“Serving Each User”: Supporting Different Eating Goals Through a Multi-List Recommender Interface

ALAIN D. STARKE, Marking and Consumer Behaviour Group, Wageningen University & Research, The Netherlands and Department of Information Science & Media Studies, University of Bergen, Norway
EDIS ASOTIC and CHRISTOPH TRATTNER, Department of Information Science & Media Studies, University of Bergen, Norway

Food recommender systems optimize towards a user’s current preferences. However, appetites may vary, in the sense that users might seek healthy recipes today and look for unhealthy meals tomorrow. In this paper, we propose a novel approach in the food domain to diversify recommendations across different lists to ‘serve’ different users goals, compiled in a multi-list food recommender interface. We evaluated our interface in a 2 (single list vs multiple lists) x 2 (without or with explanations) between-subject user study (N = 366), linking choice behavior and evaluation aspects through the user experience framework. Our multi-list interface was evaluated more favorably than a single-list interface, in terms of diversity and choice satisfaction. Moreover, it triggered changes in food choices, even though these choices were less healthy than those made in the single-list interface.

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

Additional Key Words and Phrases: Recommender Systems, Health, Nudges, Food, Goals, User Experience

ACM Reference Format:
Alain D. Starke, Edis Asotic, and Christoph Trattner. 2021. “Serving Each User”: Supporting Different Eating Goals Through a Multi-List Recommender Interface. In *Fifteenth ACM Conference on Recommender Systems (RecSys ’21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3460231.3474232

1 INTRODUCTION AND RELATED WORK

An increasing number of commercial recommender applications present multiple recommendation lists in a single interface [14]. So-called ‘Multi-list Recommender Interfaces’ present multiple item lists stacked on top of each other, accompanying each list with an explanation on what the items in the list represent [10, 31]. The algorithms underlying these lists are typically either based on a variety of recommendation approaches (e.g., using different similarity measures [10, 14]), or employ a single personalization algorithm that is optimized differently across different lists, by constraining the presented items to a certain tag [24], or by re-ranking the top-k set on a specific attribute (cf. [33]).

Commercial examples include video streaming services, such as Disney+ and Netflix. They present movie and TV series recommendations in an explainable multi-list interface [10], typically providing multiple lists that relate to a user’s preferences but which are limited to or optimized for a specific attribute, tag, or genre. For example, these lists would be explained as ‘Drama TV Series’ (genre constraint), ‘Oscar-winning movies’ (movies with a specific tag), or

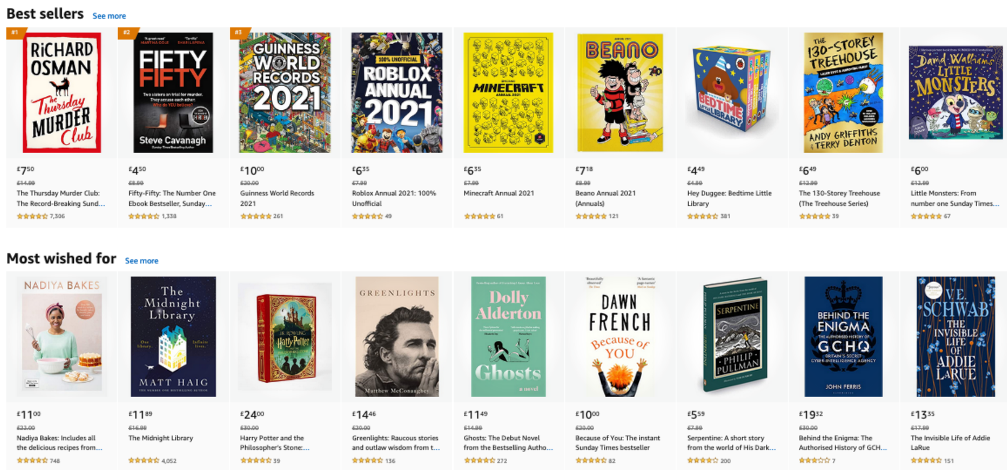

Fig. 1. Amazon utilizing explanations in a multi-list recommender interface.

