  $t = -1.56$ ,  $p = 0.20$  Q:  $t = 0.06$ ,  $p = 0.96$  CQ:  $t = -0.64$ ,  $p = 0.58$  C:  $t = -0.07$ ,  $p = 0.95$ ). For intervention group marginally significant difference in kCal (I:  $t = 2.26$ ,  $p = 0.06$  Q:  $t = 0.86$ ,  $p = 0.41$  CQ:  $t = 0.39$ ,  $p = 0.73$  C:  $t = 0.53$ ,  $p = 0.61$ )

### Full size image

Figure [Za](#) shows how the energy intake decreased after the study according to the FFQ. The strongest effect is visible in the intervention group, which used the *Nutrilize* systems for 3 months. The weakest effects are visible in the control group and the control-qualitative group. The qualitative group, which was using the application for two months, has a similar but weaker effect compared to the intervention group, with most participants reducing energy intake. In line with the small changes in energy intake, the dependent physique measures by BMI and WHR showed almost no difference. Figure [Zb](#) shows the difference in BMI for the different groups after the intervention. All groups show a slight decrease in the median BMI. There are no strong differences between the groups. Figure [Zc](#) shows there is almost no difference in WHR for the different groups after the intervention.

The nutrition ability is calculated as shown in Schäfer and Willemsen ([2019](#)), by summing up the Rasch difficulty level of the six worst-performing nutrients of a particular user's FFQ. The minimal possible ability would thus be 21 and the maximal potential ability of 159. In the rare cases where less than six nutrients are critical, the scale is continued with hypothetical nutrients 24–29. Figure [8](#) shows the initial ability, the final ability, and the changes of ability. The user's abilities spread between 48 and 125 at the beginning of the study (Fig. [8a](#)) and between 52 and 125 at the end (Fig. [8b](#)). It is interesting to note that the groups that interacted with the application have a low but positive median change of ability. In contrast, the control group has a

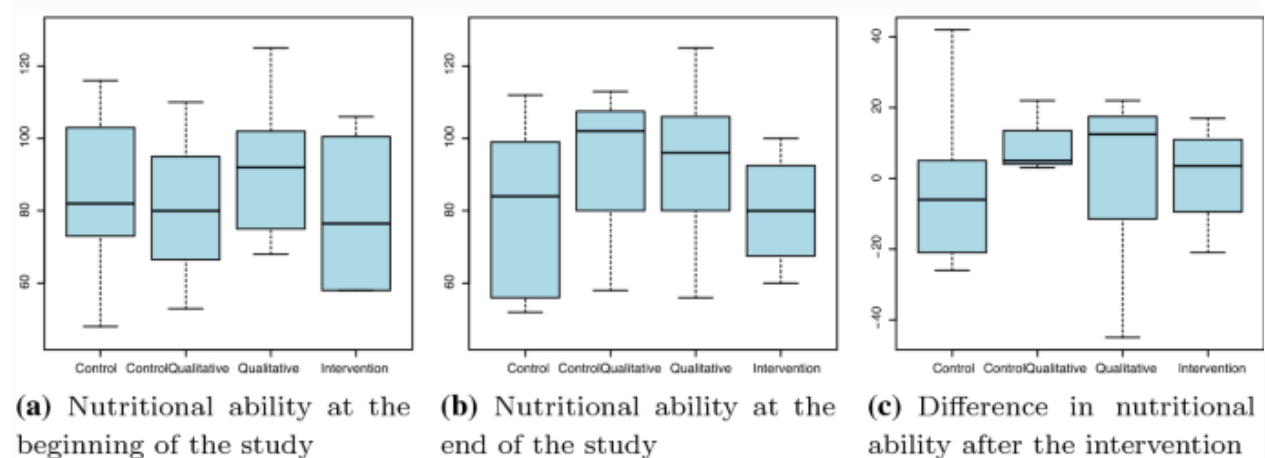
negative median change of ability. However, the small number of participants makes a statistical estimate of this difference unreasonable.

The physical activity was collected from the Norman and Baecke questionnaires. Almost all groups have a median change of their physical activity level around 0. According to the Baecke questionnaire, most groups show a slight increase in activity, while the intervention group shows a slight decrease in activity with large deviations for single participants.

### Effects of the system on nutrition behaviour

The daily nutritional data give us a rich insight into the behaviour of each user over a long-term intervention. First of all, we will compare the energy intake and the nutritional ability with the data from the standardized FFQ that we discussed in the previous section.

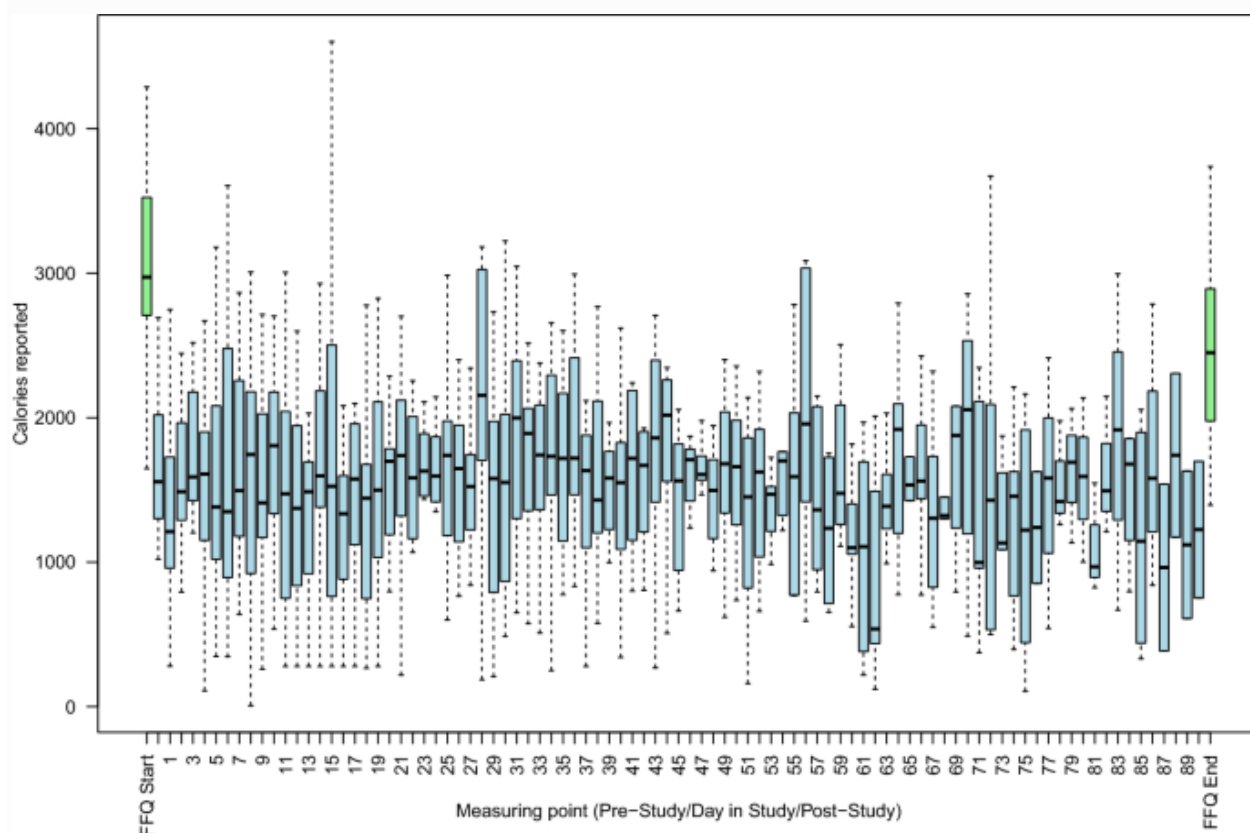
Fig. 8



Nutritional ability over the study in each group. No significant change in ability (I:  $t = -0.05$ ,  $p = 0.96$  Q:  $t = -0.17$ ,  $p = 0.86$  CQ:  $t = -0.43$ ,  $p = 0.69$  C:  $t = 0.09$ ,  $p = 0.92$ )

Full size image

Fig. 9



History of energy consumption in the pre-study FFQ, during the study, and in the post-study FFQ

### Full size image

Figure 9 describes the energy intake in both surveys and from the daily tracking tool. In this figure, we combine data from all participants that were using the application (I, Q, CQ). As in our previous studies (Leipold et al. [2018](#)), the energy measurement is higher in both FFQs than in the daily tracking. While the energy calculated based on the FFQ shows a decrease between the beginning of the study and the end, the daily tracking is much more consistent over time with similar median intakes at the beginning and end of the study. However, the energy intake seems to fluctuate strongly between days. Furthermore, the fluctuation changes around day 60, when the group Q and CQ stopped using the application, and the data only represents the I group or those participants that have not dropped out at that stage of the study.

