UMUAI manuscript No.
(will be inserted by the editor)

Intra-List Similarity and Human Diversity Perceptions of Recommendations: The Details Matter*

Mathias Jesse · Christine Bauer ·
Dietmar Jannach

Abstract The diversity of the generated item suggestions can be an important quality factor of a recommender system. In offline experiments, diversity is commonly assessed with the help of the *intra-list similarity* (ILS) measure, which is defined as the average pairwise similarity of the items in a list. The similarity of each pair of items is often determined based on domain-specific meta-data, e.g., movie genres. While this approach is common in the literature, it in most cases remains open if a particular implementation of the ILS measure is actually a valid proxy for the human diversity perception in a given application. With this work, we address this research gap and investigate the correlation of different ILS implementations with human perceptions in the domains of movie and recipe recommendation. We conducted several user studies involving over 500 participants. Our results indicate that the particularities of the ILS metric implementation matter. While we found that the ILS metric *can* be a good proxy for human perceptions, it turns out that it is important to individually validate the used ILS metric implementation for a given application. On a more general level, our work points to a certain level of oversimplification in recommender systems research when it comes to the design of computational proxies for human quality perceptions and thus calls for more research regarding the validation of the corresponding metrics.

[*] This paper or a similar version is not currently under review by a journal or conference. This paper is void of plagiarism or self-plagiarism as defined by the Committee on Publication Ethics and Springer Guidelines.

M. Jesse
University of Klagenfurt, Austria
E-mail: mathias.jesse@aau.at

C. Bauer
Utrecht University, Netherlands
E-mail: c.bauer@uu.nl

D. Jannach
University of Klagenfurt, Austria
E-mail: dietmar.jannach@aau.at

Keywords recommender systems · diversity perception · intra-list similarity · metric validation · user study

1 Introduction

The main task of a recommender system is to surface items that are relevant for users in their current context. However, it is well known that in many cases, being accurate in terms of predicting which items are relevant may not be enough (McNee et al, 2006). One established additional quality criterion that is important in many application domains is that of *diversity*. Non-diverse recommendation lists may not only appear monotone to users, but they may also lead to limited discovery if they only cover a limited part of the available catalog, e.g., only movies from the most preferred genre. Therefore, various algorithmic approaches were proposed over the years to ensure a certain level of diversity in the recommendations, see Kaminskas and Bridge (2016) and Kunaver and Požrl (2017) for related surveys.

As diversity may be considered a subjective concept, several *user studies* focused on understanding in which ways the *perceived* diversity of a set of recommendations may affect other quality factors such as perceived accuracy (Pu et al, 2011; Ekstrand et al, 2014; Willemsen et al, 2016; Nilashi et al, 2016). However, probably a much larger number of publications on diversity uses offline experiments as a research methodology and therefore rely on objective, computational metrics to quantify the extent of diversity of a given recommendation list. A very common approach to quantify a list's diversity is to consider the pairwise similarities of the items. Early proposals were made over 20 years ago in the context of retrieval-based and conversational recommenders (Bradley and Smyth, 2001; McGinty and Smyth, 2003). Soon after, Ziegler et al (2005) popularized this approach under the term *intra-list similarity* (ILS) in their early work on topic diversification in a book recommendation setting.

Technically, the ILS of a set P of (recommended) items is defined in Ziegler et al (2005) as follows:

$$ILS(P) = \frac{\sum_{p_i \in P} \sum_{p_j \in P, p_i \neq p_j} sim(p_i, p_j)}{2} \tag{1}$$

where *sim* is an arbitrary function that returns a similarity score for two items. Often, the score is standardized to lie in $[-1, 1]$ or $[0, 1]$. Note that in Ziegler et al (2005) the sum of the pairwise similarities is divided by two. Reporting the *average* pairwise similarity, as done in earlier in Bradley and Smyth (2001), is however more common today, and the denominator in this case would be the number of comparisons, i.e., $(|P|(|P|-1))/2$.

One important aspect of the technical formulation of the ILS metric is that the order of the elements in a list does not matter. The same ILS value will be returned when all similar items are dispersed across the list or when they are clustered, e.g., at the beginning or end of the list.[1] Another feature of the ILS metric definition is that it is *generic* in a sense that is not tailored to a particular application setting. Depending on a specific application, any suitable similarity function can be plugged in. Ziegler et al (2005) used Amazon's book taxonomy to diversify the results of their recommender. Later works relied on various other types of (meta-)data, for example, movie genres (Vargas et al, 2012), food ingredients (Hauptmann et al, 2021), artist similarity based on social tags (Jannach et al, 2017), or latent topic models from the users' interactions (Shi et al, 2012).