'Recommended for you' (CF 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 [10].

Multi-list interfaces may also promote items that are not necessarily personalized. An example taken from Amazon is depicted in Figure 1, showing two lists. One list is optimized for overall popularity of items, while the other comprises items that are often put on other users' wish list. Such lists can be inferred without any user history, yet still lead to changes in user preferences by presenting a larger number of items in an organized manner (cf. [24]).

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 [10, 25], research on its use in domains where users have specific behavioral goals is missing [9, 31]. Food is such a domain, where multi-list interfaces have the potential to steer user preferences towards a specific eating goal. In particular, the promotion of healthy food choices has hardly been examined in food recommender studies [26], because many approaches are popularity-based and lead to unhealthy outcomes [8, 36]. Since a user's profile becomes less relevant when she wishes to change her current eating habits [1, 28], 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 [18]), other studies have shown that increasing recommendation diversity could better serve a user's interests [25]. 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).

The commercial multi-list 'benchmark' has yet to be evaluated in a user-centric approach [14]. Whereas its merits are clear in terms of user retention and click-through rates [10], 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?

This paper presents a novel multi-list food recommender interface that is evaluated through a user-centric approach. We employ the user experience recommender framework [16, 17] 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 [24] 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, follow-up studies using eye-tracking methods show higher levels of intention to re-use the interface [4], while a related study by Nanou et al. [20] 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 [11]. For the food domain, we expect that a multi-list recommender system can overcome algorithmic biases towards unhealthy foods 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 [12, 19], the introduction of list-specific explanations is expected to boost its understandability, also given earlier findings on the reduction of cognitive load [20, 25]. 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.

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. For the user-centric evaluation of our multi-list food recommender system and whether it can support healthy eating goals, we propose the following research questions:

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

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

2 METHOD

2.1 Dataset

We developed a food recommender system to address our research questions. It employed recipes from Allrecipes.com, a popular recipe website on which users can upload their own recipes. From a larger database of around 58,000 recipes (which was also used in [35, 37–39]), 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.

2.2 Recommendation Approach and Lists

The recommendation approach we implemented was based on the similar item principle [38]. Hence, given a recipe r_i, we find all top-k most similar recipes r_j. Formally, this can be expressed as follows:

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

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

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, as these are rather representative of human similarity judgments [38], using Term Frequency-Inverse Document Frequency (TD-IDF). We implemented the test-bed as a PHP online application, using the Zend framework and Apache's Lucene search framework [23]. 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 trial in our user study, we randomly selected a reference recipe that matched the predetermined search queries for that trial.

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 (cf. [35]), 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 feature-based explanations (e.g., 'Similar, but with fewer calories'), 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.

2.3 Participants

A total of 366 participants ($M_{age} = 34.24$ years, $SD = 13.23$; 52% male) completed our user study. 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]

2.4 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, their self-reported health, 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 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.

2.5 Research Design

The recommender interface's list design and the inclusion of explanations was subject to a 2x2-between user design. Per trial, users were either presented a single list of 5 recipes or an interface that comprised 5 lists of 5 recipes (25 in total). Moreover, the presented recommendations were either annotated with an overall explanation 'Similar Recipes' (which

[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.