Since the nutrient success (daily intake within optimal personal range) is dependent on a lot of other variables, as shown in Schäfer and Willemsen ([2019](#)), we looked at that dependency. The number of measurements decreases strongly over time. This decrease was not visible in the short-term study (Schäfer and Willemsen [2019](#)) since it only covered the first 12 days, which are also quite homogeneous in this study. Nutrients, which have no determined focus are mostly appearing at the beginning of the study. This is reasonable since the NoFocus state only appears when there is a lack of data to determine the focus. Both successful nutrients and unsuccessful nutrients are equally distributed over the time, indicating that the dropout is not influencing or biasing the success measurement. The distribution of focused and unfocused nutrients over time, groups and energy intake is always comparable. The only difference we note is that unsuccessful nutrients are occurring more strongly in the focused nutrients. This is to be expected since the algorithm determines the six worst-performing nutrients of the previous

three days for the focus list. We further see that both qualitative and intervention group have more successful than unsuccessful nutrients. The CQ group, however, has an almost even split between the two. Finally, higher energy intake coincide with higher success rates as indicated in the daily energy and nutrient timelines.

To get a closer insight into the influence of the different parameters on the overall success, we built a multilevel logistic regression model. We decided to use a multilevel model to incorporate user differences without losing the power of modelling all collected interactions. The logistic regression is due to the binary nature of the outcome variable of optimal/successful vs non-optimal/-successful intake of a nutrient. We want to show that while on short-term, the application did not yet affect the nutritional ability over time, it will do so in the long-term. First, we want to prove the comparability of the short-term and the long-term dataset, in order to exclude any bias given by the different country, different nutritional database and different recipe dataset. Thus, we focus on the first 12 days of the dataset and compare the same model as in our short-term study (Schäfer and Willemsen [2019](#)).

Table 2 Comparison of short-term study with first 12 days of long-term study

**Full size table**

Table 3 Comparison of first 12 days of long-term study with full assessment of study duration

**Full size table**

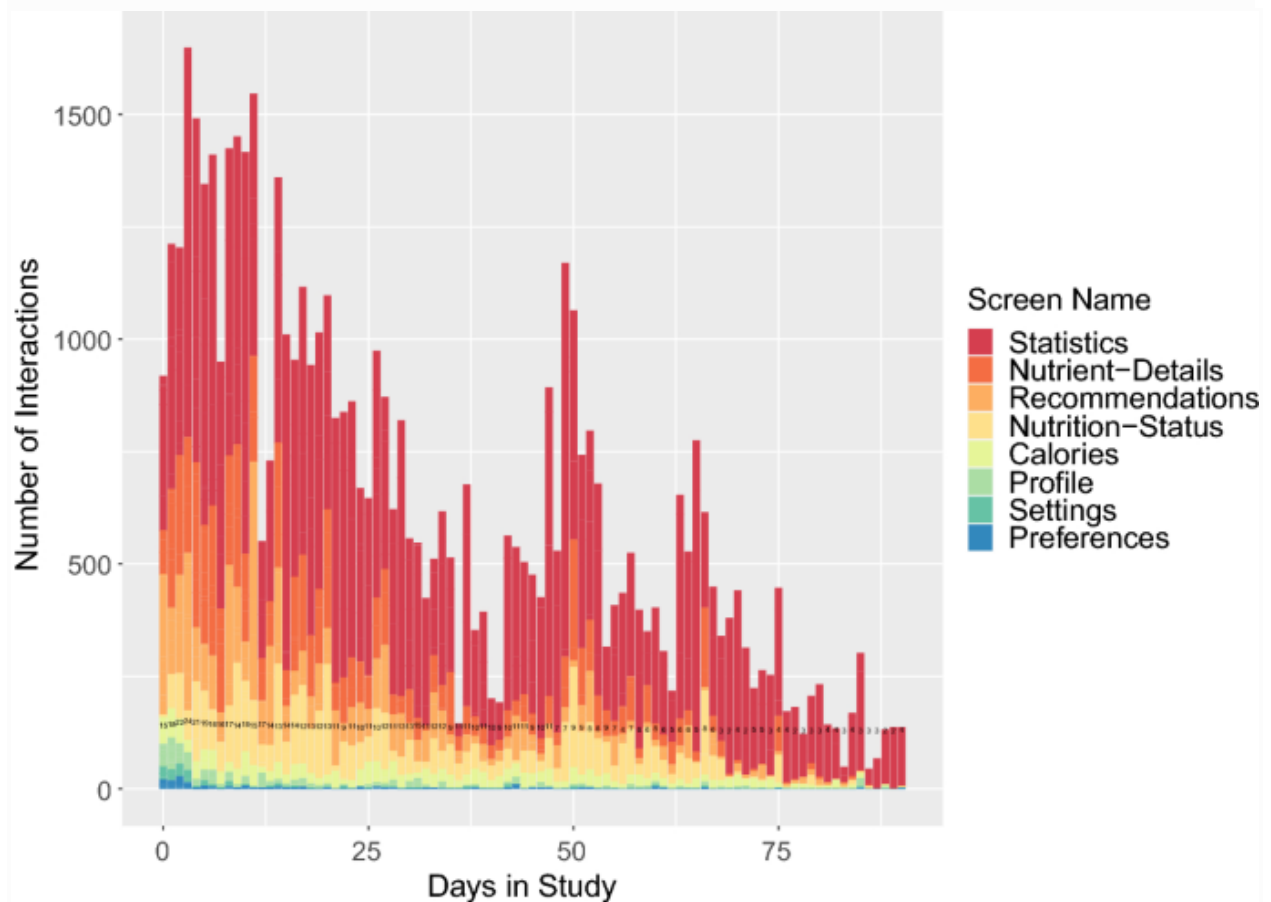
Table 2 shows the comparison of the short-term dataset and the first 12 days of the long-term dataset. Although the estimates are different, in most cases the variables show the same effect on the success in both models. Schäfer and Willemsen (2019) give a detailed discussion of these influences and coefficients. One important difference is that while the recommendations were a significant factor in the Dutch study, they are not in the German study. Also, in both studies, the time factor was not significant during the first 12 days. Now that we have shown the data behave in a similar way, we extend the model to the full long-term study. Table 3 shows the comparison of the first 12 days of the study with the full duration dataset. We can immediately see that with the additional available data, all the factors of the model become highly significant ( $p < 0.001$ ). Thus, we can focus on the direction of the effects. As before, the Rasch scale is loosely represented in the tracked data, except for the fifth level that seems to be easier to achieve than expected, but as discussed in Schäfer and Willemsen (2019) also is highly influenced by the energy intake. Most importantly, the effect of time, that was not visible before is now a strongly significant positive effect. Also, the recommendation interactions, as the only representative for interactions in this model, are significant in the long term but not in the short term. As our previous work (Schäfer and Willemsen 2019) has shown, the recommendations may not be the best representation of interactions since the system has many more features intervening with the user's decision. Thus, the next section will consider the user's interaction with these features and their influence on the behaviour change.

