Examining the User Evaluation of Multi-list Recommender Interfaces in the Context of Healthy Recipe Choices

ALAIN D. STARKE, Amsterdam School of Communication Research, University of Amsterdam, Netherlands and MediaFutures, Department of Information Science and Media Studies, University of Bergen, Norway
EDIS ASOTIC, University of Bergen, Norway
CHRISTOPH TRATTNER, MediaFutures, Department of Information Science and Media Studies, University of Bergen, Norway
ELLEN J. VAN LOO, Marketing and Consumer Behaviour Group, Wageningen University & Research, Netherlands

Multi-list recommender systems have become widespread in entertainment and e-commerce applications. Yet, extensive user evaluation research is missing. Since most content is optimized towards a user's current preferences, this may be problematic in recommender domains that involve behavioral change, such as food recommender systems for healthier food intake. We investigate the merits of multi-list recommendation in the context of internet-sourced recipes. We compile lists that adhere to varying food goals in a multi-list interface, examining whether multi-list interfaces and personalized explanations support healthier food choices. We examine the user evaluation (i.e., diversity, understandability, choice difficulty and satisfaction) of a multi-list recommender interface, linking choice behavior to evaluation aspects through the user experience framework.

We present two studies, based on 1) similar-item retrieval and 2) knowledge-based recommendation. Study 1 ($N = 366$) compared single-list (5 recipes) and multi-list recommenders (25 recipes; presented with or without explanations). Study 2 ($N = 164$) compared single-list and multi-list food recommenders with similar set sizes, but varied whether presented explanations were personalized. Multi-list interfaces were perceived as more diverse and understandable than single-list interfaces, while results for choice difficulty and satisfaction were mixed. Moreover, multi-list interfaces triggered changes in food choices, which tended to be unhealthier but more goal-based.

CCS Concepts: • **Applied computing → Consumer health**; • **Information systems → Recommender systems**.

Additional Key Words and Phrases: recommender systems, health, recipes, user evaluation, multi-list recommendation, explanations, choice overload, food choice

ACM Reference Format:
Alain D. Starke, Edis Asotic, Christoph Trattner, and Ellen J. Van Loo. 2022. Examining the User Evaluation of Multi-list Recommender Interfaces in the Context of Healthy Recipe Choices. *ACM Forthcoming* 1, 1, Article 1 (January 2022), 31 pages. https://doi.org/10.1145/XXXXXXX.XXXXXXX

Authors' addresses: Alain D. Starke, alain.starke@uib.no, Amsterdam School of Communication Research, University of Amsterdam, P.O. Box 15791, Amsterdam, Netherlands, 1001 NG and MediaFutures, Department of Information Science and Media Studies, University of Bergen, Lars Hilles gate 30, Bergen, Norway, 5008; Edis Asotic, edis.asotic@gmail.com, University of Bergen, P.O. Box 7802, Bergen, Norway, 5020; Christoph Trattner, christoph.trattner@uib.no, MediaFutures, Department of Information Science and Media Studies, University of Bergen, Lars Hilles gate 30, Bergen, Norway, 5008; Ellen J. Van Loo, ellen.vanloo@wur.nl, Marketing and Consumer Behaviour Group, Wageningen University & Research, P.O. Box 8130, Wageningen, Netherlands, 6700 EW.

XXXX-XXXX/2022/1-ART1 $15.00
https://doi.org/10.1145/XXXXXXX.XXXXXXX

1 INTRODUCTION

An increasing number of commercial recommender applications present multiple recommendation lists in a single interface [34]. So-called 'Multi-list Recommender Interfaces' present item lists stacked on top of each other, accompanying each list with an explanation on what the items in the list represent [22, 68]. The algorithms underlying these lists are typically either based on a variety of recommendation approaches (e.g., using different similarity measures [22, 34]), or employ a single personalization algorithm that is optimized differently across different lists, by constraining the presented items to a certain tag [58], or by re-ranking the top-k set on a specific attribute (cf. [70]).

Commercial examples include video streaming services, such as Disney+ and Netflix. They present movie and TV series recommendations in an explainable multi-list interface [22], mostly providing multiple lists that relate to a user's preferences and which are limited to or optimized for a specific attribute, tag, or genre. For example, lists in Netflix would be explained as 'Drama TV Series' (genre constraint), 'Oscar-winning movies' (movies with a specific tag), or 'Recommended for you' (Collaborative Filtering with no constraints). The 'sub-lists' presented within a multi-list recommender interface can be extensive: Netflix presents approx. 40 different lists on a user's page with up to 75 recommendations per list [22].

Multi-list interfaces may also promote items that are not personalized. For example, e-commerce platforms such as Amazon may display lists that prioritize items based on their overall popularity, because they are often purchased in the past 24 hours, or because many users have added them to their wish lists. Such lists are inferred without any user history, yet can still lead to changes in user preferences by presenting a larger number of items in an organized manner (cf. [58]).

An illustrative example of what constitutes differences between single-list and multi-list interfaces is depicted in Figure 1 [34]. The main distinction between them in this study is the use of multiple recommender algorithms per multi-list interface, which is also how the landing pages of some commercial applications are designed [22]. It should be noted that what we describe as single-list interfaces is at times referred to as 'grid user interfaces' or 'grid UIs' in other domains [39], while other studies use the concept of a 'single list' or 'single-list interface' for lists in which only a single option is presented per line [9], as is common in search box applications (cf. [70]).

The application of multi-list interfaces has particularly expanded in commercial domains. Whereas their use in online retail and on video streaming platforms has become more prevalent [22, 59], research on its use in domains where users have specific behavioral goals is missing [20, 68]. Food is such a domain, where multi-list interfaces have the potential to steer user preferences towards a specific eating goal. Whereas in the movie domain, the explanations may signal a particular genre or mood that may be appealing to a user [22], food choices often face a tradeoff between health and popularity (or taste) [64]. Explanations may help to mitigate this ambiguity, in attempt to align with user goals.

The promotion of healthy food choices has hardly been examined in food recommender studies [60], because many approaches are popularity-based and lead to unhealthy outcomes [18, 76]. Since a user's profile becomes less relevant when she wishes to change her current eating habits [1, 63], it is often hard to generate relevant recommendations when, for example, a user takes up a new weight-loss goal or starts to attain a vegetarian diet. While providing more control could be one way to circumvent unhealthy recommendations (e.g., in the medical domain [47]), other studies have shown that increasing recommendation diversity could better serve a user's interests [59]. This can be provided by multi-list interfaces that optimize for different types dietary restrictions (e.g., lactose-free and vegan) or nutrient intake (e.g., fewer kcal or more fiber). Moreover, providing appropriate, personalized explanations or justifications might further steer users towards healthier options [50].

The commercial multi-list 'benchmark' has yet to be evaluated in a user-centric approach [34]. Whereas its merits are clear in terms of user retention and click-through rates [22], much less is known about how users perceive the different aspects of multi-list recommender systems and how this is related to their choices. For example, do users understand the recommendation lists presented to them and does this affect from which list they choose an item? And, are multi-list interfaces *only* evaluated more favorably, or do they also lead to healthier choices and choices that match a user's eating goals? And, does this depend on the type of explanation presented?

This paper presents two novel studies with multi-list food recommender interfaces, which are each evaluated through a user-centric approach. We employ the user experience recommender framework [43, 44] to asses whether the use of multiple lists in a single interface, along with explanations, leads to changes in user choices and whether these are linked to changes in how users perceive and experience the multi-list interface. To date, only a few studies have examined the relation between user evaluation aspects and multi-list interfaces. Pu and Chen [58] compare a single-list interface with simple explanations to a category-based interface in the personal computer domain, accompanying each list with an explanation on its contents. They show that a multi-list interface is perceived as more helpful, as users could compare items more easily, even if time spent on making a decision was equal across both interfaces. Moreover, a related study by Nanou et al. [52] shows that a genre-grouped movie recommender interface is evaluated as easier to use, due to a reduced cognitive load.

The premise of earlier work on multi-list recommender interfaces is to increase diversity while reducing choice overload [30]. However, it is to date unclear how specifically choice difficulty and satisfaction are affected by the presented recommender interface, akin to work of Bollen et al. [4]. For the food domain, we expect that a multi-list recommender system can overcome human biases towards unhealthy foods through their justifications, and lead to more satisfactorily choice outcomes by increasing the diversity of the presented recipes. Since many people lack the sufficient nutritional knowledge to make healthy food choices [31], the introduction of list-specific

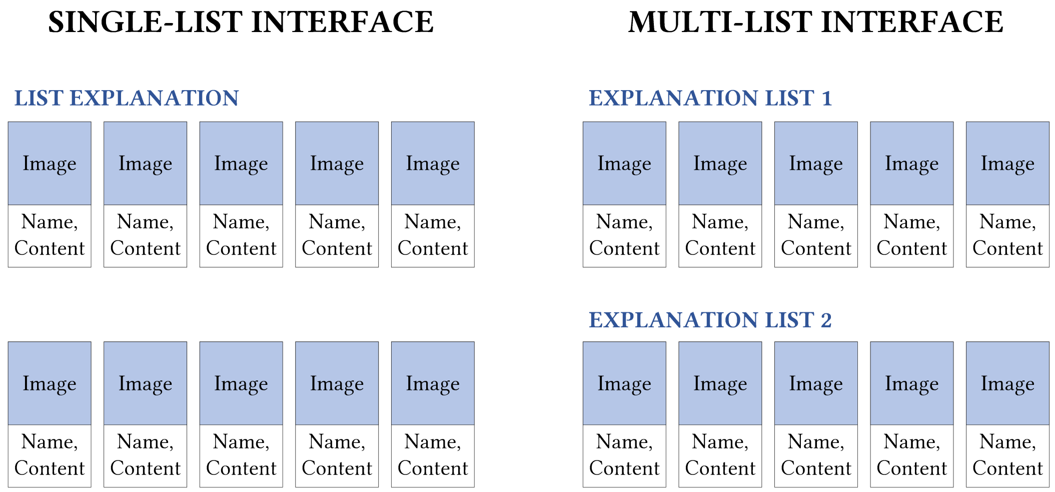


Fig. 1. Schematic mock-up of what is referred to as single-list interfaces and multi-list interfaces in this paper. The main distinction is the use of multiple recommender algorithms (one per list) in the multi-list interface.

explanations is expected to boost its understandability, also given earlier findings on the reduction of cognitive load [52, 59]. Moreover, attribute framing theory [2] suggests that nutrition-based explanations could make users pay more attention to healthy or nutrition-related aspects when choosing a recipe, particularly if they are related to personal characteristics (cf. [50, 71]). There has, however, been little attention for how explanations in such an interface should be designed and to what extent they can persuade users to consider other content, such as healthier recipes.

We expect that a multi-list interface, bringing forth a more diverse recommendation set, is more likely to cater towards eating goals that are not yet part of the user's profile. In terms of the interface, the most important contribution is that we can highlight different nutrient-specific eating goals, by presenting lists that optimize for recipes with fewer calories, less fat, or more fiber. Moreover, based on earlier work with explanations that personally relevant explanations (e.g., 'healthy recipes that are in line with your healthy eating habits') rather item-focused explanations (e.g., 'healthy recipes that meet dietary intake guidelines') positively affect a user's evaluation of a recommender system [46], we examine the merits of using personalized explanations in a multi-list interface. For the user-centric evaluation of our multi-list food recommender system and whether it can support healthy food choice and eating goals, we propose the following research questions:

[RQ1]: To what extent is a multi-list recommender interface with (personalized) explanations evaluated more favorably in the context of the user experience recommender framework, compared to a single-list interface without (personalized) explanations?

[RQ2]: To what extent can a multi-list recommender interface with (personalized) explanations support different user goals and healthy food choices, compared to an interface without (personalized) explanations and a single list?

We present two studies. In Study 1, we examine both research questions in a recommender system based on similar-item retrieval, using a US-based dataset from the recipe website AllRecipes.com [75]. We compare single-list (5 items) and multi-list (5x5 items) interfaces, either accompanied by list-based explanations or not, by assessing a user's evaluation and choices through Structural Equation Modelling. We find that multi-list interfaces are evaluated more favorably in terms of diversity and choice satisfaction, but also lead to a higher choice difficulty. In addition, they support healthy eating goals for specific users. In Study 2, we expand our findings for both research questions by developing a knowledge-based recommender system, using an Italian-based dataset from the recipe website GialloZafferano.it. We compare single-list and multi-list interfaces of similar length (25 items), which are accompanied by either non-personalized or personalized explanations. We find that multi-list interfaces increase the perceived diversity and understandability, but slightly decrease choice satisfaction and the healthiness of chosen recipes. In contrast, choice difficulty is unaffected across single-list and multi-list interfaces with similar set sizes.

2 BACKGROUND

This section provides more background on the key themes in this paper. We describe studies from the recommender systems domain, highlighting contributions on explanations, food recommendation and multi-list recommendation. Moreover, we highlight work on choice overload, mainly in the context of recommender systems.

2.1 Food Recommender Systems

Food recommender systems have been used for decades. The earliest meal-planning systems stem from the 1980s [24, 27]. The more contemporary food recommender systems can be categorized into three types of approaches [49]. The first type optimizes recommendations towards a user's preferences, i.e., based on ratings, likes or bookmarks [76], and is similar to common approaches in other recommender domains [45]. The second type of approach is arguably more distinct for the food domain, for it focuses on meeting the nutritional needs of the user [51, 61]. A third type of approach aims to balance user preferences and nutritional needs [6, 74], often by combining user preference data and user constraints, such as food allergies [50].