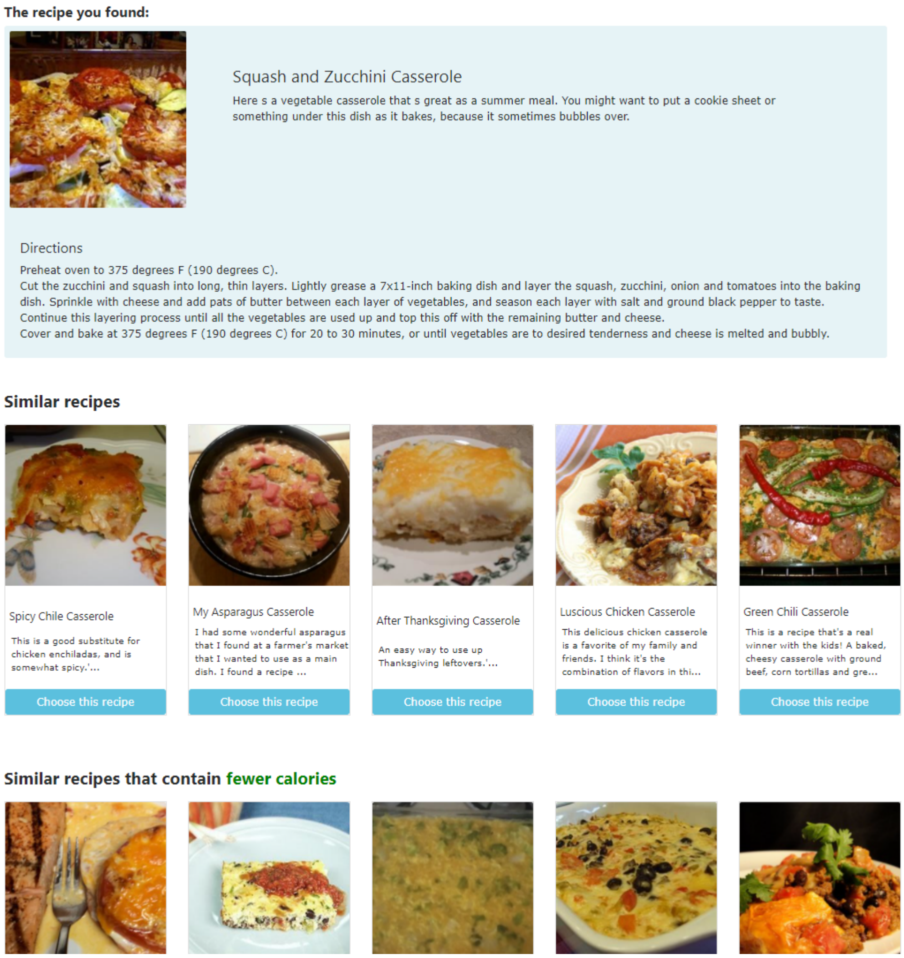

Fig. 2. Partial screenshot of our recommender interface. Depicted at the top is the reference recipe, for which a similar-item recommendation set is retrieved. Depicted here is the multi-list condition with explanations, presenting multiple lists simultaneously.

was considered as a ‘no explanation’ baseline), or shown with list-specific explanations. For the single-list condition, each of the 5 lists in our system were presented once to a user, in a randomized order. For the multi-list condition, all 5 lists in our system were presented simultaneously on each trial, but the vertical order was randomized.

2.6 Measures

2.6.1 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 [29, 32]). At the end of each study, we inquired on their perceived choice difficulty (items adapted from [17]), their perception of the diversity among presented recipes (items adapted from [17]), and how understandable each list was. All items, listed in Table 1, were evaluated through 5-point Likert scales.

2.6.2 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 [21] and was used in earlier studies [22, 30, 33, 37]. In short, a recipe's FSA score was higher if the fat, saturated fat, sugar, or salt content was higher per 100g (cf. [33] 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 RESULTS

3.1 Confirmatory Factor Analysis

We compared a user's evaluation of our single and multi-list interfaces through the recommender system user experience framework [17]. 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 [16].[4]

3.2 Structural Equation Modeling

We organized objective constructs, subjective constructs, and relevant interactions into a path model using Structural Equation Modeling (SEM). As suggested by Knijnenburg and Willemsen [16], 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 larger than the square root of each aspect's AVE (cf. Table 1) [16].[5]

3.2.1 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.

[4]Although some SEM guidelines recommended to use at least three items per latent aspect for small SEM analyses (e.g., [16]), Kline [15] 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.
[5]This model was inferred using all users. We also tested a model in which we excluded users who had not passed the attention check, but this did not lead to significant changes in the path model.