### **Long-term system interaction**

Besides the nutrition from the tracking tool of our application, we have tracked the interactions with the system over time. Since our previous work (Leipold et al. 2018) shows that the most time-consuming activities are the tracking actions, we first want to analyse the interactions on a high level and then zoom down into the detailed feature interactions.

Among the high-level screen categories (Tracking, Recommendations, Visualizations and Administration), the tracking feature comprises around 58–93% of all interactions with an average of 83%. The overall interaction further decreases drastically over time, with only half of the initial actions per day after a period of four weeks. This is partially due to less than daily usage, and partially due to shorter usage sessions.

Fig. 10



History of interactions during the study excluding tracking actions. The additional black numbers on each bar indicate the number of unique users on that day

#### Full size image

To get a closer look into the distribution of feature usage, we will exclude all tracking interactions in Fig. 10. The first thing to note is the continuous decrease of interactions over time. This decrease behaved differently for the types of activities. While the visualization interactions decrease very slowly, the tracking activity increases drastically. The recommendation interactions stay on a similar level over the first weeks but then decrease into almost no interaction. The administrative screens have most of their interactions in the first week, which is to be expected due to their functionality. As shown in our previous work (Leipold et al. 2018), the statistics screen has the most interaction of all features. Considering the long term changes in interactions, we can also see that the statistics screen is one of the few to be still used at the end of the study. From the other visualizations, the nutrient details and the nutrition status, which are both accessible over the home-screen visualization, have the strongest number of interactions and even a long persistence over time, while the energy visualization is of minor importance. The usage of the administrative screens decreases very quickly after the start of the study. The only screen that has some longer activities is the profile screen, where users can enter changes in weight and other body measures. Finally, the recommendations have a fair share in interactions at the beginning of the study but quickly decrease into a minor percentage of the interactions around the middle of the intervention.

Table 4 Extended multilevel logistic regression with one random intercept for users and the successful nutrient intake as an outcome variable, including interaction categories (left) and details for each screen (right)



## Full size table

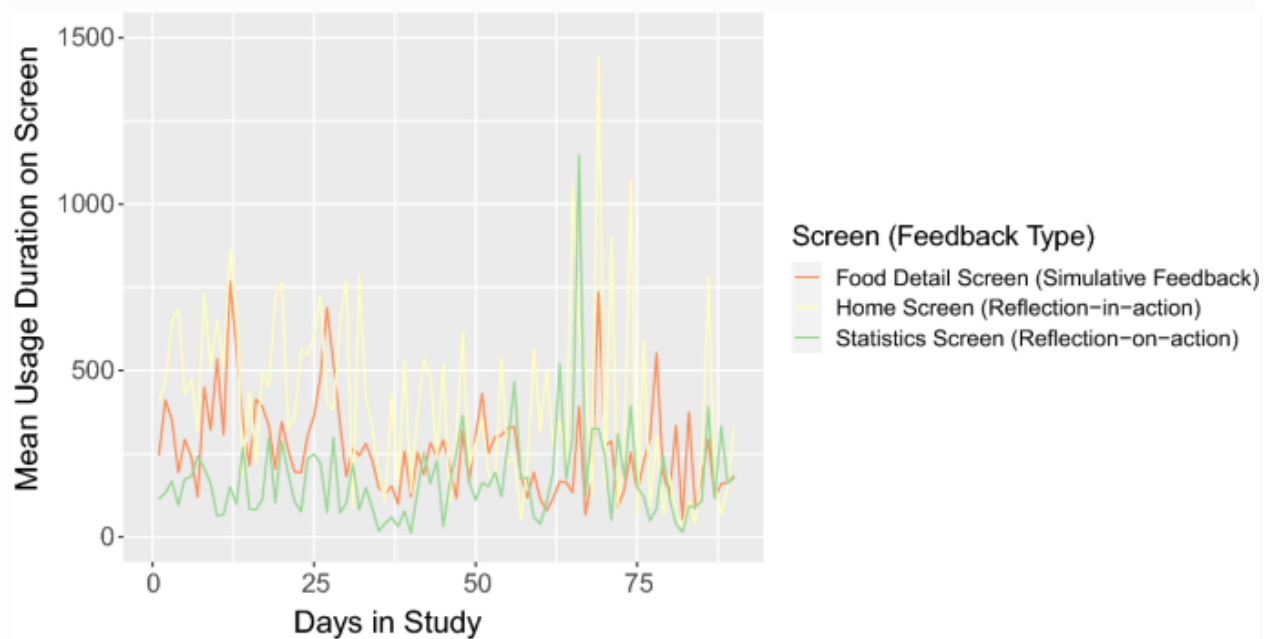
To see, not only which screens the users liked to interact with most, but also which screens finally influenced the users' nutritional behaviour, we excluded the interactions of the current day into the previous multilevel logistic regression model on nutrition success. Table 4 shows the results of these models. As with the visual insight, we first derived a high-level model and then moved into a model considering each screen. The first thing to note is that all the previous effects of the nutrition Rasch-level, the energy intake and the nutrient focus have stayed similar to the previous model. In the high-level model, we see that while the tracking and visualization screens have a positive impact on the nutritional behaviour, the recommendations have no significant impact and the administrative features have a negative impact. Since the recommendation interactions are not further distinguished, they still have no significant effect on the behaviour in the detailed model. Regarding the tracking interactions, the screens responsible for the positive impact on the nutritional behaviour are the food-search and even more so the food details. This is to be expected since the food search indicates a diligent tracking and thus well adjusted nutritional information for the model. In contrast, the food details might even indicate a planning/ decision process during the tracking action. Out of the visualization screens, the most influential scenes are, in line with the number of interactions, the statistics screen, the food-details screen, and the home screen. The visual feedback interactions are reviewed in more detail in the following section.

## Usage behaviour related to feedback types

The tracked usage data provide us with detailed usage durations for each screen. We specifically analyse and compare the usage duration for the most influential visual feedback screens such as the home screen, the food-details screen and the statistics screen. These screens relate to different visual feedback types rooted in reflective practise such as reflection-in-action feedback

by the home-screen and a specific variant of reflection-in-action by the food/recipe details screen that we call simulative feedback as well as reflection-on-action feedback with the statistics screen.