Food recommender systems that focus on user preferences generally suggest foods that a user is most likely to enjoy. Many approaches involve collaborative filtering and content-based recommendation, leveraging user ratings, likes, or similar-item retrieval [76]. Freyne and Berkovsky [18] evaluate the performance of different approaches: content-based, collaborative filtering, and hybrid algorithms. The best performing approach is found to be content-based, which deconstructs recipe ratings into ingredient ratings. It shows that users tend to prefer recipes with similar ingredients (e.g., potatoes) as ones liked in the past. Approaches employed in other studies have also incorporated negative feedback, for example by using a hybrid approach of Singular Value Decomposition with user and item biases [26].

Recommenders that are optimized towards the nutritional needs of the users tend to employ different recommender methods [61, 74, 76]. As shown by the popularity of unhealthy recipes [75], many users do not consider the nutritional content of internet-sourced recipes, which could be supported by recommender systems [74]. Mankoff et al. [48] present an early example of a nutrition-based food recommender, generating food recommendations using supermarket receipts. Other methods are more goal-based, by allowing users to disclose specific health issues, which in turn triggers the recommender to avoid nutrients that co-occur with such health issues [81]. To this end, a suitable recommender approach is knowledge-based recommendation, which matches recipes to users based on the recipes' nutritional content and the user's health-related characteristics and needs, according to a specific set of rules [51]. Although most work in this area is new, recent studies show the possibilities to support healthier food choices through the use of explanations, particularly when it can draw upon a multitude of user and food-related features [50].

Many of the recent food recommender studies aim to optimize between nutrient intake and food preferences [64, 76]. Whereas some approaches balance these two factors simultaneously when retrieving recipes [21], most approaches rely on either pre-filtering or post-filtering based on one or more health indicators [3]. Pre-filtering approaches impose constraints before user preferences are considered, such as by filtering for specific dietary restrictions (e.g., halal or vegetarian) [73, 85]. A more common approach is to apply post-filtering in food recommender systems [76], retrieving first a set of recipes that are relevant to a user's food preferences, after which a nutrient-based or health-based re-ranking can be applied [66, 75].

In our experimental evaluations, we apply two different recommendations approaches. In Study 1, we retrieve recipes on title-based similarity, which are post-filtered based on specific nutrients. In Study 2, we present a knowledge-based recommender that generates either personalized or non-personalized explanations in a multi-list context.

2.2 Choice Overload in Recommender Systems

Choice overload refers to experienced decision-making difficulty due to an overabundance of rather equivalent options to choose from [62]. Although there has been discussion about the main mechanisms of choice overload [10, 11], and under what circumstances the general effect replicates [11], an tradeoff often observed is that between the need for a large choice set and the difficulty in choosing from such a large choice set.

Although the very first claims about choice difficulty stem from the 1300s [62], a contemporary milestone study was conducted by Iyengar and Lepper [32]. They examine the choice overload hypothesis by conducting three studies in both field and laboratory settings. They demonstrate the tradeoff between the desirability of larger choice sets and the negative consequences with regards to choice difficulty, choice deferral, and a decreased experienced satisfaction. This can be attributed to the positive expectations an individual has for finding a product in large assortments, which can in turn lead to a stronger disappointment if a match is not found [14]. In later research, the choice overload phenomenon is claimed to depend on four factors [11]: choice set complexity, task difficulty, the extent to which one's preferences are set, and one's decision goal.

A few studies have directly addressed choice overload in the recommender systems literature. Bollen et al. [4] have conducted online user evaluation studies to investigate to what extent users of movie recommender systems with different algorithms (top-N lists vs lists with mixed quality) and set sizes (5 vs 20 items) also suffer from choice overload. In doing so, they inquire on the user's perceived variety, attractiveness of the recommendation set, choice difficulty and satisfaction with chosen item. They find that larger sets with only good movies (Top-N list) does not necessarily lead to higher levels of choice satisfaction when compared to smaller set sizes. Instead, there is a trade-off between the increase in perceived set attractiveness and the increase of choice difficulty. Other studies have also found that higher-quality personalization can lead to information overload, stimulating the user to seek more alternative before making a final decision [12]. It must be noted that satisfaction and choice deferral are among the main measures with which choice overload can be assessed across all domains [11].

2.3 Multi-list Recommender Systems

Although recommender systems are touted as a general solution for choice overload [33], some studies have examined it specifically. To this end, multi-list recommender systems are thought to be solution to the dilemma of providing more content at the cost of choice difficulty [68], which are thought to increase perceived diversity and choice satisfaction without drastically increasing the difficulty of making a decision. The latter would be at odds with the findings of, for example, Bollen et al. [4].

Although the topic is still new [34, 68], a few researchers have studied multi-list or organization-based interfaces before their commercial use became widespread. Most notably, Pu and Chen [58] show that an interface similar to what we consider a multi-list interface today, e.g., Netflix [22], is perceived as more helpful than a single-list interface, as it allows for easier comparison between items, while inducing less cognitive strain, even if time spent on making a decision was equal between the two interfaces. Such multi-list interfaces are often accompanied with short explanations on what each list contains [7, 52, 58], or the difference between them. For example, Netflix's landing page consists of approximately 40 different lists, which each contain up to 75 items per list, retrieved by a single algorithm per list [22]. Each list is annotated by an explanation that can describe a myriad of aspects related to the presented content and the user, such as describing the user's past behavior or highlighting a movie genre. Similar landing page design is now used at other video-streaming platforms, such as Amazon Prime [84].

It is to date, however, unclear what the exact merits are of such a multi-list design, particularly when combined with personalized explanations. Academic research is scarce, for most publications on such multi-list home page design mainly covers the algorithmic intricacies, not the user's evaluation compared to other interfaces. Since 2020, research has started to compare the merits of single-list and multi-list interfaces. Jannach et al. [34] have explored the effects of multi-list interfaces for similar-item recommendations in the movie domain. In their online evaluation study, they compare a long single-list with line breaks to a multi-list interface separated by labels. They find that single-list interfaces are less effortful to use when choosing an movie, compared to a multi-list interfaces. In contrast, multi-list interfaces do allow for more exploration, leading to a higher levels of perceived diversity and novelty.

In our experimental evaluations, we examine to what extent multi-list interfaces are evaluated more favorably than single-list interfaces in the context of food recommendation. In Study 1, we compare interfaces with different set sizes, presenting either with or without explanations. In contrast, we compare interfaces with similar set sizes in Study 2.

2.4 Designing Explanations

As users might have varying eating goals, designing effective explanations is arguably important in the context of food multi-list recommender systems. Tintarev and Masthoff [71] provide guidelines on designing a 'good' explanation, as well as on how to evaluate them. It is important to note that the evaluation of explanations is confounded on the recommender approach, as well as on how the presentation of recommendations interacts with them. Moreover, the relationship between the algorithm and the type of explanations to be generated also needs to be considered.

In the context of supermarket food purchases, most countries require food products to disclose ingredients and nutritional content, such as the amount of calories and sugar [83]. This often involves a detailed description on the back of a product. Although such product information helps experiences customers [23], it is often simplified through the use of Front-of-Package (FoP) labels with health scores, such as the Nutriscore [35]. Although the use of FoP labels in online contexts is still not at the level of brick-and-mortar supermarkets, an increasing number of studies are conducted at the intersection of health-based explanations and online food content [25]. Recently, health-based explanations or justifications have been used in a knowledge-based recommender context to support healthier recipe choices [50]. These justifications explain how the nutritional content of the presented recipe relates to the user, as well as help users who have health-related goals to make better decisions.

In our experimental evaluations, we focus on nutrition and health-related explanations for single-list and multi-list interfaces. In Study 1, we design simple explanations that typically optimize for a single nutrient, such as saturated fat. In Study 2, we develop a knowledge-based approach in line with Musto et al. [51], and compare explanations that only describe the nutritional content of recipes to explanations that also relate a recipe's nutrients to a user's food-related goals, such as "Recipes Low in Saturated Fat That Match Your Weight-Loss Goal".

3 STUDY 1

To address [RQ1] and [RQ2], Study 1 examined how a user's evaluation and healthy recipe choice related to each other in the context of a multi-list recommender system. We describe our 2 (single- vs multi-list) x 2 (with or without explanations) between-subject design study, including how our recipes, explanations, and measures were evaluated in the context of the user experience framework [44].

3.1 Methods

3.1.1 Dataset. We developed a food recommender system based on similar-item retrieval. It employed recipes from Allrecipes.com, a popular recipe website to which users can upload their own recipes. From a larger database of around 58,000 recipes (which was also used in [75, 77–79]), we determined five different categories from which we sampled a total of 935 recipes:[1] Casseroles, Roasts, Salads, Pasta, and Chicken dishes. In turn, a subset of 28 recipes was randomly selected from this dataset (5 to 6 per dish type) to serve as 'reference recipes' in our study, on which the different recommendation lists would be based.

3.1.2 Recommendation Approach and Lists. The recommendation approach we implemented was based on the similar-item principle [78]. Hence, given a recipe r_i, we find all top-k most similar

[1] The full list of recipes, including features, can be obtained here: https://osf.io/cpfwj/.

recipes r_j. Formally, this can be expressed as follows:

$$rec@k(r_i) = \underset{r_j \in R \backslash r_i}{\overset{k}{\operatorname{argmax}}}\{sim(r_i, r_j)\}, \tag{1}$$

where $R\backslash r_i$ denotes the set of all recipes without r_i and $sim(r_i, r_j)$ is a similarity function. In our case, similarities were calculated based on recipe titles. Previous studies on similar-item retrieval have found these to be representative of human similarity judgments [78]. Titles were favored over using ingredients, as a recipe's ingredients were not depicted in the interface. Moreover, titles were found to be the most used cue in similar-item judgment tasks [78].

To compute pairwise recipe similarity based on titles, we used Term Frequency-Inverse Document Frequency (TD-IDF). We implemented the test-bed as an PHP online application, using the Zend framework and Apache's Lucene search framework [57]. This framework indexed all recipes in our dataset to allow for similar-item retrieval and recommendation, based on a given reference recipe (see above).

For each list presented in our user study, we randomly selected a reference recipe that matched five predetermined search queries for that trial ('Casserole', 'Roast', 'Salad', 'Pasta', and 'Chicken'). Recommendation sets were populated using similar-item retrieval. To create explainable sub-lists for the multi-list interface, we first retrieved the top-40 recipes in terms of title-based similarity. Subsequently, we applied a post-filtering approach (based on [75]), by re-arranging the retrieved recipes on a specific feature per list and presenting the top-5, i.e., $k = 5$. In total, we designed and displayed five different recommendation lists with the following re-sorting criteria:

- Similar Recipes: Similar recipes sorted from most to least similar (without resorting).
- Fewer Calories: Recipes were re-sorted on their calorie content, from lowest to highest.
- Fewer Carbohydrates: Recipes were re-sorted on their carbohydrate content (per 100g), from lowest to highest.
- Less Fat: Recipes were re-sorted on their fat content (per 100g), from lowest to highest.
- More Fiber: Recipes were re-sorted on their fiber content (per 100g), from highest to lowest.

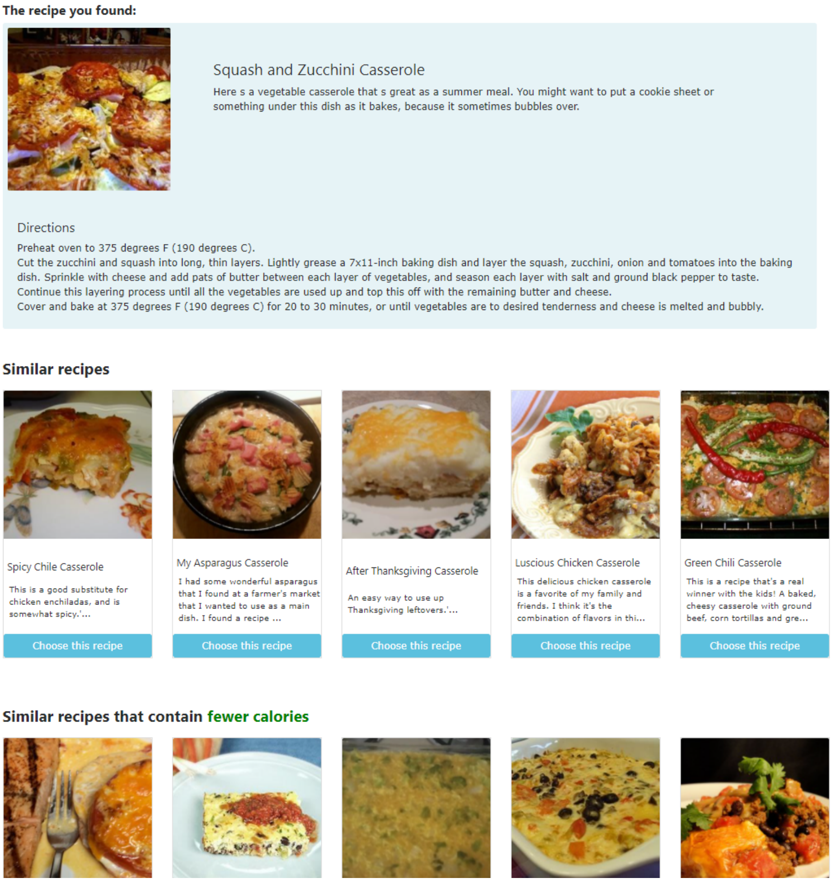
If an explanation was shown above such a list, it would either state "Similar Recipes" or "Similar, but with [fewer/less/more] [calories/carbs/fat/fiber]". Hence, explanations would emphasize similarity, while indicating nutrition-specific characteristic of the list if applicable.