Table 1. Results of the confirmatory factory analysis on user experience aspects. 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

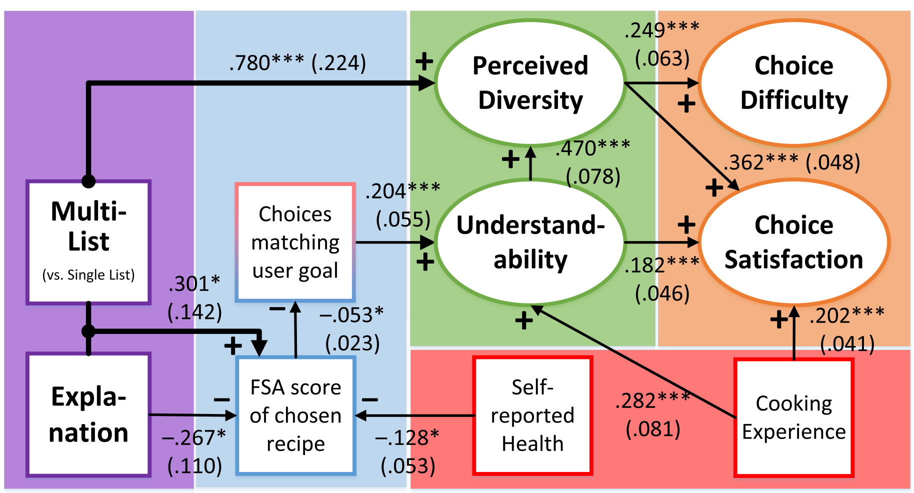

Fig. 3. Structural Equation Model (SEM). Numbers on the arrows represent the β-coefficients, standard errors are denoted between brackets. Effects between the subjective constructs are standardized and can be considered as correlations, other effects show regression coefficients. Aspects are grouped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

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

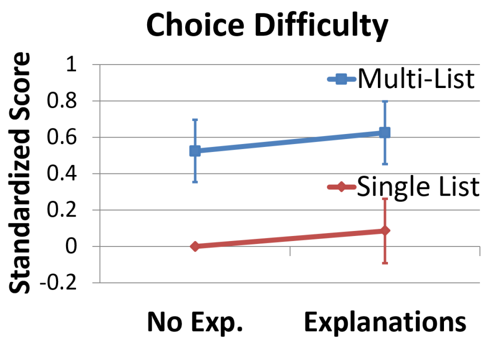

Fig. 4. Standardized scores for the choice difficulty experience aspect across conditions. Errors bars represent 1 S.E.

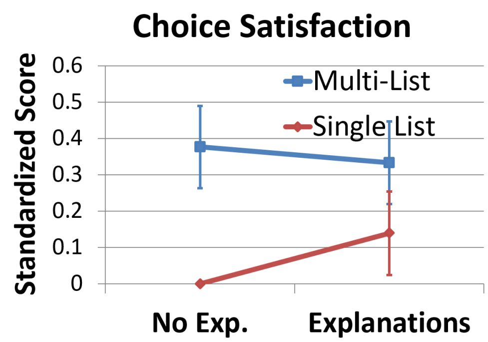

Fig. 5. Standardized scores for the choice satisfaction experience aspect across conditions. Errors bars represent 1 S.E.

could be observed. Second, diversity was also positively related to choice satisfaction: $\beta = .362$, $p < 0.001$, leaving 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.

3.2.2 Choice Metrics (RQ2). The second main path in Figure 3 stemmed from both objective system aspects and followed through choice metrics towards 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, 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. Since this only applied to the multi-list conditions, we depicted the distribution of lists from which a recipe was chosen per multi-list condition 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 (cf. Figure 3) shows that more choices that matched a user's goal led to a higher 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 towards this experience aspect was not significantly mediated by the aforementioned interaction metrics, indicating that they were related but not causally mediated.

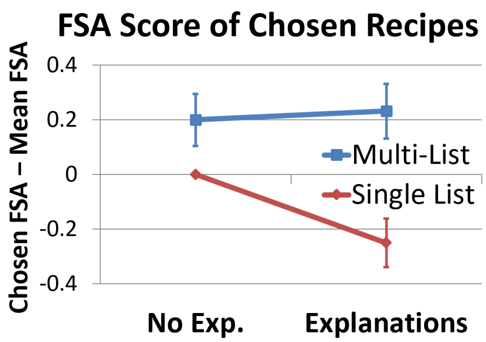

Fig. 6. FSA score of chosen recipes, relative to the mean FSA score of the presented recommendations. Negative values indicate that a relatively healthy recipe was chosen, vice versa for positive values. Errors bars represent 1 S.E.