Fig. 11

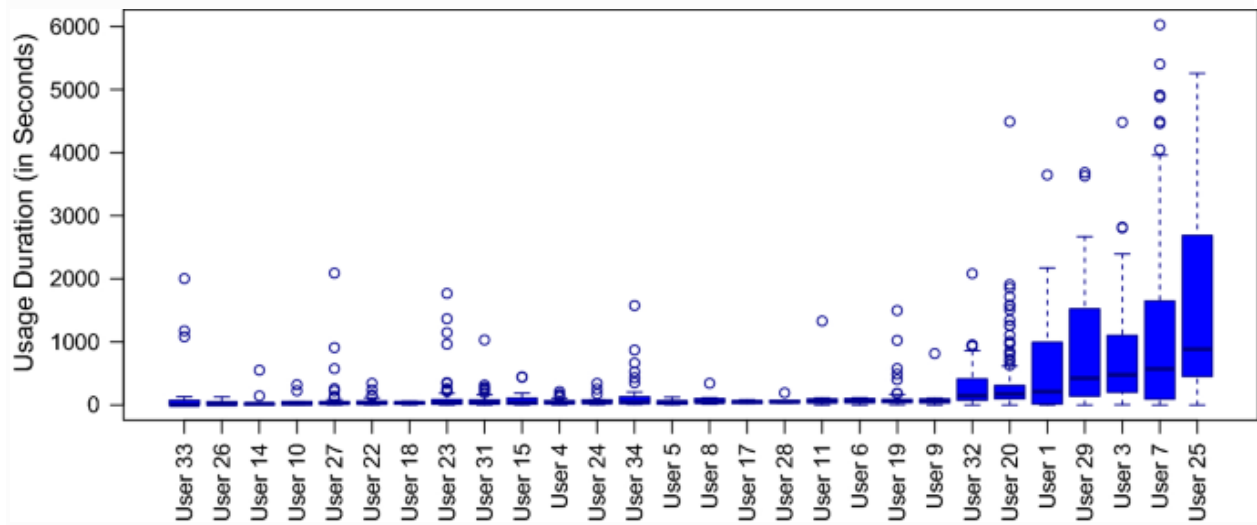


Mean Usage Duration over Study Period per Feedback Type

**Full size image**

Figure 11 shows all mean values for all feedback types in a timely perspective. Most obvious from this figure is that all three features are used over the whole period of study. In the first two-thirds of the study period, there is a more prominent use of the home screen as well as the food detail screen than of the statistics screen. We explain this behaviour due to the possible interests of users in their actual status of diet and potential ways to improve it. While the utilization of the statistics screen remains relatively low compared to the usage of the other two screens, it starts to increase with the second half of the study. This behaviour might further be explained by the increasing amount of tracking data available in the historical self-monitoring later in time.

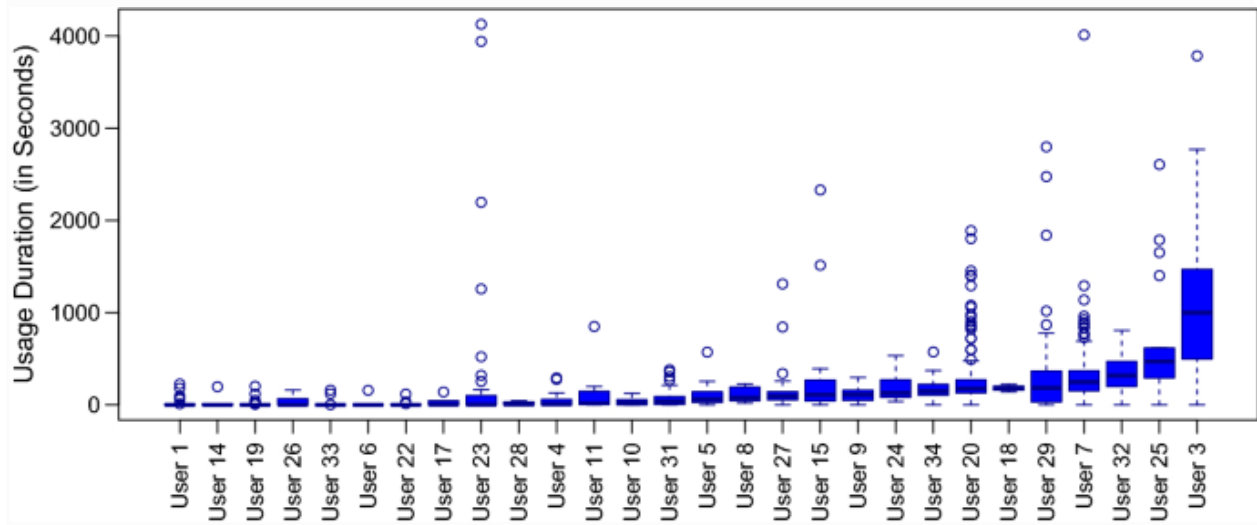
Fig. 12



Distribution of total home screen usage duration per user

Full size image

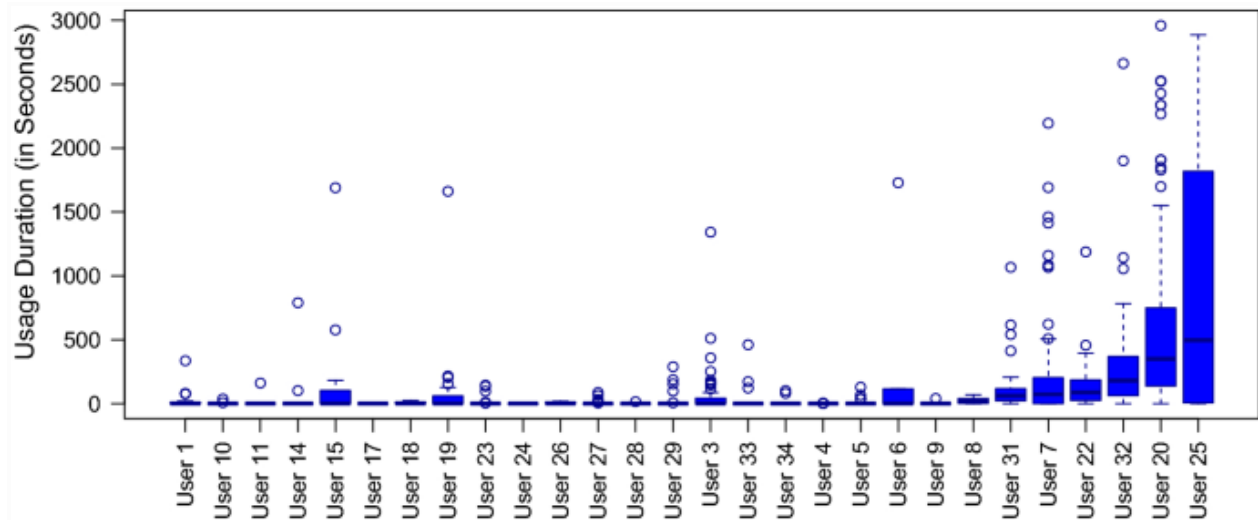
Fig. 13



Distribution of total food details screen usage duration per user

Full size image

Fig. 14



Distribution of total statistics screen usage duration per user

### Full size image

We further evaluated the absolute accumulated usage duration from all users for all three screen types. Primarily, the Figs. [12](#), [13](#) and [14](#) depict a high difference between usage durations of users within each screen type. All three analysed screen types show longer usage durations for 4–5 users, but not all screens are used highly by the same users. User 25 seems to be very active in general (883 s (

~ 15 min), 471 seconds (

~

~ 8 min), and 496 s (

~

~ 8 min)), user 7 seems to favor the home screen with 571 s (

~

~ 10 minutes), user 20 the statistics screen 350 s (

~

~ 6 min), and user 3 the food details 1000 s

~

~ approx. 17 min. Most users have only very short usage durations on all three screens.

Besides the duration, we can also investigate the type of actions conducted frequently on each of these screens. Looking at the most frequently used features within the home screen, we can observe that the most used functions are of administrative nature like the “Add meal button”

and “Navigation item”. The visualized features result in higher frequent use than the “Recommendation button”, but less frequent than central functions like adding meals or using the navigation menu. The most frequently used function in the food/recipe detail screens scrolling the food details view, followed by adding the standard or custom portion sizes of a food item. However, functions related to recipe details or an extended view of influenced nutrients or the info button appear less frequent. For the statistics screen, no actions besides the time frame selection were possible, leading to a clear visual feedback intent for all statistics actions. Regarding this time frame, the most used screen was the weekly overview with 54.8%, followed by daily with 27.4%, and monthly with 17.8%.

Overall, despite the positive usage duration for the three central visual features, the attention on the actual visual feedback seems to be limited. We thus want to compare how the users perceived the screens which we have shown to be both popular and influential. The next section shows the system perception that we examined during the final survey.

### **Perception of the system after long term usage**

System perception was part of our final survey. Since we can only examine the survey by the participants that both finished the survey and used the application, the number of samples is reduced to 12.