3.1.3 Participants. A total of 366 participants ($M_{age} = 34.24$ years, $SD = 13.23$; 52% male) completed our user study. Among them, 182 participants with no dietary restrictions were recruited from the crowdsourcing platform Prolific, who were compensated with 1.25 USD. The 184 other participants were recruited from Amazon Mechanical Turk, who had completed more than 500 HITs and were compensated with 0.75 USD.[2] On average, participants took 321 seconds or around 3.5 minutes to finish the study.

3.1.4 Procedure. Participants were invited to join a study in which they could find interesting recipes to cook, including suggestions for healthier choices. After disclosing demographics and their self-reported healthiness and cooking experience, users were instructed to imagine that they had used five different search terms to look for recipes: 'Casserole', 'Roast', 'Salad', 'Pasta', and 'Chicken'.[3] Subsequently, they were presented five trials in our recommender interface, of which an example is depicted in Figure 2. In each trial, users were presented a 'reference recipe' at the top of the screen that matched one of the five search queries. Underneath it, a recommendation set was presented that contained recipes that were similar to the reference recipe at the top, either

[2] The research conformed to the ethical standards of the Norwegian Centre for Research Data (NSD).

[3] The order in which recipes were presented was counterbalanced to mitigate order effects.



The recipe you found:

Squash and Zucchini Casserole

Here s a vegetable casserole that s great as a summer meal. You might want to put a cookie sheet or something under this dish as it bakes, because it sometimes bubbles over.

Directions

Preheat oven to 375 degrees F (190 degrees C).
Cut the zucchini and squash into long, thin layers. Lightly grease a 7x11-inch baking dish and layer the squash, zucchini, onion and tomatoes into the baking dish. Sprinkle with cheese and add pats of butter between each layer of vegetables, and season each layer with salt and ground black pepper to taste.
Continue this layering process until all the vegetables are used up and top this off with the remaining butter and cheese.
Cover and bake at 375 degrees F (190 degrees C) for 20 to 30 minutes, or until vegetables are to desired tenderness and cheese is melted and bubbly.

Similar recipes

Spicy Chile Casserole

This is a good substitute for chicken enchiladas, and is somewhat spicy.'...

My Asparagus Casserole

I had some wonderful asparagus that I found at a farmer's market that I wanted to use as a main dish. I found a recipe ...

After Thanksgiving Casserole

An easy way to use up Thanksgiving leftovers.'...

Luscious Chicken Casserole

This delicious chicken casserole is a favorite of my family and friends. I think it's the combination of flavors in thi...

Green Chili Casserole

This is a recipe that's a real winner with the kids! A baked, cheesy casserole with ground beef, corn tortillas and gre...

Similar recipes that contain fewer calories

Fig. 2. Partial screenshot of our recommender interface. Depicted at the top is the reference recipe, for which a similar-item recommendation set is retrieved. Depicted here is the multi-list condition with explanations, presenting multiple lists simultaneously.

presented in a single list or across multiple lists (cf. Figure 2). For each trial, users were asked to choose the recipe they liked most and would like to prepare at home. In addition, they were asked to evaluate how much they liked the chosen recipes and the presented recommendations. After going through five trials, users were then asked to evaluate the recipe sets recommended to them, in terms of their experienced choice difficulty, and perceived diversity and understandability.

3.1.5 *Research Design.*

The recommender interface's list design and the inclusion of explanations were subject to a 2x2-between user design. On the one hand, users were presented either a single list of five recipes or a multi-list interface that comprised five lists of five recipes each (25 recipes in total) across all trials. In the single-list condition, all of the mentioned lists in Subsection 3.1.2 would be presented across all five trials, such as a list re-ranked on fewest carbohydrates, presenting one per trial in a randomized order.

On the other hand, users were randomly assigned to either the 'no explanation' baseline or the explanation condition. For the 'no explanation' baseline, recommended recipes were only annotated with the overall explanation 'Similar Recipes', which would be presented at the top of the interface. In contrast, users assigned to the explanation condition were presented list-specific explanations.

This meant for the single-list condition that one specific explanation was presented per trial above each list of five recipes, such as 'Similar recipes that contain fewer calories'. As the presented lists differed per trial, the explanations would also do so, mentioning either similarity, calories, carbohydrates, fat or fiber depending on the list content. In the multi-list condition, each trial or interface comprised five explanations, explaining the contents of each of the sub-lists in the interface. The vertical order of lists and accompanying explanations was also randomized.

3.1.6 Measures. To relate changes in the interface design to interaction data and how users evaluated an interface, we examined different types of measures.

User Evaluation Metrics. To examine whether a multi-list interface was evaluated more favorably than a single-list interface (RQ1), we asked users to reflect on their chosen recipes, the presented recommendations, and the overall interface. Per list of recommended recipes, we asked users whether they liked the recipes they've chosen (i.e., Choice Satisfaction; items adapted from [65, 69]). At the end of each study, we inquired on their perceived choice difficulty (items adapted from [37]), their perception of the diversity among presented recipes (items adapted from [44, 69]), and how understandable each list was. All items, listed in Table 1, were evaluated through 5-point Likert scales.

Choice Metrics & User Characteristics. To examine possible changes in user choices (RQ2), we represented the healthiness of each recipe through its 'FSA score'. This score, ranging from 4 (healthiest) to 12 (unhealthiest), was based on nutritional guidelines of the UK Food Standards Agency [53] and was used in earlier studies [56, 66, 70, 77]. In short, a recipe's FSA score was higher if the fat, saturated fat, sugar, or salt content was higher per 100g (cf. [70] for computational details). Since there were slight variations in the average FSA score across conditions, we considered the FSA score of chosen recipes relative to the mean of the recipes presented (i.e., the FSA score of the chosen recipe minus the mean FSA score of the presented recipes).

To relate users' choices to their eating goals, we considered from which list a recipe was chosen. In addition, we asked users whether they had one or more specific goals when choosing a recipe. They could indicate to look for similar recipes, recipes they liked, recipes with more fiber, or recipes with lower fat and kcal. In our analysis, we tallied the number of lists for which the chosen recipes matched a user's recipe or eating goal. For example, a choice was counted as a match if a user had indicated to look for recipes with more fiber and chose a recipe from the 'More Fiber' sub-list. Finally, we asked users to rate their self-reported health and cooking experience, which were captured on 5-point scales, as well as to disclose some demographical details, such as age and gender.

3.2 Results

3.2.1 Confirmatory Factor Analysis. We compared a user's evaluation of our single and multi-list interfaces in Study 1 through the recommender system user experience framework [44]. We submitted the responses to our questionnaires to a confirmatory factor analysis (CFA) using ordinal dependent variables. Table 1 shows that we could reliably distinguish between four different aspects: Choice Difficulty, Perceived Diversity, Understandability, and Choice Satisfaction. Items that did not explain sufficient variance of their respective latent aspects were removed from further analysis. Eventually, the resulting aspects all met the guidelines for convergence validity, as the average variance explained of each aspect was larger than 0.5 [43].[4]

[4] Although some SEM guidelines recommended to use at least three items per latent aspect for small SEM analyses (e.g., [43]), Kline [40] describes that the use of two items per latent aspect is sufficient, as long as the model's degrees of freedom are sufficiently high; which was the case here.

Table 1. Results of the confirmatory factory analysis on user experience aspects for Study 1. The analysis was clustered at the user level, as the items for choice satisfaction had five observations per user. All aspects met the requirements for convergent validity ($AVE > 0.5$). Items in grey and without factor loading were omitted from the final Structural Equation Model.

Aspect	**Item**	**Loading**
Choice Difficulty	I changed my mind several times before choosing a recipe.	.755
$AVE = .53$	I think I selected the most attractive recipe from each list.	
$\alpha = .71$	I was in doubt between multiple recipes.	.769
	The task of choosing a recipe was overwhelming.	.548
Perceived Diversity	The lists of recommended recipes were varied.	.689
$AVE = .58$	The recommendation lists included recipes from many different categories.	.655
$\alpha = .69$	Several recipes in each list differed strongly from each other.	
	Most recipes were of the same type.	
Understandability	I understood why recipes were recommended to me.	.825
$AVE = .61$	The explanations of recipes, such as ‘similar recipes’, were clear to me.	.652
$\alpha = .67$	I did not understand the presented explanations.	
Choice Satisfaction	I like the recipe I’ve chosen.	.804
$AVE = .72$	I think I will prepare the recipe I’ve chosen.	.751
$\alpha = .85$	I like the list of recommended similar recipes.	.610

3.2.2 Structural Equation Modeling. We organized the objective constructs, subjective constructs, and relevant interactions into a path model using Structural Equation Modeling (SEM). As suggested by Knijnenburg and Willemsen [43], we first tested a fully saturated model and performed stepwise removal of non-significant relations afterwards. Figure 3 depicts the resulting model, which had good fit statistics: $\chi^2(100) = 177.130, p < 0.001, CFI = 0.982, TLI = 0.977, RMSEA = 0.021$, $90\%-CI$: $[0.015, 0.025]$. Our path model met the guidelines for discriminant validity, as the correlations between latent constructs were smaller than the square root of each their AVEs (cf. Table 1) [43].[5]

User Experience of Multi-List vs Single-List Interfaces (RQ1). Figure 3 depicts two types of ‘main’ paths between the objective changes in our interfaces (i.e., multi-list vs single list, use of explanations) towards our evaluation aspects (i.e., choice difficulty, choice satisfaction). The first path, running at the top of Figure 3, showed that multi-list interfaces (with and without explanations) led to higher levels of perceived diversity ($\beta = .780$, $p < 0.001$). This indicated that presenting more recipes from multiple lists to users led them to perceive a list as being more varied.

In turn, diversity affected two user experience aspects. First, higher levels of diversity came at the cost of higher levels of choice difficulty: $\beta = .249$, $p < 0.001$. A test of indirect effects showed that the path from multi-list to choice difficulty was mediated by diversity ($coef. = .194$, $p < 0.01$). This effect is also depicted in Figure 4: choice difficulty was significantly higher in the multi-list condition (compared to single lists), while no interaction effect of explanations could be observed. Second, diversity was also positively related to choice satisfaction: $\beta = .362$, $p < 0.001$, leaving

[5] This model was inferred using all users. We also tested a model from which we excluded users who had not passed the attention check, but this did not lead to significant changes in the path model.

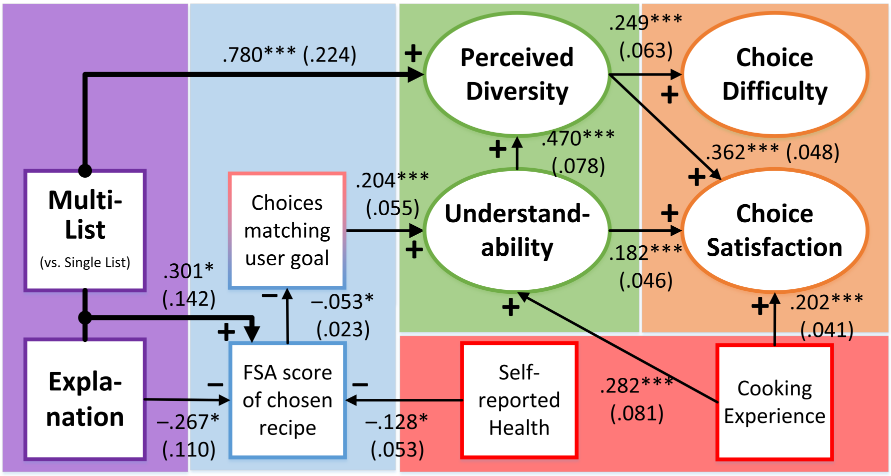


Fig. 3. Structural Equation Model (SEM). Numbers on the arrows represent the β-coefficients, standard errors are denoted between brackets. Effects between the subjective constructs are standardized and can be considered as correlations, other effects show regression coefficients. Aspects are grouped by color: Personal characteristics are red, objective system aspects are purple and behavioral indicators are blue. Experience aspects are orange, perception aspects are green. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

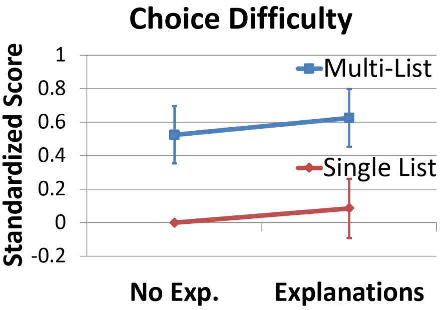


Fig. 4. Standardized scores for the choice difficulty experience aspect across conditions. Errors bars represent 1 S.E.

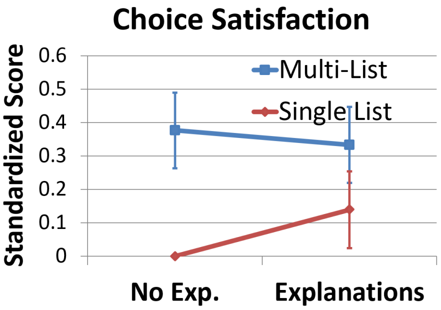


Fig. 5. Standardized scores for the choice satisfaction experience aspect across conditions. Errors bars represent 1 S.E.

users more satisfied with the recipes they had chosen if the recommendation sets were perceived as diverse. The path from multi-list towards choice satisfaction was also significantly mediated by diversity ($coef. = 0.282$, $p < 0.01$), which can be understood by inspecting Figure 5. Whereas choice satisfaction levels were higher for multi-lists, both with and without explanations, we did not observe an interaction effect with the use of explanations.