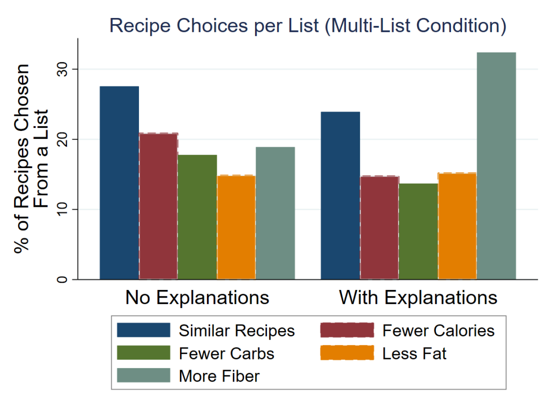

Fig. 7. Distribution of individual lists from which recipes were chosen – for the Multi-List condition only. In the Single List condition, the distribution was flat (0.2 for each list).

3.2.3 User Characteristics. Finally, two user characteristics (in red) also significantly affected a user's choices, perception, and evaluation.[6] 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 ($\beta = .202$, $p < 0.001$). 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.

4 DISCUSSION

Recommender interfaces that present multiple item lists in a single interface are being used in an increasing number of commercial applications [10]. Nonetheless, studies on how they are evaluated by users are limited to specific domains [24, 25], while its current use in consumer and leisure domains (e.g., e-commerce, movies) 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 [31].

The current study is the first to empirically examine 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 [16]. In performing such a user-centric evaluation, we have examined whether a multi-list interface can support healthier recipe choices and user food goals, which we have examined by designing nutrient-specific recommendation lists. Whereas other studies are based on single-item evaluations [14] or analyses in which latent aspects are evaluated separately [24, 25], we have linked different latent evaluation aspects in a path model.

With regard to [RQ1], we find that users are more satisfied with recipes they have chosen from a multi-list interface, compared to a single interface. Moreover, they also report higher levels of perceived diversity. At the same time, we

[6]We had also explored possible interaction effects between user characteristics and interaction metrics and evaluation aspects, but found none.

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. [3, 27]). These findings are consistent with earlier studies on choice overload [13], which describe that people evaluate larger choice sets more favorably, but also have a harder time in making a decision, which sometimes leads to choice deferral [6]. An important finding is 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 [20, 24]. 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.

With regard to the chosen recipes, we have observed a variety of choices from non-similar 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. [5]).

Although we do not find clear advantages of the use of explanations in multi-list interfaces, it must be noted that current study only put forth recommendation sets of 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 [10]. 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 [24, 34], but might not significantly affect choice-related outcomes.

Furthermore, it could be argued that the use of a recommendation approach that is not user-personalized is a limitation. However, many recipe websites and recommender systems use similar-item recommendation approaches that are much like our study design [38]. Moreover, the findings from our similar-item approach is 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 the current study 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 [7], which can even lead to healthier choices [33]. Due to our controlled between-subject design, 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) [33], image attractiveness can be added as an additional feature to a recipe database and be used to further personalize recommendations, as also done in industry applications [10].

Future studies should test our findings in a more naturalistic setting. The number of recipes recommended should not necessarily be limited to 25, while the evaluation of a personalized scenario would add to both the recommender system literature, as well as to the digital food literature. Moreover, the current study has merely focused on lists that optimize for a single recipe nutrient (e.g., fat), while multi-list interfaces that consider dietary restrictions would also be relevant for users with specific eating goals.

ACKNOWLEDGMENTS

This work was 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.

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] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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

[22] 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.

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

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

[25] 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.

[26] 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.

[27] 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.

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

[29] 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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

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

[39] 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.