First, we look at the popularity of the different feature screens. We showed in the previous sections that the statistics, the nutrient status, and the food details, reachable over the home screen, are the most used screens. During the survey, the users gave the following ratings for how much they liked each screen on a 1–5 Likert scale (Table 5). We see that the frequently used screens are also the ones that reached a rating of around 4. Additionally, the energy overview and the profile were liked but not used that often. The recommendations were the least

popular feature, which might explain the difference in impact between the model of this study's first 12 days and the Dutch study (Schäfer and Willemsen [2019](#)).

Table 5 Popularity and helpfulness of screens on a 1–5 Likert scale

**Full size table**

Second, we look at the perceived helpfulness of screens. We showed in the previous sections that the most influential positive screens were the statistics, home-screen, and the food-details within the food-search. During the survey, the users gave the following ratings on a 1–5 Likert scale (Table [5](#)). The home screen and the statistics have high ratings, confirming their helpfulness. Instead of the food-search and food-details, the users perceived the diary and the nutrient status list as helpful, which had no significant effect in our model.

Table 6 Quality assessment of visualization screens with a combination of several items on a 1–5 Likert scale

**Full size table**

Third, we wanted to take an even closer look at the visualizations and how they were perceived. The users answered several questions to each visualization about perspicuity, helpfulness, and similar items. The overall scores are shown in Table [6](#). As indicated in the previous measures, all the screens received excellent feedback, with the highest rating going to the home screen, the



nutrient status screen, and the statistics screen. We draw a relation between the ratings and the usage behaviour in terms of timely duration on the screens (see Sect. 5.3.4) as the home screen achieves the highest rating in terms of understanding and helpfulness and further belongs to the screens with the longest usage duration. Here, we could associate a longer usage duration with a higher user satisfaction when using the features of the screen. However, this does not hold true for the food-details screen, which receives the lowest rating for understanding/helpfulness compared to the home screen but achieves the second-highest usage duration after the home screen. Here, a longer usage duration could mean that users had problems in quickly understanding the screen. Further, the statistics screen achieved the highest average rating in the quality assessment, although it reached the last place in the overall usage duration. This could probably mean that users appreciated the screen but did not use it longer since the interface contains commonly known graphics (trend lines), which mostly every user can interpret quickly.

Table 7 Quality assessment of recommendations aspects

#### Full size table

Fourth, we wanted to take a closer look at the recommendations and how they were perceived (Table 7). For this, we used the same questionnaire as in Schäfer and Willemsen (2019). The highest positive feedback was given to diversity and overall system effectiveness. Indecisive feedback was given to the difficulty of the recommendations and the system effort, with an average of 3 and a high deviation in answers. Health, taste, and personalization were all graded slightly negatively with values around 2,5 on a scale of 1–5. These constraints might explain the negative feedback on the recommendations. We added the recommendation perception from the short-term study (Schäfer and Willemsen 2019) in Table 7 as a comparison point. Most

perceptions were slightly better in the short-term than in long-term usage. In particular, health, taste, and personalization were perceived better during the shorter study. However, the system effectiveness was regarded better in the long-term, which would be in line with the assessment of behavioural and physical changes.

Finally, we asked users for feedback on the overall application and their intention to use it further. The SUS score of 59 (min. 30, max. 77) is higher than in the pilot study (Leipold et al. [2018](#)). While 59 is lower than the average SUS score of 65, Ferrara et al. ([2019](#)) report SUS scores of comparable commercial applications between 46.7 for MyDietCoach and 89.2 for LifeSum. The users agree the strongest to the application giving insights into their diet (3.9). They agree slightly (3.4) to the app being a useful tool and making it easy to eat healthily. The lowest agreement (3.3) is to the fact that the application improved the diet. This is in line with the standardized measures before and after the study, which showed little effect on visible measures, but some effect on the nutrient intake, which would be less tangible. Overall, half of the 12 survey participants would buy the application. Most of them would agree to pay a one-time fee of 2–5€.

## **Evaluation of recommendations**

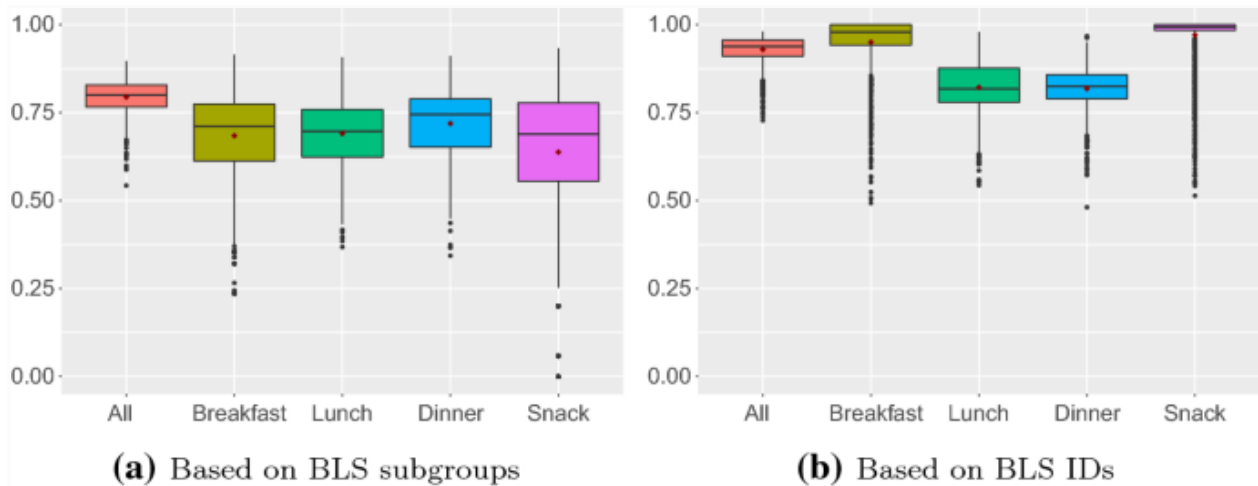
Results from the survey about system perception indicate that users did not perceive the recommendations very positively. One of the critiques is the diversity as only partially diverse, as shown in Table [7](#). Therefore, we evaluated this diversity quantitatively, both on the intra- and inter-recommendation level. We considered a total of 1873 recommendations divided among 28 users with varying levels of engagement with the application. We limited the considered recipes per recommendation to top-ten in each of the four meal categories (breakfast, lunch, dinner, and snack), since depending on screen size three or four recommendations per category are visible, and we assumed users did not scroll through the list indefinitely. The system automatically recreates the recommendation each day, but users could also manually

regenerate. Such a new recommendation would differ from the first one if the food diary entries of the previous three days changed in the meantime. Thus, we considered the last recommendation for each day.

The first consideration was the diversity of items within one recommendation, which we assessed using the Intra-List Diversity (ILD) (Vargas and Castells [2011](#)). As distance metric, we used the cosine distance between vector representations of food items which were based on a binary encoding of the recipe ingredients' identifiers in the BLS database. For this representation, two encoding approaches were used: one only incorporated the first two digits of the BLS code (e.g. "F2"), indicating category and subcategory of the item (in this case "fruit" with the subcategory "stone fruit"), while the other approach considered the entire BLS code, identifying individual food items. In consequence, the former approach considered a wider range of items as similar, as all food items within the same subcategory occupy the same vector position.

The resulting ILD scores for the first encoding approach, computed over all recommendations and divided by meal category, can be seen in Fig. [15a](#). As a comparison, top-10 ILD in a recent music dataset analysis has reached values between 0.39 and 0.57 depending on the reranking strategy (Lu and Tintarev [2018](#)). The ILD score calculated over all items of a recommendation regardless of their assigned meal is higher than the scores computed for meal categories separately, indicating that items within a meal category are more similar to each other. The lower subgroup ILD scores in the snack category are likely caused by a prevalence of fruit in the recommended snack items. At the same time, the snack category has a high ILD for individual IDs due to the majority of snack recommendations consisting of one distinct food item. The ILD scores are stable across users and over time.