Choice Metrics (RQ2). The second main path in Figure 3 stemmed from both objective system aspects and followed through choice metrics towards the perception and evaluation aspects. We observed two contrasting effects of our research design on the healthiness of chosen recipes, relative to the mean in a recommendation set: while the addition of explanations led users to choose relatively healthy recipes (i.e., with lower FSA scores): $\beta = -.267$, < 0.05, an interaction effect between multi-list (vs single list) and explanations led to relatively unhealthy choices (i.e., recipes with higher FSA scores): $\beta = .301$, $p < 0.05$. This effect was understood by inspecting Figure 6,

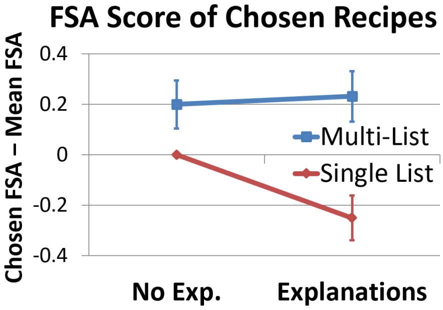


Fig. 6. FSA score of chosen recipes, relative to the mean FSA score of the presented recommendations. Negative values indicate that a relatively healthy recipe was chosen, vice versa for positive values. Errors bars represent 1 S.E.

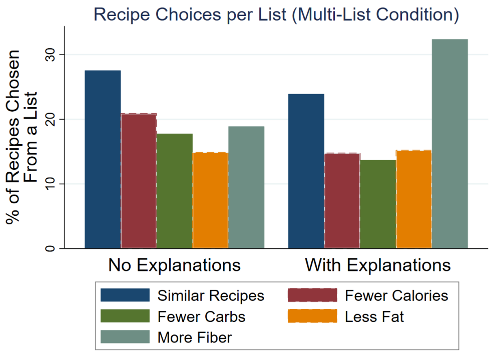


Fig. 7. Distribution of individual lists from which recipes were chosen – for the Multi-List condition only. In the Single List condition, the distribution was flat (0.2 for each list).

which on the one hand depicts that the addition of list-specific explanations (instead of 'Similar Recipes') led to lower, healthier FSA scores of the chosen recipe in the single-list conditions. On the other hand, it shows that recipe choices in the multi-list conditions were unhealthier than in the single-list conditions, which was not further affected by the use of explanations.

Furthermore, Figure 3 shows that the chosen FSA score was negatively related to the number of choices that matched a user's goal: $\beta = -.053$, $p < 0.05$. This meant that users who had chosen relatively healthy recipes were also more likely to have chosen recipes that matched their recipe goals. The distribution of lists from which recipe were chosen, per multi-list condition[6], is depicted in Figure 7. The presence of explanations led users to choose fewer 'Similar' and low-calorie recipes, but more recipes that were rich in fiber. Since the FSA score of most fiber-rich recipes was higher ($M_{fiber} = 7.94$) than the recommendation set's average ($M_{Multi-List} = 7.44$), it seemed that users had chosen healthier recipes from other lists. Although there were little changes in the relative chosen FSA score for 'Similar', 'Fewer Calories' and 'More Fiber' recipes, the addition of explanations led to lower FSA scores for the 'Fewer Carbs' list (a drop from +.50 to +.15) and the 'Less Fat' list (going down from +.21 to -.42).

The last part of the path in Figure 3 shows that more choices that matched a user's goal led to a higher level of understandability: $\beta = .204$, $p < 0.001$. This, in turn, was positively related to higher levels of choice satisfaction through two pathways: one direct path and one mediated by perceived diversity. However, a test of indirect effects showed that the total path from the objective system aspects towards choice satisfaction was not significantly mediated by the aforementioned interaction metrics, indicating that they were related but not mediated causally.

User Characteristics. Finally, two user characteristics (in red) also significantly affected a user's choices, perception, and evaluation.[7] First, a user's self-reported health was negatively related to the FSA score of chosen recipes ($\beta = -.128$, $p < 0.05$), showing that users who rated themselves as healthy had also chosen healthier recipes. Second, a user's cooking experience was positively related to the perceived understandability ($\beta = .282$, $p < 0.001$) and experienced choice satisfaction

[6] The distribution of lists from which recipes were chosen in the single-list condition is flat, as each list was presented once to each user/ Therefore, we only depict the findings for the multi-list conditions.
[7] We had also explored possible interaction effects between user characteristics and interaction metrics and evaluation aspects, but found none.

($\beta = .202, p < 0.001$). This suggested that our recommender interface, averaged across all conditions, was more suitable for experienced users than novices. A test of indirect effects indicated that this path was significantly mediated by both understandability and perceived diversity, indicating that experienced users better understood our interfaces and, in turn, perceived them as more diverse and were more satisfied with the recipes they had chosen.

3.3 Summary and Conclusion

We examined to what extent multi-list interfaces were evaluated more favorably than single-list interfaces (RQ1), as well as whether they could support users with healthy eating goals (RQ2). We did so by comparing single-list interfaces with 5 items (with or without explanations) to multi-list interfaces of 25 items (with or without explanations).

With regard to [RQ1], we found that users were more satisfied with recipes they have chosen from a multi-list interface, compared to a single interface. They also reported higher levels of perceived diversity, which in turn positive affected choice satisfaction. At the same time, users experienced higher levels of choice difficulty when using a multi-list interface, compared to a shorter list that did not trigger choice overload. These findings were in line with preceding research [4], except that our path model showed that the detrimental effects of the larger set size were fully mediated by diversity.

With regard to [RQ2], we observed that on average users made unhealthier decisions in multi-list interfaces. Although an interaction effect was revealed between the use of a multi-list interface and explanations, this only led to relatively healthier choices in the single-list conditions. However, our path model did reveal that users who had chosen relatively healthy recipes, typically chose them from lists that matched one or more healthy eating goals. In turn, this was also related to higher levels of understandability, suggesting that the multi-list interface was able to support users with specific nutrition-related goals. Moreover, the unhealthiness of recipe choices was exacerbated by the poor FSA scores of the 'more fiber' recipes.

The main limitation of Study 1 was that our comparison of single-list and multi-list interfaces was confounded on differences in set size (5 vs 25). To control for this, we performed Study 2, in which we compared single-list and multi-list interfaces with 25 recipes each.

4 STUDY 2

We developed a knowledge-based food recommender system[8] for Study 2. We compared single-list and multi-list interfaces with the same set size, as well as examined the merits of personalized explanations compared to non-personalized, recipe-focused explanations, akin to the ones used in Study 1. This was assessed in terms of the user's evaluation (RQ1) and the healthiness of recipes chosen (RQ2).

4.1 Methods

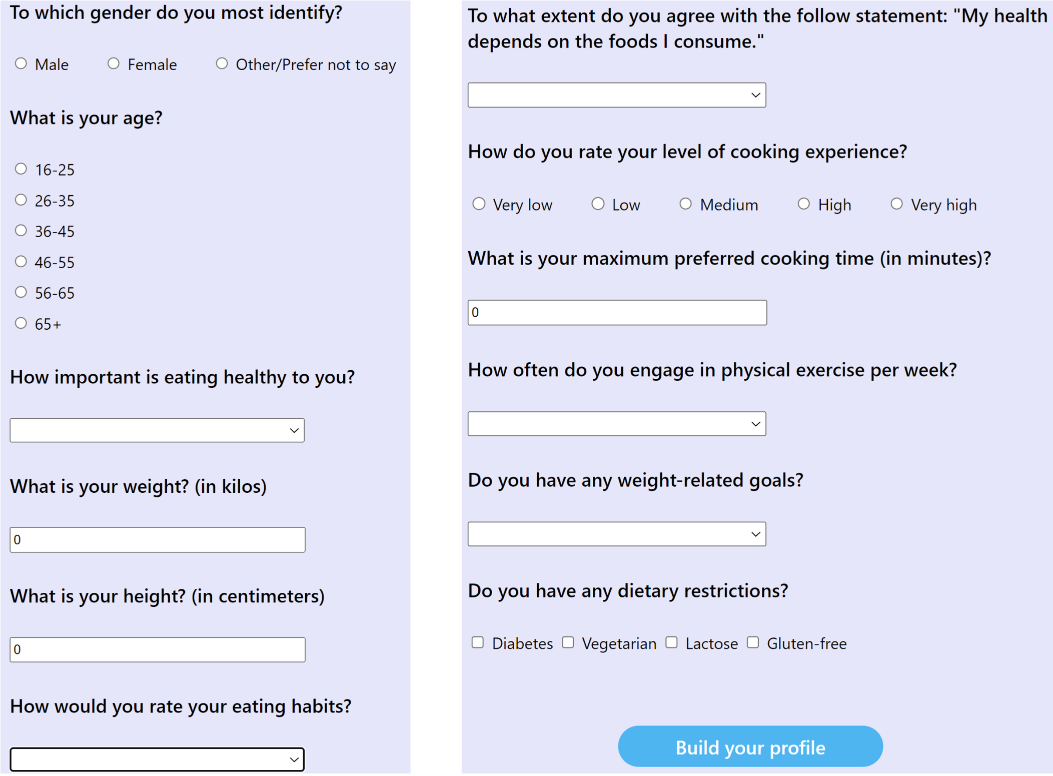
4.1.1 Dataset. We used a database from the Italian food community platform GialloZafferano. It contained 4,671 recipes, used in previous studies on food recommendations [50, 51], which were each translated to English. For this study, we only used main course recipes, which eventually resulted in a dataset of 1,190 recipes. The dataset included descriptive details, such as the recipe's name, images, cooking time, and allergents. In terms of nutritional content, we could access the weights (in g) per serving for carbohydrates, fat, saturated fat, sugar, fiber, and sodium, as well as the amount of calories. However, we could not compute the nutritional content per 100g, as the serving size was missing.

[8] Please refer to our prototype: https://github.com/larsholth97/personalized-exp-multilist.

4.1.2 Participants. A total of 164 participants (75% below 35 years old; 55.2% female) completed our user study. Participants were recruited online through the crowdsourcing platform Prolific and were compensated with 0.75 GBP, as we estimated that completing the study would take 5 to 6 minutes[9]. Participants needed to have at least a 95% approval rate and to be fluent in English. Eventually, participants took on average 500 seconds (just over 7 minutes) to complete the study (SD = 293 seconds).

4.1.3 Procedure. Participants were first asked to disclose personal information and preferences. This included demographic details, questions about their weight status, and information about their cooking experience. A screenshot of the questions posed to users is depicted in Figure 8. Thereafter, they would be presented two personalized sets of 25 recipe recommendations, which is partially depicted in Figure 9. Each set would include one or more explanations and a few features of a recipe: name, photo, cooking time, and a short description. For each recommendation set, they were asked to choose one recipe they liked the most and would like to prepare at home in the near future. Moreover, they were asked to evaluate their perception of the recipes, the interface and how they experienced choosing a recipe.

[9] The research conformed to the ethical standards of the Norwegian Centre for Research Data (NSD).



To which gender do you most identify?

○ Male ○ Female ○ Other/Prefer not to say

What is your age?

○ 16-25
○ 26-35
○ 36-45
○ 46-55
○ 56-65
○ 65+

How important is eating healthy to you?

What is your weight? (in kilos)

0

What is your height? (in centimeters)

0

How would you rate your eating habits?

To what extent do you agree with the follow statement: "My health depends on the foods I consume."

How do you rate your level of cooking experience?

○ Very low ○ Low ○ Medium ○ High ○ Very high

What is your maximum preferred cooking time (in minutes)?

0

How often do you engage in physical exercise per week?

Do you have any weight-related goals?

Do you have any dietary restrictions?

□ Diabetes □ Vegetarian □ Lactose □ Gluten-free

Fig. 8. Profile builder for the knowledge-based recommender system used in Study 2. Depicted here are two screenshots from the same page in which user preferences were elicited. Details on the options displayed in drop-down menus are discussed in the subsection 4.1.6.

Table 2. Knowledge-based relations between elicited user characteristics and recipe features. Recipes are assigned a higher score if they meet specific thresholds. For details about scoring, please refer to our GitHub repository: https://github.com/larsholth97/personalized-exp-multilist.

User Characteristics	Recipe Features	Scoring Mechanism
Cooking Experience	Preparation Difficulty	Distance Metric
Cooking Time	Preparation Time	Cooking Time>Preparation Time: ↑
Healthiness of Eating Habits	Calories & Nutrients	Higher Calorie & Nutrient allowance for scoring ↑ if Eating Habits are healthier
Physical Activity & BMI	Calories & Nutrients	Low BMI + Physical Activity: Protein ↑; High BMI: Low-Calorie and low-Fat ↑
Weight-Gain Goal	Calories & Nutrients	Calories<550, Carbs<23g, Fat<15g, Sat.Fat & Sugar<8g: ↑
Weight-Loss Goal	Calories & Nutrients: ↑	Calories>550, Fat>15g: ↑

4.1.4 Recommendation Approach and Explanations. We employed a knowledge-based recommendation approach to generate recipe recommendations. Our approach leveraged elicited user characteristics on the one hand, and recipe features on the other hand.

The recommendation pipeline comprised three steps. To support all steps, user characteristics were employed as preferences and were used to score different recipe features. The knowledge-based relations between users and recipes are described in Table 2, which were used to score recipes and select possible lists. Details on scoring can be retrieved from our GitHub repository[10].

As step 1, list algorithm candidates, for single and multi-list alike, would be selected from the sample of lists that were compatible with a user's preferences. Instead of generating list explanations using NLP methods as in a previous knowledge-based recommender study [51], we pre-designed list explanations based on general dietary guidelines and common sense matching. For example, users who wished to lose weight could be presented a list with 'low-fat' recipes, while users who indicated to be novice cooks would *not* be exposed to algorithms that selected difficult recipes.