Fig. 15



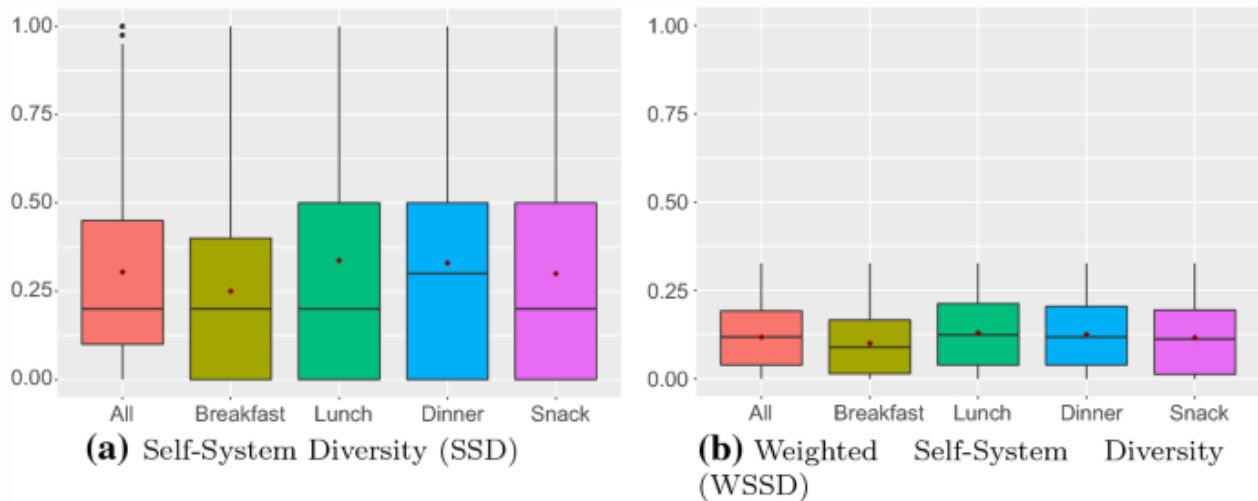
ILD scores over all recommendations and for the meal categories breakfast, lunch, dinner, and snack

**Full size image**

Since users receive recommendations daily, we also consider diversity over time. First, we calculated the Self-System Diversity (SSD) as proposed by Vargas and Castells ([2011](#)), which represents the diversity between two subsequent recommendations for a user. Since the SSD score of the first recommendation of each user would be 1.0 due to a lacking previous recommendation, we omit this value to avoid bias. In the context of meal recommendations, diversity is also essential in a medium-term setting, so we adjusted the SSD measure to incorporate the period of one week. This Weighted Self-System Diversity (WSSD) employs the idea of past recommendations becoming gradually less important for the current recommendation, which is represented through a weighting function based on the physical

concept of atomic decay. For our setting, we chose to consider a period of seven days, i.e. seven previous recommendations, and a half-life of 3 days, meaning that after three days, the weight of the recommendation has halved. The weight indicates how much overlaps in recommended items penalize the diversity score. Similarly to the SSD, we do not consider the WSSD score for the first recommendation of each user, as it has a high positive bias due to missing previous recommendations.

Fig. 16



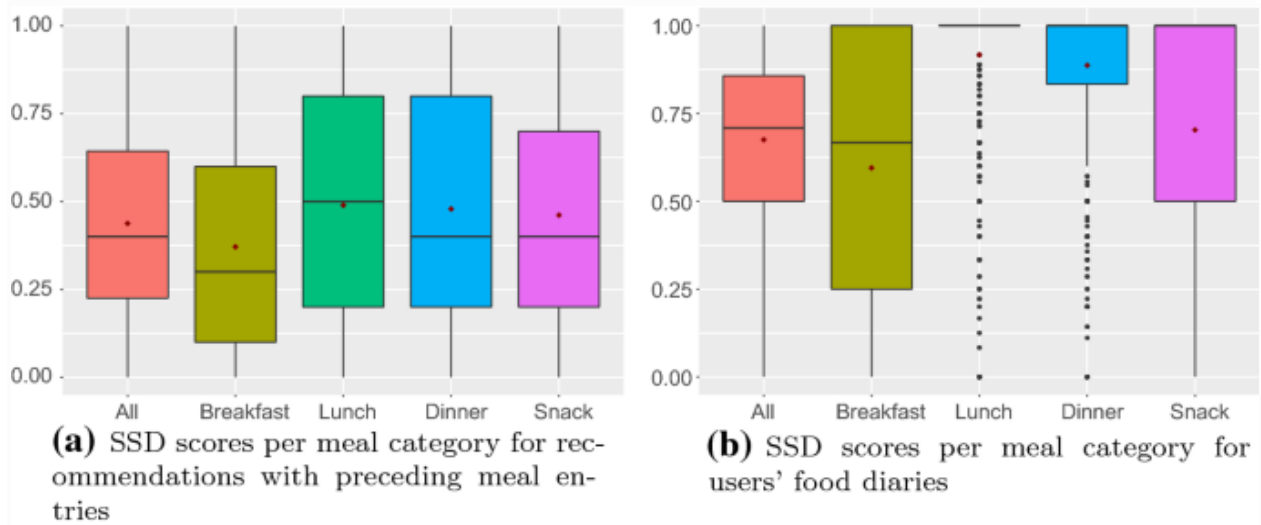
Diversity over time over all recommendations and for the meal categories breakfast, lunch, dinner, and snack

Full size image

Overall, our system achieves a mean SSD of 0.30 and a median SSD of 0.2 over all recommendations. Considering the different meal categories, we can see in Fig. 16a that dinner recommendations had the highest median SSD with a score of 0.3, while the other meal categories achieve a mean of 0.2. As a comparison, temporal top-10 diversity analysis on movie datasets has reached values between 0.1 and 0.4 depending on the used algorithm (Lathia et al. 2010). The mean values, indicated by red markers, show that breakfast recommendations are on average less diverse than other categories. This is less critical since German breakfast habits are generally undiverse as well. Overall, we see that both mean and median values are rather low, showing that diversity over time is not as high as the diversity within recommendations. We achieved both a mean and median WSSD of 0.12 over all categories and can see in Fig. 16b that the WSSD behaves very similar to the SSD, but is lower in general. This is due to the larger number of recommendations being considered, and the, therefore, higher chance of overlaps between recommendations. It indicates that the recommendations of our application were more repetitive in the long run. Interestingly, when considering meal categories separately, we see differences to the SSD metric. While for the SSD, dinner recommendations achieved a higher median value, the WSSD is not higher compared to other categories, with median scores rounding to 0.12 for lunch and dinner, and 0.11 for snacks. This indicates that while dinner recommendations are more diverse in the short term, there are repetitions over longer time periods.

There is a clear relationship between user activity and SSD. When the user's meal tracking activity, here measured in food diary entries, stops, the SSD score decreases with a slight delay, since the recommender system considers the user's previous nutrient intake over three days. As soon as user activity resumes, the SSD increases again. Looking at all participants, we also see that users with a higher number of active days (days on which they tracked at least one meal) also tend to have a higher mean SSD. The WSSD decreases in the period of inactivity, albeit more slowly than the SSD score since the score takes a longer time period into account. The same trend as for mean SSD scores for all users can also be observed for mean WSSD scores, with more active users having higher scores.

Fig. 17



Comparison of SSD between cleaned set of recommendations and users' food diaries

**Full size image**

We have seen that diversity over time is related to user activity, specifically meal tracking. We have also observed that some users tracked their meals sporadically, which is bound to impact recommendation diversity. As discussed in Sect. 4, the recommender system relies on the users' food diaries to gauge both nutrient intake and dietary preferences, considering the past three and seven days, respectively. To illustrate the dependency on users' meal tracking, we calculate

the SSD as before, but only considering recommendations on days that have at least one food diary entry in the preceding seven days. This decreases the number of recommendations considered from 1873 to 1046 for the calculation over all meal categories. Figure [17a](#) shows the SSD scores per meal category based on the selected recommendations. If we compare the results in SSD and WSSD scores to the previous ones, we see an increase in both despite the low minimum requirement of one meal entry within the previous week. The mean SSD score over all recommendations is 0.44 (previously 0.30), and the mean WSSD score over all recommendations is 0.18 (previously 0.12). Finally, we compare this diversity with the diversity of the users' actual dietary behaviour by applying the Self-System Diversity to users' food diaries. Since missing diary entries on the previous day would result in an SSD score of 1.0, we only consider entries with preceding meal entries. Figure [17b](#) shows a generally diverse diet with apparent differences between meal categories. Revisiting the SSD scores in Fig. [17a](#), we can see that the trends in users' eating habits are reflected by the recommendations' diversity score.

### **Contexts and motives from interview data**

This section describes the interview results. The order of topics is aligned with the previous sections. This section will name the most frequent and relevant topics from coding the interviews. Frequencies are always shown as (topicCount/interviewCount). Additionally, important topics are exemplified by: *literal citations translated to English*. First, we want to look into the effects of the application by analysing the interview data on perceived physical, psychological, and cognitive changes. Second, we want to look into the users' dietary change and their perception of it by analysing the interview data on their nutrition changes and their difficulties in combination with the dietary change questionnaire (Shannon et al. [1997](#)). Third, we want to look into the users' interaction patterns by analysing the interview data on their interaction habits and their system feedback. Fourth, we want to take a deeper look into the



application perception and intention of use by analysing the interview data on comparing the application to previous or future systems.

Regarding the knowledge on nutrition, the users report having learned about their own nutritional needs (4/8) and the nutritional content of food items (4/8): *I noticed that when you eat a lot of liver or paprika, vitamin A rises brutally*. Three participants (3/8) even gained transfer knowledge on which food items to include into their daily habits: *the consciousness, yes now you could add a spinach, or maybe leave it out, something like that*. For physical changes, (2/8) participants report weight loss, while (2/8) participants report no changes at all. (3/8) participants report increased physical activity ("the app also encourages you to do 30 minutes instead of 20 minutes, and then you get a few green calories on top of it."). Most participants report no psychological changes due to the application (5/8). Some participants report being annoyed by the tracking (4/8). Still, others are reporting (3/4) feeling more empowered: *I think because of the app I now know sometimes automatically, what is good and what isn't*.

The three most important changes in diet reported in the interviews or surveys are eating more vegetables (7/13 and 3,6), eating more fruits (6/13 and 3,5), and eating less food in general (6/13 and 4,1): *I always just looked what kind of nutrients were missing, and if there was any orange there, I ate another orange*. After the middle of the study, almost half of the participants reported having no or minor difficulties with keeping these changes (4/9). The most common reasons for relapse were different types of extraordinary situations (4/9) such as events, illness, holidays, and weekends, favourite food items (3/9), social factors (3/9), and general high effort (3/9): *So I always found it difficult, for example, when we were visiting my parents because there is always cake and so on*. At the end of the study, some participants started to feel a sense of helplessness when trying to reach healthy nutrition (3/12): *And that you never manage to bring everything into line. This is an impossibility in practice*. The relapse reasons are also similar to before with (5/12) extraordinary situations, (4/12) social factors, (3/12) general high effort, and (2/12) favourite food items.

Regarding the interaction timing, the participants are mostly split between entering their meals directly after the consumption (7/13) and entering everything in the evening (5/13). Regarding the features, both the tracking (6/13) and the feedback on the home screen (6/13) are commonly used. Other important features that are mentioned by multiple participants are the energy history (3/13), the food item information (4/13), and the nutrient history/statistical view (3/13). Most of the participants used the application for feedback on their past nutrition, mostly with the home screen (7/12): *Well, it wasn't bad, it always said that you should eat a little more in that area because I need more calcium, or this and that. So then you were also happy when that maybe changed from red to yellow.* Others used the application to decide on the next meals with both the recommendations (2/12) and the energy overview (2/12) or for longer-term planning by looking into the suggested food items for each nutrient (3/12): *And in the evening I always looked again to see where the nutrients were deficient. And then for the next day, I said, come on, if there is a fish in the canteen, then I have a fish.* The participants report the visualization being helpful as feedback (6/12), as decisional support on choosing food items (5/12), and as a source of detailed information (3/12). The participants used the recommendation either as inspiration (4/13), as a fitting suggestion (2/13), or as a literal guideline (2/13): *Maybe tried two, three, four things and then I saw that there was a lot with potatoes. Then I thought I'll have something with potatoes. You already got ideas like that.* Some participants thought the recipes were strange (6/12), while others said they did not meet their preferences (6/13): *Otherwise I should always eat tarte flambée rolls or something. I haven't done that yet, because I don't quite understand why, I think it's about a vitamin A story.* Others explained that they could not use the recipes because of external circumstances (4/13) or lack of understanding (3/13). Regarding changes in interaction, participants reported having reduced the amount of time they spend with the application both in the middle of the study (4/8) and at the end (4/13). Almost all participants reported problems with finding food items in the BLS database (11/13): *For example, when I ate in the restaurant, it was sometimes difficult to enter what I actually ate.* Many (9/13) complained about the general effort of tracking their diet.

Most of the participants have not used an application for nutrition before (8/12) since they either have not used any tracking before the study (4/12) or have tracked manually (4/12). The ones that had prior experience list the detailed tracking functionality (3/8) and the nutrient optimality feedback (3/8), the tracking of physical activity (2/8), and the statistics overview (2/8) as unique selling points of *Nutrilize*: *So I liked this (nutrient) optimization very much because I haven't found it with the other apps until now.* (9/12) participants would use *Nutrilize* in the future, but less frequently (3/9) or not always (2/9): *Yes, I would very much like to keep making the entries and see how that develops, with my weight and the other indicators.* Suggested improvements are faster performance (5/12), more content (4/12), and adding own recipes (4/12): *But it would make the whole thing more charming if you could say in one fell swoop, I eat this cereal every morning, so I can pull it over like that.*

## Discussion and limitations

The following section discusses the implication of our results on the initial research questions about the system's influence on the users' (a) physique, (b) nutrition behaviour, (c) system interaction, and (d) system perception, as well as the contextual limitations in real-life settings.

### Changes in physique

There were only minimal changes in the physique of participants. This is partially due to the duration of the intervention and partially due to the focus on nutritional health instead of weight loss. With only minimal changes in energy intake, no physical changes were to be expected. The effect is further similar to Celis-Morales et al. (2016), where the mean BMI at month three is even increased compared to the beginning of the study.

### Changes in behaviour

Time played an important role in the long-term regression on the optimal nutrient intake, which was not the case for the short-term study. There is also a difference in the nutritional ability changes between the control group and the intervention groups. These observations should be taken with care, though, since the large dropout might have led to a survivor's bias in the evaluation, as shown in previous nutrition interventions (Wu et al. [2009](#); Mateo et al. [2015](#)). The evaluation of optimal intake is further based on the daily tracking data, which had been shown to deviate largely from the standardized food frequency questionnaires in both energy and nutritional content. While the behaviour was adapted, the in-depth interviews revealed that the personalized micro-nutrient feedback in both visual and recommendation form was rarely followed consciously. Instead, participants report changes such as eating more fruits or lower portion sizes. Such changes could have been suggested to the user without the need for personalized suggestions that require high user effort. Future systems should make complex personalized recommendations more operationalizable with explicit change suggestions based on habits, such as your nutrients would improve if you leave out the butter in your regular breakfast sandwich, and even stronger explainability and transparency features, such as your calcium intake is always too low because you rarely eat dairy products.