Second, we would filter out recipes that did not meet specific criteria. Most notably, users could disclose dietary restrictions, such as gluten-free, for which recipes could be filtered, such as recipes that contain wheat. This also included constraints that applied to multiple recipe features, such as users indicating to suffer from diabetes. For the latter, we would only present low-fat, low-sugar, and low-sodium recipes, with specific cut-off values.

Third, the candidate lists would be filled with recipes. To do this, recipes in the dataset were scored based on the elicited user preferences and the features of the recipes. Scoring was done in line with the knowledge-based relations described in Table 2. We used a 'positive scoring' approach, in which the baseline score of a recipe would be zero, but was increased if applicable. For each relevant user characteristic, recipe scores were increased if they would meet specific feature values. For example, if a user indicated to be willing to lose weight, a recipe's score was increased if the recipe's fat or caloric content would fall below the British recommended dietary intake guidelines. Eventually, a random sample from the candidate lists would be presented to the user, i.e., the top-1 in the single-list condition and the top-5 in the multi-list condition, in a randomized order from top to bottom.

[10] https://github.com/larsholth97/personalized-exp-multilist

Table 3. Explanations and lists presented in the food recommender interface of Study 2. In single-list interfaces, only one of the lists would be presented, while five lists were presented in multi-list interfaces.

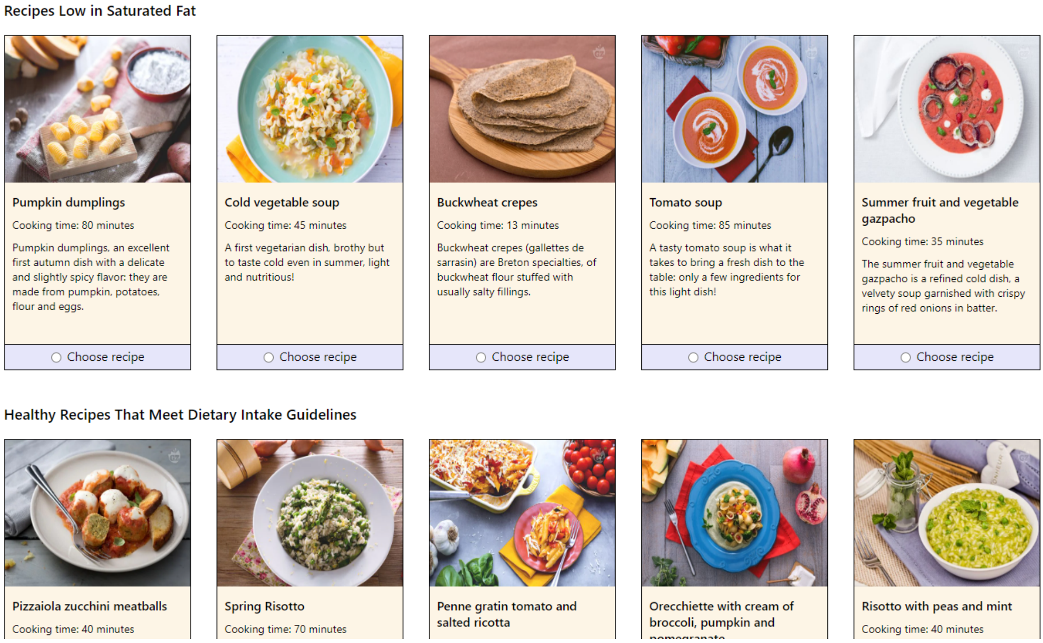
Non-personalized Explanations	**Personalized Explanations**
Healthy Recipes That Meet Dietary Intake Guidelines	Healthy Recipes That Could Improve Your Unhealthy Eating Habits
	Healthy Recipes That Are in Line With Your Healthy Eating Habits
	Healthy Recipes That Also Match Your Vegetarian Dietary Preferences
	Healthy Recipes That Fit Your Lactose-Free Dietary Restriction
	Healthy Recipes That Fit Your Gluten-Free Dietary Restriction
Recipes With a Short Cooking Time	Recipes With a Rather Short Cooking Time, Which Matches Your Preferences
Recipes With a Long Cooking Time	Recipes With a Long Cooking Time, but No Longer Than Your Preferred Cooking Time
	Recipes With a Long Cooking Time That Match Your Preferred Cooking Time and Your Diabetes Dietary Restriction
Recipes That Are Easy to Cook	These Recipes Are Very Easy to Prepare, Which Matches Your Low Level of Cooking Experience
	These Recipes Are Easy to Prepare, Which Matches Your Low Level of Cooking Experience
Challenging Recipes to Try	These Recipes Are Rather Challenging to Prepare, but Match Your Level of Cooking Experience
High-Protein Recipe	High-Protein Recipes That Match Your Body Mass Index and High Level of Physical Activity
	High-Protein Recipes That Match Your Weight-Gain Goal
High-Fat Recipes	High-Fat Recipes That Match Your Body Mass Index and High Level of Physical Activity
	High-Fat Recipes That Match Your Weight-Gain Goal
Low-Fat Recipes	Low-Fat Recipes That Match Your Body Mass Index and High Level of Physical Activity
High-Fiber Recipes	High-Fiber Recipes That Match Your Body Mass Index and High Level of Physical Activity
Low-Sugar Recipes	Low-Sugar Recipes That Match Your Body Mass Index and High Level of Physical Activity
Recipes Low in Saturated Fat	Recipes Low in Saturated Fat That Match Your Weight-Loss Goal
Low-Calorie Recipes	Low-Calorie Recipes That Match Your Body Mass Index and High Level of Physical Activity
	Low-Calorie Recipes That Also Fit Your Diabetes Dietary Restriction
	Low-Calorie Recipes That Also Fit Your Diabetes Dietary Restriction
	High-Calorie Recipes That Match Your Weight-Gain Goal

An overview of all the possible lists and explanations are presented in Table 3. The recipes eligible for the personalized lists would not only be personalized to the user's profile, but would also meet constraints applicable to that list. For example, 'High-Fat Recipes That Match Your Weight-Gain Goal' only considered recipes with high levels of fat, showing the recipes with the highest overall score. Furthermore, even though the explanations in the non-personalized condition did not relate to user characteristics, the content in those lists was still optimized towards the user's elicited preferences. Hence, the manipulation mainly focused on how the lists were explained. Note that duplicates across multiple lists within a single interface were possible, but not across both multi-list interfaces.

4.1.5 Research Design. Recommendations presented to users were subject to a 2x2-mixed research design. All users were presented two sets of 25 recommendations, of which one set presented non-personalized explanations and the other presented personalized explanations. In contrast, only half of users was presented recipes in two single-list interfaces, while the other half faced two multi-list interfaces.

4.1.6 Measures. To assess how changes in the interface design led to alterations in choice behavior and a user's evaluation, we examined different types of measures.

User Evaluation Metrics. To examine whether multi-list interfaces (with or without personalized explanations) were evaluated more favorably than single-list interfaces (RQ1), we inquired on different user evaluation aspects. Table 4 outlines the five user evaluation aspects examined in this study that were related to choice overload and the effectiveness of the used explanations, which were all evaluated on 7-point Likert scales. Questionnaire items used were in part based on previous studies: choice difficulty [37], choice satisfaction [44], perceived support [29, 65], understandability



Recipes Low in Saturated Fat

Pumpkin dumplings
Cooking time: 80 minutes
Pumpkin dumplings, an excellent first autumn dish with a delicate and slightly spicy flavor: they are made from pumpkin, potatoes, flour and eggs.

Cold vegetable soup
Cooking time: 45 minutes
A first vegetarian dish, brothy but to taste cold even in summer, light and nutritious!

Buckwheat crepes
Cooking time: 13 minutes
Buckwheat crepes (gallettes de sarrasin) are Breton specialties, of buckwheat flour stuffed with usually salty fillings.

Tomato soup
Cooking time: 85 minutes
A tasty tomato soup is what it takes to bring a fresh dish to the table: only a few ingredients for this light dish!

Summer fruit and vegetable gazpacho
Cooking time: 35 minutes
The summer fruit and vegetable gazpacho is a refined cold dish, a velvety soup garnished with crispy rings of red onions in batter.

Healthy Recipes That Meet Dietary Intake Guidelines

Pizzaiola zucchini meatballs
Cooking time: 40 minutes

Spring Risotto
Cooking time: 70 minutes

Penne gratin tomato and salted ricotta

Orecchiette with cream of broccoli, pumpkin and pomegranate

Risotto with peas and mint
Cooking time: 40 minutes

Fig. 9. Partial depiction of the recommender interface used in Study 2. Depicted here is the multi-list interface condition with non-personalized explanations.

[42], and perceived diversity [37, 44, 65]. We asked users about their perceived diversity after both sets had been presented, while we inquired on the other aspects after each set.

We performed a principal component factor analysis with non-orthogonal, promax rotation on the different aspects. Unlike in Study 1, we did not perform a confirmatory factor analysis, as we could not eventually infer a reliable structural equation model due to fit issues related to within-user and between-user parameters and divergence validity (cf. [43, 44]). Instead, we analyzed the influence of a multi-list representation and personalized explanations per user evaluation aspect separately. The principal component factor analysis revealed that perceived support did not form a reliable latent aspect and was omitted from further analysis. The internal consistencies of choice satisfaction and understandability were good (α>0.8), while those of choice difficulty and perceived diversity were acceptable (α>0.6).

Choice Metrics. We assessed which recipe was chosen from each list. Unlike in Study 1, we could not express the healthiness of chosen recipes through the FSA score, as we did not have information on a recipe's serving size. Instead, recipe healthiness was expressed through the 'WHO Score', following a method proposed by Howard et al. [28], validated further by Trattner et al. [75], that was based on recommended daily intake levels for six nutrients and calories [54]. The score

Table 4. Results of two principal component factor analyses on user experience aspects in Study 2. As perceived diversity did not have multiple observations per users, it was analyzed in a separate analysis. Factor loadings were obtained after performing promax rotation. Aspects in grey were omitted from further analysis as the items did not form a sensible factor and yielded low values for Cronbach's Alpha. Items without factor loading were omitted due to low factor loadings.

Aspect	Item	Loading
Choice Difficulty	The task of choosing a recipe was overwhelming.	.852
	I changed my mind several times before choosing a recipe.	.801
α = .66	Comparing the recommended recipes was easy.	-.624
Choice Satisfaction	I like the recipe I've chosen.	.789
	I think I will prepare the recipe I've chosen.	.881
α = .83	I know many recipes that I like more than the one I have chosen.	
	I would recommend the chosen recipe to others.	.933
Perceived Support	I could easily find recipes on this page.	
	This page helped to discover new recipes.	
	A page like this helps me make better recipe choices.	
Understandability	I understood why the recipes were recommended to me.	.893
	I could understand how the recipes were based on my preferences.	.896
α = .85	The recommendation process was NOT clear to me.	-.852
Perceived Diversity	Several recipes in each list of recommended recipes differed strongly from each other.	.875
	The recommendation lists included recipes from many different categories.	.875
α = .69	Both interfaces contained recipes that were similar to each other.	
	No two recipes seemed alike.	

runs from 0 to 7, tallying the calories and nutrients for which the intake guidelines were met, and was evaluated as an absolute, continuous score. Besides healthiness, we also tracked from which lists recipes were chosen, as well as the position in the list. The latter was expressed as the vertical 'List Position', with the lowest values indicating items presented in the top row.

User Characteristics. We inquired on various user characteristics. Besides demographics and BMI, we asked users how important they found healthy eating (on a 5-point scale), to what extent their eating habits were healthy (5-point scale), their health consciousness (1 item, 5-point Likert scale), and their level of cooking experience (5-point scale). Users could also indicate their maximum preferred cooking time, to what extent they engaged in physical activity (3-point scale), and whether they had a weight-gain or weight-loss goal. Among these characteristics, cooking experience and health consciousness were considered in our analyses as continuous predictors. Finally, we also inquired on dietary restrictions, which were used to filter recipes from the recommendation sets and to impose nutrient-based constraints (e.g., for diabetes).

Descriptive Statistics. We report the descriptive statistics of the elicited user preferences here; see Figure 8. On a scale ranging from very unimportant to very important, 53.3% of recruited participants ($N = 163$) across both conditions evaluated the importance of eating healthily as important. In contrast, 49.7 percent of respondents self-assessed their own eating habits as neither healthy nor unhealthy, which was also the middle option. Moreover, given the statement, "My health depends on the foods I consume", 61.3% of respondents agree on a scale ranging from entirely disagree to completely agree. In terms of physical activity and diet goals, 52.1% of respondents reported 3 hours of weekly physical activity, while 40.5% reported 6 hours, which was considered as an average level of physical activity in the study. In addition, slightly more than half of the participants (50.9%) had a diet goal of weight loss, whereas 36.8% had no diet goal. In terms of personal preferences related to cooking, 39.2% of respondents rated their experience as medium, while 37.4% rated it as high. The average cooking time requested by participants was 53.37 minutes.

Participants could tick a box if they had any dietary limitations. Although we had applied a filter in Prolific during data collection to rule out dietary constraints, 18.4% percent of participants indicated to have dietary restriction, with lactose-free diet being the most ticked option. The latter may have been due to such specific dietary restrictions not being featured in Prolific's system.