### Interaction with system

The largest part of system interactions is concerned with the daily tracking of nutrition. The interactions show the gradual dropout of participants with only half as many unique users per day after around four weeks. Concerning the feature screens, a strong focus is set on the visual feedback screens. This is further reflected in the impact of different screens on the optimal nutrient intake where tracking and visual feedback showed a significant positive influence. The home screen, statics screen, and food details screen had the strongest influence on optimal nutrient intake within the visual feedback. The timely usage behaviour of the three analysed feedback types showed us that both feedback types based on reflection-in-action (home screen, food/recipe details) resulted in a higher overall usage duration than the statistics screen that is

based on reflection-on-action. We see that besides one consistently active user, the preferences for these visualizations differed between users. We further observed that both the home-screen and the food-details screen were mainly used for tracking purposes while only briefly interacting with the actual visualizations. While the statistics showed the lowest usage duration of all three screen types, it is the only purely visual interaction and therefore highly relevant. This is reflected very positively in terms of estimated usefulness in the interviews. While the interactions are in line with the perception feedback and the impact on behaviour, this in-depth analysis of individual interaction paths revealed false-positive associations with features that were not consciously used. Future systems should incorporate visual feedback in an isolated fashion to further confirm their positive impact on behaviour change.

#### Perception of system

We can conclude from our results that the perceived usefulness of the visualizations was higher than of the recommendations, although carefully considering design elements such as personalization on different levels like personal preferences related to taste and individual health parameters. Based on our results on system perception, we can assume that the visualizations represent a proper means to communicate personalized data related to an individual diet. Rooting the design of visualizations in theory-related concepts like reflective practice may further contribute to the positive influence on nutritional ability, positive perceptions, and observed user behaviour. Specifically, the possibility for reflection on personal behaviour through the offered visualizations could positively influence users' perception of usefulness since direct feedback on their own behaviour creates awareness, understanding, and a basis for future actions. Different visual features as complementary components that support transparency, trust, and empowerment for the recommender system should be systematically evaluated in the future.

We can relate the low perceived helpfulness of the recommendations in part to their diversity evaluation. The measured intra-list diversity is good, which is reflected in the users' perception. On the other hand, the diversity between recommendations over time is low. Some users have mentioned this repetitiveness as a reason for not using the recommendations. However, we could see a relation between diversity over time and the users' meal tracking activity. Users who tracked their dietary intake more diligently received more diverse recommendations, as the recommender system heavily relies on knowledge of the user's dietary intake to recommend personalized meals both in terms of taste preferences and nutritional needs. Regular tracking of dietary intake is laborious (Maia and Ferreira [2018](#)). As a result, some users do not maintain their food diary, which was also shown by drastically decreasing tracking activity in our study. This does not only lead to a low recommendation diversity over time but also reduced personalization, which matches the users' slightly negative perception of this aspect. The dependency on well-kept food diaries, albeit theoretically achieving high degrees of personalization and accuracy, is arguably the main drawback of *Nutrilize's* recommendation feature. Future systems should consider implementing temporal and intra-list diversity metrics for long-term usage as well as approximating missing dietary diary information by default or substitute learning (Liu et al. [2019](#)).

#### External influence factors in long-term usage

Dietary preferences are complex and based on diverse and dynamic factors, such as cooking method, cost, availability, and type of cuisine, as well as social, psychological, cultural, and environmental factors (Freyne and Berkovsky [2010b](#); Maia and Ferreira [2018](#); Min et al. [2019a](#)). While the recommendation feature of *Nutrilize* took dietary preferences into account, aspects like cost, availability, or social or psychological context were not incorporated. This might explain why some users stated a mismatch with their preferences as a reason for not using the recommendations. During our in-depth interviews, the participants report that factors limiting the app's success are stress, tracking effort, social situations, and exceptional events (e.g.

holiday/birthday). All the reported issues lead back to a higher need for personalization and context awareness, as has been suggested by previous work (Rokicki et al. [2016](#); Kusmierczyk et al. [2015](#)). They further stated to not feel constrained by the application and to focus on substituting individual food items, especially vegetables and fruits, rather than focusing on improving single nutrients. They felt that the application enhances the pleasure of eating and their knowledge of nutrition. To avoid relapse and dropout due to social and resource-based context, future systems should aim at sensing, predicting, and proactively incorporating such contexts. For example, negotiations of meals before or after the context could be offered based on a predicted event.

## Limitations

The *Nutrilize* application has constraints regarding usability and feature availability. First, our aim at high precision nutritional content has led to using the BLS as a food item database. This has led to issues with non-layman terms and thus low searchability for certain food items. Furthermore, the restriction to open source data has led to the integration of recipes from the KochWiki database, which covers many basic recipes, but not as many “currently popular” ones. We also see in the effect analysis that the underestimation in daily tracking is propagating errors to the feedback users are receiving. We try to correct for this in our models by including the daily calorie count. Additionally, the study we conducted only had a limited number of participants which does not allow us to conclude or generalize any results with certainty. Finally, while the duration of the study is uniquely long for mobile applications, it is still relatively short compared to traditional interventions. Despite these limitations, our study reveals some systematic challenges in the long-term usage of food recommender systems that are common to other similar applications.

## Conclusion and future work

---

In this paper, we presented a long-term mixed-method study of the usage of our nutrition assistance system *Nutrilize*. Our *Nutrilize* system provided dietary tracking, prospective, retrospective, and perspective visual feedback, as well as personalized recipe recommendations. We investigated the changes in physique, the nutritional behaviour changes, the system interaction, and the perception of users for 34 participants (11 control, 11 3-month, 12 2-month).

Our results suggest that the system is effective in changing behaviour but struggles with high dropout due to different contextual factors. The changes in physique were minor and without clear differences between the groups due to the low sample size and short study period. As our results have shown that participants specifically appreciated the visual feedback, we suggest that recommender systems should consider visual feedback as an integral part to serve awareness, reflection on behaviour, and educational content to enhance nutrition-related knowledge. Further investigations into the potential of automated personalized feedback towards users' action (Rabbi et al. [2015](#)) would be helpful to determine which visual feedback serves users with the most decisional support. The personalized recommendations provided diverse daily sets of recipes but were repetitive over time and could not always meet user preferences. The repetitiveness is partially attributed to the dependency on users' food diaries, which were sometimes incomplete due to high tracking effort. Alternative methods to elicit user information could help mitigate this issue if they are less laborious for users but might come with a decrease in accuracy. Such a method could, for instance, be dietary tracking based on photographs of meals, as proposed by Ming et al. ([2018](#)). To adapt the recommendations more to dynamic user preferences, methods to circumvent the problem of uncertainty and missing information as presented by Schäfer et al. ([2017a](#)), as well as approaches to model more context factors such as social context could be incorporated. Alternatively, the system could account for needed flexibility by letting users choose dishes according to their current preferences and then adapting them based on individual nutritional needs by substituting ingredients or varying ingredient quantities. This results in a trade-off between maintaining the recipe's culinary quality and achieving a nutritional impact.



In summary, our study shows different challenges that health-focused nutritional assistance systems face when being used in the long term. Our findings can be used to improve future system regarding their impact in the long-term and to postulate more long-term evaluation of recommender approaches.

## References

- 
- Achananuparp, P., Weber, I.: Extracting food substitutes from food diary via distributional similarity (2016). arXiv preprint [arXiv:1607.08807](https://arxiv.org/abs/1607.08807)
  - Alrige, M., Chatterjee, S.: Easy nutrition: a customized dietary app to highlight the food nutritional value. In: Chatterjee, S., Dutta, K., Sundarraj, R.P. (eds.) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). LNCS, vol. 10844, pp. 132–145. Springer, Berlin (2018). [https://doi.org/10.1007/978-3-319-91800-6\\_9](https://doi.org/10.1007/978-3-319-91800-6_9)
-