4.2 Results

We examined the user evaluation (RQ1) and recipe choices (RQ2) across knowledge-based multi-list and single-list recommender interfaces, with or without personalized explanations. Although we could not infer a path model due to fit issues, we approached the method of the user experience recommender framework [44], by breaking down the analysis into multiple parts. We investigated changes in perception aspects (i.e., diversity, understandability) and choice metrics (i.e., WHO score) as a function of the research design, while changes in experience aspects (i.e., choice satisfaction and choice difficulty) were in turn predicted using the research design factors, perception aspects, choice metrics, and user characteristics.

4.2.1 Diversity, Understandability and Recipe Healthiness. We first examined whether our research design affected the perceived diversity and understandability of our interface, as well as the healthiness of chosen recipes. For perceived diversity[11], we performed a two-sample t-test and found higher levels of diversity for multi-list interfaces ($M = .18$, $SD = .85$), than for single-list interfaces ($M = -.18$, $SD = 1.11$): $t(162) = -2.30$, $p < 0.05$. This suggested that using multiple algorithms in a recommender interface led users to perceived the content indeed as more varied.

[11] Note that Perceived Diversity was only measured between users and, thus, cannot be used to examine differences between the non-personalized and personalized explanation conditions.

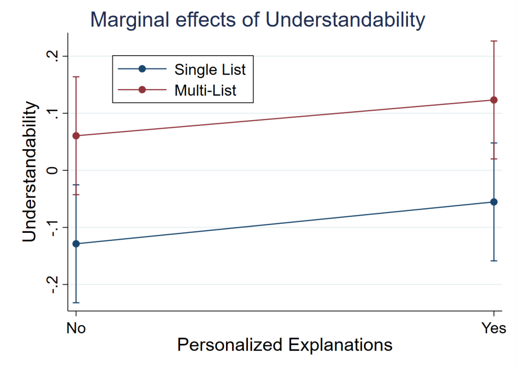


Fig. 10. Marginal effects for the understandability perception aspect across conditions. Errors bars represent 1 S.E.

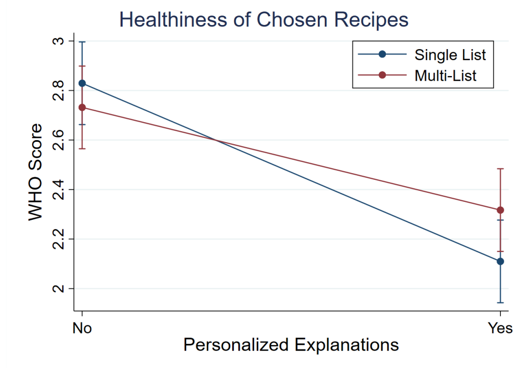


Fig. 11. Marginal effects for the WHO Score across conditions, representing the overall healthiness of chosen recipes. Errors bars represent 1 S.E.

For understandability and the WHO score, we performed two-way mixed-model ANOVAs as a function of the research design, with the personalized explanation condition as a repeated measures term. Our analysis revealed a small increase in understandability for multi-list interfaces ($M = .092$, $SD = 1.03$), compared to single-list interfaces ($M = -.092$, $SD = .96$): $F(1, 162) = 4.92$, $p = .028$). This suggested that presenting multiple list in an interface rather than content from a single algorithm helped to users to understand the recommendation process and why recipes were recommended to them. In contrast, the use of personalized explanations only led to a small, but non-significant increase in understandability ($M = .034$) compared to the non-personalized explanations ($M = -.034$), while no interaction effect was observed. These effects can be understood by examining Figure 10.

The WHO score was significantly affected by our personalized explanation manipulation. The healthiness of the chosen recipes significantly decreased when personalized explanations were presented (i.e., that related to the user; $M = 2.21$, $SD = 1.47$), compared to non-personalized ones ($M = 2.78$, $SD = 1.75$): $F(1, 162) = 10.19$, $p = .0017$. This suggested that users facing interfaces with explanations that also related to their characteristics and healthy eating needs, were less likely to actually choose a healthier recipe than for interfaces with explanations that only described recipe characteristics. In contrast, we did not observe any significant changes in the chosen WHO score across single-list and multi-list interfaces, nor an interaction effect between personalized explanations and multi-list. This can also be observed in Figure 11.

4.2.2 Choice Difficulty and Satisfaction. We further analyzed whether our research design, perception and interaction variables (i.e., perceived diversity and understandability, WHO health score, list position), and personal characteristics (i.e., cooking experience and health consciousness) affected a user's experience choice difficulty and choice satisfaction. We predicted each aspect using a multilevel linear regression model, clustered at the user level.

The results are presented in Table 5. It was revealed that choice difficulty was not reduced due to the use of a multi-list interface, nor because of the presence of personalized explanations or any interaction effect (all predictors: $p > 0.05$). Choice satisfaction did change across single-list and multi-list interfaces, but the direction of the effect was inconsistent with Study 1: it was lower in the multi-list condition than in the single-list condition: $\beta = -0.39$, $p = .001$. Taken together, this

Table 5. Results of multi-level linear regression analyses on choice difficulty and choice satisfaction, clustered at the user level. For interpretability, the Multi-list, Personalized Explanations, and the 'Multi X Personalized' interaction were each coded as continuous variables (taking values of either +0.5 or -0.5), and can be considered as additive effects. Beta coefficients are denoted by β, S.E. denotes the standard error. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

	Choice Difficulty		**Choice Satisfaction**	
	β	S.E.	β	S.E.
Multi-list	-.14	.14	-.39**	.12
Personalized Explanations	.018	.068	.053	.070
Multi X Personalized	-.071	.13	-.14	.14
Perceived Diversity	.026	.072	.20**	.060
Understandability	-.16**	.057	.33***	.052
Choice Difficulty			-.15**	.051
WHO Score	.030	.027	.0077	.026
List Position	-.069*	.029	-.0036	.028
Cooking Experience	.17*	.082	.16*	.069
Health Consciousness	-.17	.10	.19*	.087
Intercept	.20	.48	-1.26**	.41
R^2 (Between)	.084*		.30***	
R^2 (Within)	.033*		.11***	
R^2 (Overall)	.074*		.26***	

suggested that having multiple lists to choose from did not affect choice difficulty, but did decrease the level of satisfaction.

With regard to the perception and evaluation predictors, we found that perceived diversity positively affected choice satisfaction ($p < 0.01$), but did not affect choice difficulty. This suggested that the variety in recipes, which was higher in the multi-list condition, did not made the task of finding a suitable recipe more difficult. It was only partially consistent with our findings in Study 1, as we previously found that diversity positively affected choice difficulty. Our findings for understandability were more in line with Study 1, as we now found that higher levels of understandability were associated with lower levels of choice difficulty ($\beta = -.16$, $p = .006$), as well as with higher levels of choice satisfaction ($\beta = .34$, $p < 0.001$). This suggested that users who understood why recipes were recommended to them found it easier to choose a recipe, as well as were more satisfied with the chosen recipe. This was in spite of understandability not being significantly different across experimental conditions, which was also consistent with Study 1.

We further examined how interaction data affected our evaluation aspects. It was found that the healthiness of chosen recipes (i.e., WHO score) was not related to choice difficulty, nor to choice satisfaction: both $p > 0.05$. In contrast, choice difficulty was related to where a recipe was chosen. Users that chose items listed first or higher up in an interface were more likely to experience higher levels of choice difficulty ($\beta = -.068$, $p = .017$), while this did not affect choice satisfaction. This suggested that top items were not necessarily chosen due to convenience, but because of an increase in decision-making difficulty.

Finally, we found a few user characteristics that affected the overall user experience. Users with higher levels of cooking experience reported both higher levels of choice difficulty ($\beta = .17$, $p = .033$) and choice satisfaction ($\beta = .18$, $p = .010$). This could suggest that experienced users considered more recipes and therefore faced a higher choice difficulty, yet were more satisfied about their chosen recipe. Moreover, while health consciousness was not related to choice difficulty, it was observed to positively affect choice satisfaction ($\beta = .22$, $p = .013$), suggesting that users who felt that their health depended on the foods consumed, seemed to be more satisfied with using our knowledge-based recommender system, compared to users who are less health conscious.

4.3 Summary and Conclusion

We examined changes in user evaluation and recipe choice in Study 2, based on n Study 2: 1) whether using multi-list interfaces rather than single-list interfaces, and 2) presenting detailed personalized explanations instead of non-personalized, recipe-focused explanations led to changes in a user's evaluation and recipe choices. An important aspect of this study, compared to Study 1, was that the set sizes were consistent across conditions, while a knowledge-based recommender engine was used instead of content-based similar item retrieval.

Contrary to our findings in Study 1 for [RQ1], we observed a small, direct negative effect on choice satisfaction for using a multi-list rather a single-list interface. This is not consistent with previous work on category-based interfaces with related evaluative aspects [7, 58], nor with previous work from Jannach et al. [34] that observed no differences in the level of satisfaction across interface conditions. This finding was hard to attribute to the measured user perception and experience aspects. For one, presenting a multi-list did increase the user's perceived diversity and understandability, which were also found to positively affect choice satisfaction. Even though we could not test this in a path model as in Study 1, it seemed that presenting a multi-list interface had a direct negative effect on choice satisfaction, while it also positively affect it through understandability and diversity.

Furthermore, whereas we found in Study 1 that choice difficulty increased in the multi-list condition, mediated by diversity, we observed no effect in this Study. Instead, understandability was negatively related to choice difficulty, suggesting it was easier to choose a recipe if the interface was perceived as understandable – which seemed to apply to the multi-list condition. Moreover, choice difficulty and choice satisfaction were negatively related, revealing another possible 'path' for the positive effect of using a multi-list interface. This showed that presenting a multi-list rather than a single-list interface positively affected a user's evaluation, but mostly in terms of the inquired perception aspects.

For [RQ2], we found that the healthiness of chosen recipes decreased because of the use of personalized explanations. For multi-list interfaces, it was suggested that highlighting how nutritional features of recipes in a list were connected to user characteristics led users to unhealthier options or lists that were not nutrition-focused (e.g., 'These recipes are very easy to prepare, which matches your low level of cooking experience'). Hence, the healthiness of chosen recipes was higher for interfaces with explanations that described recipe features only. We also found such user choices to not relate to the examined user experience aspects, suggesting that a user's perception of the interface and experienced difficulty and satisfaction were not related to recipe health in Study 2.

5 DISCUSSION

In this paper, we have examined an emerging topic in the context of recommender systems. Multi-list interfaces are being used in an increasing number of commercial applications [22]. Nonetheless, studies on how they are evaluated by users typically do not involve a user-centric evaluation [34]. That is, research has yet to examine how changes in a multi-list interface relate to choice data and perception and experience aspects. Moreover, their use is limited to specific domains [22, 34, 58], mostly consumer and leisure domains (e.g., e-commerce, movies), that do not correspond to domains where behavioral change plays a role. In fact, the interplay between multi-list interfaces and user goals, such as healthy eating, has not yet been examined empirically [68].

This research involves one of the first empirical examinations of multi-list interfaces in the food domain. Moreover, it is also the first to have investigated to what extent a multi-list recommender interface is evaluated more favorably than a single-list interface, in the context of the user experience recommender framework of Knijnenburg and Willemsen [43]. In performing such a user-centric evaluation, we have examined whether a multi-list interface can support healthier recipe choices and user food goals (Study 1), by designing nutrient-specific recommendation lists. Whereas other studies are based on single-item evaluations [34] or analyses in which latent aspects are evaluated separately [58, 59], we have linked different latent evaluation aspects in a path model. For Study 2, we could not infer a path model due to fit validity issues, but have instead presented analyses with which we show how interaction data and user perceptions are related to user experience aspects.

5.1 User Evaluation (RQ1)

With regard to [RQ1], both studies reveal for most of the inquired user experience aspects that users evaluate multi-list recommenders more favorably than single-list interfaces. However, there are a few contrasting results between Study 1 and Study 2 regarding a user's choice experience, which may have arised due to a few differences in their respective designs.

In Study 1, we have found that users are more satisfied about the recipes they have chosen from a long multi-list interface, compared to a short single-list interface. Moreover, they also report higher levels of perceived diversity. At the same time, we find that users experience higher levels of choice difficulty when using a multi-list interface, compared to a shorter list that does not trigger choice overload (cf. [4, 62]). These findings are consistent with earlier studies on choice overload [32], in which people evaluate larger choice sets more favorably, but also find it more difficult to make a decision, which can also lead to choice deferral [11]. On top of that, we have found that the addition of explanations to an unlabeled multi-list interface does not reduce this experienced choice overload, nor does it significantly increase choice satisfaction. This partially contrasts with earlier findings that an 'organized view' of multiple item lists reduces the perceived cognitive effort or load [52, 58]. It is possible that the addition of explanations does not have an impact if numerous other modalities are presented in the interface, such as a recipe's title, photo, and description.

The main limitation to Study 1 is the difference in set sizes across single-list and multi-list interfaces. While the set size in the multi-list conditions was 25, comprising 5 algorithms in a single interface, we only showed five recipes in the single-list conditions that stemmed from a single algorithm. In itself, this seems to have also contributed to the higher levels of choice difficulty and choice satisfaction in the multi-list interface conditions, as such a list length effect is also suggested in Bollen et al. [4]. In Study 2, we have aimed to mitigate these limitations by presenting a fairer comparison, examining the user evaluation across different interfaces that each present 25 recipes. There, we no longer find differences in choice difficulty between single-list and multi-list interfaces, while the perceived diversity and understandability is higher for the multi-list conditions.

Contrary to Study 1, we have found choice satisfaction to be lower for our multi-list interfaces in Study 2. On top of that, choice difficulty is found to increase in Study 1, while it is not affected in Study 2 across the single-list and multi-list conditions. This finding is difficult to explain using the evaluation aspects considered in this study. For one, we have observed positive relations between understandability and choice satisfaction, and between diversity and choice satisfaction. Although the recommender approaches differ across both studies (similar-item retrieval vs knowledge-based), they have been consistent within studies, and are unlikely to have led to this contradictory result.

Moreover, a decrease in satisfaction due to users taking more time to find an appropriate recipe, as is shown in Jannach et al. [34], should have also been reflected in higher levels of choice difficulty, but this was found to not differ across conditions. Instead, a possible factor may be that the multi-list interface, which presented at times rather extensive explanations, is more appropriate to use for experienced users, for we have observed positive relations between a user's levels of health consciousness and cooking experience, and choice satisfaction. Earlier research has suggested, albeit in the context of preference elicitation methods, that how users interact with an interface may be moderated by that user's domain knowledge [41]. Future research should reveal whether this also applies to our findings.

While the experience aspects show mixed results, the perception aspects are more consistent across both studies. We find higher levels of diversity for the multi-list interfaces, even when controlling similar set sizes. Moreover, understandability has been found to be higher for multi-list interfaces in both studies, with the clearest results in Study 2. This indicates that multi-list food recommender systems contribute positively to the user's perception of a system, even if the experience may not improve, compared to a single-list approach.

5.2 Recipe Healthiness (RQ2)

With regard to the healthiness of chosen recipes (RQ2), we are faced with mixed findings. What stands out across both studies is that the average healthiness of recipes chosen decreases when using a multi-list interface, which is not further affected by explanations. What could underpin this finding is that the multi-list interface design has empowered users with a lower level of health-related interests to seek out unhealthy foods. Another explanation may be that the used explanations are rather salient about a persuasive intent of the system to steer users towards healthier choices [19], which might have led to reactance among users [15]. Particularly in Study 2, some of the explanations are rather explicit about the relation between nutrition and the user's characteristics, could have also led to negative feelings about the self [55]. It has been argued in the context of personalized nudging that being aware of this can cause resistance among users [82], particularly in the context of health promotion [13, 17]. A third, and more practical explanation is that multi-list interfaces have made it too easy to locate foods that a user likes, and that taste-based preferences tend to trump nutrition-based preferences among most users [49], also given the popularity of unhealthy recipes on the internet [75]. The latter would be consistent with the lack of changes in choice difficulty.

We do not find clear advantages of the use of explanations in multi-list interfaces regarding recipe healthiness. In fact, in Study 2, personalized explanations actually led to unhealthier choices. It could be argued that current study only put forth recommendation sets with up to 25 items. This is much smaller than the number of items presented in multi-list interfaces in the movie domain, where each sub-list comprises 40 items [22]. Such a recommendation set size arguably better lends it itself for a well-explained multi-list interface. In a smaller 'large sets', however, explanations may only increase user trust as in previous studies [58, 72], but might not significantly affect choice-related outcomes.

In Study 1, we have teased apart user choices for specific lists. We have observed a variety of choices from non-similar, nutrition-focused lists, suggesting that different users seek out different types of recipes. Although food choices in our multi-list interface were relatively unhealthier than in single lists, we also found that the number of unhealthy 'similar recipe' choices in the multi-list conditions were significantly reduced due to the use of explanations, as many users had chosen fiber-rich recipes. We argue that the increase in recipe diversity in the multi-list condition enabled users to find the recipes they are looking for. Moreover, we found that healthy recipe choices were associated with users making more choices that match their eating or recipe goals. These findings suggest that the availability of unhealthy foods will lead to relatively unhealthy choices by users who do not have any healthy eating goals, but will support users with healthy eating goals nonetheless. Moreover, one observed shift in user choices was from recipes that were optimized for similarity, to fiber-rich recipes that had a relatively high FSA score. Future studies should attempt to pin this down more precisely, by incorporating explicit user goals in a recommendation approach, possibly through a critiquing approach (cf. [8]).

5.3 Limitations and Future Work

A question that arises from the findings in this paper is to what extent the results are generalizable. Two aspects might limit the ecological validity of the performed studies. First, the use of crowdworkers across two different platforms (MTurk and Prolific) that do not necessarily need to consume the chosen recipes. The use of crowdworkers, particularly from Amazon MTurk, has received scrutiny for at times leading data generation and research results with a relatively high variability, which may depend on extraneous factors (e.g., such as a person's mood [86]). However, it also offers some advantages, such as more diverse than most recruitment pools, a relatively fast recruitment of participants and a relatively higher-quality data compared with panel providers [80]. While we have rewarded participants at a good rate, have ruled out 'speeding' and have selected on approval rate, while not finding any evidence of reduced motivation, we have not tracked whether they have consumed the chosen recipes. The latter would be an important step for future studies to include, in line with for example work from the energy domain [65].

A second limitation to the ecological validity is the relatively small set sizes that have been used for each recommender system. Whereas many contemporary multi-list systems include numerous items per list [22], our systems have only included 5 per list. Although this have led to an arguably less realistic scenario, it is in line with previous scientific work on multi-list studies [34]. Moreover, the 'reduced' interface design has allowed us to specifically address our research questions. Hence, both studies have sacrificed some ecological validity and generalizability to improve internal validity and experimental control.

It could be argued that the use of a recommendation approach that is not user-personalized in Study 1 is a limitation. However, many recipe websites and recommender systems use similar-item recommendation approaches that are much like our study design [78]. Moreover, the findings from our similar-item approach are useful for domains where personalization is harder to apply, such as on platforms where most users do not have an interaction history or user account, such as news and recipe websites that attract many users from general search engines (i.e., Google).

A limitation to Study 2 is that the analyses could not be compiled in a single path model, such as in Study 1. This can mainly be attributed to cross-correlations across multiple questionnaire items that emerged when organizing the user experience aspects in a structural equation model, which were absent in our principal component factor analysis. This led to divergence validity issues and overall model fit issues. Nonetheless, the analyses that are reported instead, in which each interaction, perception or experience aspect is examined separately, does indicate to what extent our research design has affected them. Although this approach has not allowed us to perform model-driven optimization, it has illuminated nonetheless how interaction and perception aspects and user characteristics affect experience aspects. Moreover, linear regression is consistent with how each edge is formed between two nodes in structural equation model.

A possible limitation for both studies is that we have not controlled for image attractiveness. Two recent studies show that users are more likely to choose recipes that are accompanied by attractive photos [16], which can even lead to healthier choices [70]. Due to our controlled research designs, however, we do not expect this to have affected our results in terms of user evaluation aspects and aggregate choice metrics. Nonetheless, by unpacking an image into its underlying attributes (e.g., contrast, colorfulness) [70], image attractiveness can be added as an additional feature to a recipe database and used to further personalize recommendations, as also done in industry applications [22].

In a similar vein, the systems used in either study have not been tested on usability. Since part of the 'appeal' of such multi-list systems is the ease of locating information, this could have been a useful addition. Although the explanation could have been more thoroughly, we emphasize that our interfaces are based on state-of-the-art design. Moreover, other studies have showed high levels of perceived ease-of-use for multi-list systems [67]. Moreover, other studies have provided indirect evidence that such systems have a high usability [22, 34].

Regarding the explanations used in Study 2, a limiting factor is that they have not been pre-tested. The rationale is based on earlier work on knowledge-based recipe recommender systems [50, 51], where natural language processing is used to concatenate user characteristics and recipe feature to justify healthy recipe recommendations. The lists generated in the current study were rule-based, in the sense that they were pre-defined. We acknowledge that it would have been desirable to have pre-validated their understandability and whether they support user goals. Nonetheless, we have found higher levels of perceived understandability for the detailed, personalized explanations, which were less based on earlier work that the feature-based explanations. This seems to suggest that the explanation has been satisfactory, at least on a comparative level. Nonetheless, we wholeheartedly recommend a qualitative study design regarding explanation design in a multi-list context.

Future studies should test our findings in a more naturalistic setting. For one, the number of recipes recommended should not necessarily be limited to 25. Moreover, the role of explanations could be unpacked further in this specific domain, for they can also be considered a cognitively-oriented nudge [5], while the vertical organization of lists can be regarded as a behaviorally-oriented nudge. In this sense, a multi-list recommender interface is a diverse, yet complex decision-making environment, in which both personalization and digital nudges can steer user choices (cf. [36]). The final choices made by users are likely to be a function of both the content and the choice architecture (cf. [38]).

Future research could also examine the problem of recommendations and interface aspects more generally. For example, it would be interesting to tease apart the influence of recommendation algorithms and how items are organized on a user's final choice. The top presented items might simply attract many of the choices for a specific group of users [70], such as those who found the decision-making to be difficult [11], like in our Study 2. Moreover, we propose to also address the challenge of healthy food or recipe recommendation with a more longitudinal perspective, by examining which recipes fall within the user's 'comfort zone' to try in the short term and whether this can improve the healthiness of a user's diet in the long term.

The takeaways may vary for different readers of this paper. While we encourage scholars to study the effects of multi-list recommender interfaces in more detail, also using qualitative methods and in other domains, we recommend practitioners to not necessarily adopt a multi-list homepage without further testing. For practitioners in the food domain, a similar-item retrieval page with explanations at the bottom of a recipe page should be beneficial, taking the results of Study 1 into account. Moreover, understandable explanations that link user characteristics to recipe features seem to be perceived positively.

5.4 Overall Conclusion

This study has presented two empirical studies on multi-list recipe recommender systems. Being among the first to do, we have shown that the extent to which multi-list interfaces have behavioral and evaluative benefits seem to depend on the interface design aspects. Compared to single-list interfaces, we have found that multi-list recommender interfaces have evaluative benefits in terms of their perceived understandability and diversity. The results for choice satisfaction and choice difficulty are mixed and warrant research in a more naturalistic environment, that includes longer lists. Based on our findings, we argue that ‘more lists are not always better’, when it comes to a person’s choice process and choice satisfaction, for a similar list of the same length in a grid format can yield higher levels of choice satisfaction.

Regarding user choices, the main benefits of multi-list interfaces seem to lie in helping users with specific goals to locate relevant content. This findings generalizes to domains that are similar to the food recommender domain. It seems that multi-list food recommenders are suitable for ‘more like this’ scenarios across the board (design of Study 1), but that their usefulness for homepages (design of Study 2) requires further examination.

ACKNOWLEDGMENTS

We thank Lars Giske Holth for his relentless work on the knowledge-based recommender systems. This work was in part supported by the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the Centre for Research-based Innovation scheme, project number 309339, as well as through seed funding of the Protein Transition Investment Theme at Wageningen University & Research. It also received funding from the DARS Research Group at the University of Bergen.

REFERENCES

[1] Yuki M Asano and Gesa Biermann. 2019. Rising adoption and retention of meat-free diets in online recipe data. *Nature Sustainability* 2, 7 (2019), 621–627.

[2] James R. Bettman, Mary Frances Luce, and John W. Payne. 1998. Constructive Consumer Choice Processes. *Journal of Consumer Research* 25, 3 (Dec 1998), 187–217.

[3] Devis Bianchini, Valeria De Antonellis, Nicola De Franceschi, and Michele Melchiori. 2017. PREFer: A prescription-based food recommender system. *Computer Standards & Interfaces* 54 (2017), 64–75.

[4] Dirk Bollen, Bart P Knijnenburg, Martijn C Willemsen, and Mark Graus. 2010. Understanding choice overload in recommender systems. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, New York, NY, USA, 63–70.

[5] Romain Cadario and Pierre Chandon. 2020. Which healthy eating nudges work best? A meta-analysis of field experiments. *Marketing Science* 39, 3 (2020), 465–486.

[6] Jefferson Caldeira, Ricardo S Oliveira, Leandro Marinho, and Christoph Trattner. 2018. Healthy menus recommendation: optimizing the use of the pantry. In *Proceedings of the 3rd International Workshop on Health Recommender Systems Co-Located with ACM RecSys*. CEUR, Aachen, DE, 6 pages.

[7] Li Chen and Pearl Pu. 2010. Eye-tracking study of user behavior in recommender interfaces. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6075 LNCS, June 2010 (2010), 375–380.

[8] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction* 22, 1 (2012), 125–150.

[9] Li Chen and Ho Keung Tsoi. 2011. Users’ decision behavior in recommender interfaces: Impact of layout design. In *RecSys’ 11 Workshop on Human Decision Making in Recommender Systems*.

[10] Alexander Chernev, Ulf Böckenholt, and Joseph Goodman. 2010. Commentary on Scheibehenne, Greifeneder, and Todd choice overload: Is there anything to it? *Journal of Consumer Research* 37, 3 (2010), 426–428.

[11] Alexander Chernev, Ulf Böckenholt, and Joseph Goodman. 2015. Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology* 25, 2 (2015), 333–358.

[12] Paolo Cremonesi, Antonio Donatacci, Franca Garzotto, and Roberto Turrin. 2012. Decision-Making in Recommender Systems: The Role of User’s Goals and Bounded Resources.. In *Decisions@ RecSys*. Citeseer, 1–7.

[13] Michele L Crossley. 2002. Resistance to health promotion: a preliminary comparative investigation of British and Australian students. *Health Education* (2002).

[14] Kristin Diehl and Cait Poynor. 2010. Great expectations?! Assortment size, expectations, and satisfaction. *Journal of marketing research* 47, 2 (2010), 312–322.

[15] James Price Dillard and Lijiang Shen. 2005. On the nature of reactance and its role in persuasive health communication. *Communication monographs* 72, 2 (2005), 144–168.

[16] David Elsweiler, Christoph Trattner, and Morgan Harvey. 2017. Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. ACM, New York, NY, USA, 575–584.

[17] Marieke L Fransen, Edith G Smit, and Peeter WJ Verlegh. 2015. Strategies and motives for resistance to persuasion: An integrative framework. *Frontiers in psychology* 6 (2015), 1201.

[18] Jill Freyne and Shlomo Berkovsky. 2010. Intelligent food planning: personalized recipe recommendation. In *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, New York, NY, USA, 321–324.

[19] Marian Friestad and Peter Wright. 1994. The persuasion knowledge model: How people cope with persuasion attempts. *Journal of consumer research* 21, 1 (1994), 1–31.

[20] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, ACM, New York, NY, USA, 257–260.

[21] Mouzhi Ge, Francesco Ricci, and David Massimo. 2015. Health-aware food recommender system. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 333–334.

[22] Carlos A Gomez-Uribe and Neil Hunt. 2015. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)* 6, 4 (2015), 1–19.

[23] Klaus G Grunert, Josephine M Wills, and Laura Fernández-Celemín. 2010. Nutrition knowledge, and use and understanding of nutrition information on food labels among consumers in the UK. *Appetite* 55, 2 (2010), 177–189.

[24] Kristian J Hammond. 1986. CHEF: A Model of Case-based Planning.. In *AAAI*. 267–271.

[25] Morgan Harvey and David Elsweiler. 2015. Automated recommendation of healthy, personalised meal plans. In *Proceedings of the 9th acm conference on recommender systems*. 327–328.

[26] Morgan Harvey, Bernd Ludwig, and David Elsweiler. 2013. You are what you eat: Learning user tastes for rating prediction. In *International symposium on string processing and information retrieval*. Springer, 153–164.

[27] Thomas R Hinrichs. 1989. Strategies for adaptation and recovery in a design problem solver. In *Proceedings of the Workshop on Case-Based Reasoning*. 343–348.

[28] Simon Howard, Jean Adams, and Martin White. 2012. Nutritional content of supermarket ready meals and recipes by television chefs in the United Kingdom: cross sectional study. *Bmj* 345 (2012).

[29] Rong Hu and Pearl Pu. 2010. A study on user perception of personality-based recommender systems. In *International conference on user modeling, adaptation, and personalization*. Springer, 291–302.

[30] Rong Hu and Pearl Pu. 2011. Enhancing recommendation diversity with organization interfaces. In *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, New York, NY, USA, 347–350.

[31] J.Z. Ilich, J.A. Vollono, and R.A. Brownbill. 1999. Impact of Nutritional Knowledge on Food Choices and Dietary Intake of College Students. *Journal of the American Dietetic Association* 99, 9, Supplement (1999), A89.

[32] Sheena S Iyengar and Mark R Lepper. 2000. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology* 79, 6 (2000), 995.

[33] Anthony Jameson, Martijn C Willemsen, Alexander Felfernig, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Li Chen. 2015. Human decision making and recommender systems. In *Recommender systems handbook*. Springer, 611–648.

[34] Dietmar Jannach, Mathias Jesse, Michael Jugovac, and Christoph Trattner. 2021. Exploring Multi-List User Interfaces for Similar-Item Recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 224–228.

[35] Laura Jansen, Ellen van Kleef, and Ellen J Van Loo. 2021. The use of food swaps to encourage healthier online food choices: a randomized controlled trial. *International journal of behavioral nutrition and physical activity* 18, 1 (2021), 1–16.

[36] Mathias Jesse and Dietmar Jannach. 2021. Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports* 3 (2021), 100052.

[37] Yucheng Jin, Nava Tintarev, Nyi Nyi Htun, and Katrien Verbert. 2020. Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Modeling and User-Adapted Interaction* 30, 2 (2020), 199–249.

[38] Eric J Johnson, Suzanne B Shu, Benedict GC Dellaert, Craig Fox, Daniel G Goldstein, Gerald Häubl, Richard P Larrick, John W Payne, Ellen Peters, David Schkade, et al. 2012. Beyond nudges: Tools of a choice architecture. *Marketing Letters* 23, 2 (2012), 487–504.

[39] Yvonne Kammerer and Peter Gerjets. 2014. The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *International Journal of Human-Computer Interaction* 30, 3 (2014), 177–191.

[40] Rex B Kline. 2015. *Principles and practice of structural equation modeling*. Guilford publications, New York, NY, USA.

[41] BP Knijnenburg, MC Willemsen, and R Broeders. 2014. Smart sustainability through system satisfaction: tailored preference elicitation for energy-saving recommenders. In *20th Americas Conference on Information Systems (AMCIS 2014), August 7-9, 2014, Savannah, Georgia, United States.* AIS/ICIS, 1–15.

[42] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems.* 43–50. http://dl.acm.org.dianus.libr.tue.nl/citation.cfm?id=2365966

[43] Bart P. Knijnenburg and Martijn C. Willemsen. 2015. *Evaluating recommender systems with user experiments.* Springer, New York, NY, USA, 309–352.

[44] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4–5 (2012), 441–504.

[45] J.A. Konstan and J. Riedl. 2012. Recommended for you. *IEEE Spectrum* 49, 10 (Oct 2012), 54–61. https://doi.org/10.1109/MSPEC.2012.6309257

[46] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–12.

[47] Ricardo Gomes Lage, Frederico Durao, Peter Dolog, and Avaré Stewart. 2011. Applicability of recommender systems to medical surveillance systems. In *Proceedings of the second international workshop on Web science and information exchange in the medical web.* ACM, New York, NY, USA, 1–6.

[48] Jennifer Mankoff, Gary Hsieh, Ho Chak Hung, Sharon Lee, and Elizabeth Nitao. 2002. Using low-cost sensing to support nutritional awareness. In *International conference on ubiquitous computing.* Springer, 371–378.

[49] Stefanie Mika. 2011. Challenges for nutrition recommender systems. In *Proceedings of the 2nd Workshop on Context Aware Intel. Assistance, Berlin, Germany.* Citeseer, 25–33.

[50] Cataldo Musto, Alain D. Starke, Christoph Trattner, Amon Rapp, and Giovanni Semeraro. 2021. Exploring the Effects of Natural Language Justifications in Food Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization.*

[51] Cataldo Musto, Christoph Trattner, Alain Starke, and Giovanni Semeraro. 2020. Towards a knowledge-aware food recommender system exploiting holistic user models. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization.* ACM, New York, NY, USA, 333–337.

[52] Theodora Nanou, George Lekakos, and Konstantinos Fouskas. 2010. The effects of recommendations' presentation on persuasion and satisfaction in a movie recommender system. *Multimedia Systems* 16, 4-5 (2010), 219–230.

[53] Department of Health UK and Food Standards Agency. 2016. Guide to creating a front of pack (FoP) nutrition label for pre-packed products sold through retail outlets. (2016). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566251/FoP_Nutrition_labelling_UK_guidance.pdf

[54] World Health Organization. 2003. *Diet, nutrition, and the prevention of chronic diseases: report of a joint WHO/FAO expert consultation.* Vol. 916. World Health Organization.

[55] Daniel J O'Keefe. 2002. Guilt as a mechanism of persuasion. *The persuasion handbook: Developments in theory and practice* (2002), 329–344.

[56] Florian Pecune, Lucile Callebert, and Stacy Marsella. 2020. A Recommender System for Healthy and Personalized Recipes Recommendations.. In *HealthRecSys@ RecSys.* ACM, New York, NY, USA, 15–20.

[57] Bartosz Porebski, Karol Przystalski, and Leszek Nowak. 2011. *Building PHP Applications with Symfony, CakePHP, and Zend Framework.* John Wiley and Sons, Indianapolis, IN, USA.

[58] Pearl Pu and Li Chen. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20, 6 (2007), 542–556.

[59] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems.* ACM, New York, NY, USA, 157–164.

[60] Hanna Schäfer, Santiago Hors-Fraile, Raghav Pavan Karumur, André Calero Valdez, Alan Said, Helma Torkamaan, Tom Ulmer, and Christoph Trattner. 2017. Towards health (aware) recommender systems. In *Proceedings of the 2017 international conference on digital health.* ACM, New York, NY, USA, 157–161.

[61] Hanna Schäfer and Martijn C Willemsen. 2019. Rasch-based tailored goals for nutrition assistance systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces.* 18–29.

[62] Benjamin Scheibehenne, Rainer Greifeneder, and Peter M Todd. 2010. Can there ever be too many options? A meta-analytic review of choice overload. *Journal of consumer research* 37, 3 (2010), 409–425.

[63] Alain Starke. 2019. RecSys Challenges in achieving sustainable eating habits.. In *HealthRecSys@RecSys.* CEUR-WS, Aachen, DE, 29–30.

[64] Alain Starke, Christoph Trattner, Hedda Bakken, Martin Johannessen, and Vegard Solberg. 2021. The cholesterol factor: Balancing accuracy and health in recipe recommendation through a nutrient-specific metric. In *CEUR Workshop Proceedings*, Vol. 2959.

[65] Alain Starke, Martijn Willemsen, and Chris Snijders. 2017. Effective User Interface Designs to Increase Energy-efficient Behavior in a Rasch-based Energy Recommender System. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. ACM, New York, NY, USA, 65–73. https://doi.org/10.1145/3109859.3109902

[66] Alain D Starke, Elias Kløverød Kløverød Brynestad, Sveinung Hauge, and Louise Sandal Løkeland. 2021. Nudging Healthy Choices in Food Search Through List Re-Ranking. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 293–298.

[67] Alain D Starke, Justyna Sedkowska, Mihir Chouhan, and Bruce Ferwerda. 2022. Examining Choice Overload across Single-list andMulti-list User Interfaces. In *IntRS'22: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 22, 2022, Seattle, US (hybrid event)*. CEUR-WS, 3–17.

[68] Alain D. Starke and Christoph Trattner. 2021. Promoting Healthy Food Choices Online: A Case for Multi-List Recommender Systems. In *HEALTHI'21: Joint Proceedings of ACM IUI 2021 Workshops*. CEUR-WS, Aachen, DE, 3 pages.

[69] Alain D Starke, Martijn C Willemsen, and Chris Snijders. 2020. With a little help from my peers: Depicting social norms in a recommender interface to promote energy conservation. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 568–578.

[70] Alain D Starke, Martijn C Willemsen, and Christoph Trattner. 2021. Nudging Healthy Choices in Food Search Through Visual Attractiveness. *Frontiers in Artificial Intelligence* 4 (2021), 20.

[71] Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*. Springer, 479–510.

[72] Nava Tintarev and Judith Masthoff. 2012. Evaluating the Effectiveness of Explanations for Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4–5 (Oct 2012), 399–439.

[73] Raciel Yera Toledo, Ahmad A Alzahrani, and Luis Martinez. 2019. A food recommender system considering nutritional information and user preferences. *IEEE Access* 7 (2019), 96695–96711.

[74] Thi Ngoc Trang Tran, Müslüm Atas, Alexander Felfernig, and Martin Stettinger. 2018. An overview of recommender systems in the healthy food domain. *Journal of Intelligent Information Systems* 50, 3 (2018), 501–526.

[75] Christoph Trattner and David Elsweiler. 2017. Investigating the healthiness of internet-sourced recipes: implications for meal planning and recommender systems. In *Proceedings of the 26th international conference on world wide web*. ACM, New York, NY, USA, 489–498.

[76] Christoph Trattner and David Elsweiler. 2019. Food Recommendations. In *Collaborative recommendations: Algorithms, practical challenges and applications*. World Scientific, 653–685.

[77] Christoph Trattner, David Elsweiler, and Simon Howard. 2017. Estimating the healthiness of internet recipes: a cross-sectional study. *Frontiers in public health* 5 (2017), 16.

[78] Christoph Trattner and Dietmar Jannach. 2020. Learning to recommend similar items from human judgments. *User Modeling and User-Adapted Interaction* 30, 1 (2020), 1–49.

[79] Christoph Trattner, Dominik Moesslang, and David Elsweiler. 2018. On the predictability of the popularity of online recipes. *EPJ Data Science* 7, 1 (2018), 1–39.

[80] Anne M Turner, Thomas Engelsma, Jean O Taylor, Rashmi K Sharma, and George Demiris. 2020. Recruiting older adult participants through crowdsourcing platforms: Mechanical Turk versus Prolific Academic. In *AMIA Annual Symposium Proceedings*, Vol. 2020. American Medical Informatics Association, 1230.

[81] Tsuguya Ueta, Masashi Iwakami, and Takayuki Ito. 2011. A recipe recommendation system based on automatic nutrition information extraction. In *International Conference on Knowledge Science, Engineering and Management*. Springer, 79–90.

[82] L Nynke van der Laan and Oliwia Orcholska. 2022. Effects of Digital Just-In-Time Nudges on Healthy Food Choice–a Field Experiment. *Food Quality and Preference* (2022), 104535.

[83] Ellen J Van Loo, Carola Grebitus, and Wim Verbeke. 2021. Effects of nutrition and sustainability claims on attention and choice: An eye-tracking study in the context of a choice experiment using granola bar concepts. *Food Quality and Preference* 90 (2021), 104100.

[84] Michael L Wayne. 2018. Netflix, Amazon, and branded television content in subscription video on-demand portals. *Media, culture & society* 40, 5 (2018), 725–741.

[85] Longqi Yang, Cheng-Kang Hsieh, Hongjian Yang, John P Pollak, Nicola Dell, Serge Belongie, Curtis Cole, and Deborah Estrin. 2017. Yum-me: a personalized nutrient-based meal recommender system. *ACM Transactions on Information Systems (TOIS)* 36, 1 (2017), 1–31.

[86] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 10th ACM Conference on Web Science*. 373–382.