In many works on diversity in recommendation, the rationale for using a particular similarity function is however not discussed in depth, and it might simply be based on the availability of item meta-data. In this regard, we may therefore face an oversimplification in terms of the selection and operationalization of diversity metrics. Most works come without an evaluation of the selected diversity metric—neither for a specific application nor across settings or domains. Instead, it is simply assumed—often without evidence or theoretical underpinning—that the chosen computational metric is aligned with human perceptions. Moreover, while many published works may show that a particular diversity-aware algorithm has an effect on the chosen ILS metric, it is often unclear if the algorithm would actually impact the users' diversity *perception*. Understanding the users' perception is however important because it may significantly affect the quality perception of the recommender system and the behavioral intentions of users, as mentioned above (Pu et al, 2011; Ekstrand et al, 2014; Willemsen et al, 2016; Nilashi et al, 2016).

In this work, we address this largely open research gap and investigate *to what extent different ILS metric implementations are suitable proxies for the diversity perception of users.* For that purpose, we conducted a number of user studies in two application domains, involving over 500 participants. In these studies, we presented the participants with recommendation lists that had different diversity levels according to a particular ILS metric, and we then contrasted the ILS-based diversity values with the participants' self-reported diversity perceptions. Our studies led to two main insights. First, we find that ILS *can* be a valid proxy for user-perceived diversity. In both application domains, we found a metric implementation that correlated well with user perceptions. Second, however, we found that the particularities matter: using different metric implementations for diversification result in varying diversity perceptions, and what works well in one domain does not necessarily work well in another. Overall, our work therefore confirms our conjecture regarding a certain level of oversimplification in our research practices when it comes to studying diversity-aware recommendation algorithms. Correspondingly, more research seems to be needed so that future research in this area can build on validated metrics.

[1] Ge et al (2012) reported some indications that the similarity perception of users may possibly depend on the positions of the items.

The remainder of the paper is organized as follows. After discussing previous works in Section 2, we describe the details of our user studies in Section 3. The results are presented and discussed in Section 4. We conclude our work with a summary of the main findings and an outlook on future research in Section 5.

2 Related Work

The main concepts of interest in our study are diversity metrics for recommendation lists and the diversity perceptions of humans. In terms of terminology, let us note upfront that 'similarity' and 'diversity' are often considered to be inversely related concepts in the literature and that the terms are sometimes used in an almost interchangeable manner. Technically, as mentioned above, the diversity of a list is commonly computed as a mathematical inverse of a metric that is based on similarity. In user-centric evaluations, by comparison, both questions related to the perceived similarity and the perceived diversity are sometimes used to assess the diversity of a list, e.g., in Pu et al (2011).

In the following sections, we first present related work on intra-list similarity (ILS), which is the central metric in our work, and discuss other similarity metrics used in recommender systems research (Section 2.1). Thereafter, we review previous works that studied human perceptions of similarity and diversity. First, we discuss works that focus on the similarity perception of item pairs (Section 2.2); subsequently, we present related work that focuses on the diversity perception of lists of items (Section 2.3).

Overall, our present work is different from the discussed prior works in that we aim to validate whether commonly used metrics to assess the diversity of lists of items are indeed valid proxies of human perceptions. For that purpose, we investigate to what extent the particularities of the specific implementation of ILS matter for two important application domains.

2.1 Intra-list Similarity and other Diversity Metrics

ILS is probably the most commonly used metric to capture diversity in recommender systems in the literature (Du et al, 2021). As described earlier, ILS is based on pairwise similarity comparisons, and a higher ILS score denotes a lower level of diversity. Other names for ILS are *intra-list diversity* (Vargas et al, 2014; Mauro and Ardissono, 2019) or *intra-list distance* (Lin et al, 2020).

The generic ILS definition provides the flexibility to plug in any suitable similarity function. In the literature, a variety of similarity functions was used for different applications, using different types of content information or metadata. Kaminskas and Bridge (2016) point out that there is no guarantee that lists with a high ILS metric value are also perceived as highly similar. When using a particular implementation of the ILS metric, it is in principle necessary to validate that it is indeed a good proxy for user perceptions. This validation is, however, rarely done and our present research addresses this question for two ILS implementations.

We note that—besides ILS—also various other metrics are used to capture diversity in recommender research. Kunaver and Požrl (2017) provide a detailed literature review of diversity metrics. However, it turns out that several of these metrics are variations or extensions of each other, which means only a small number of distinct metrics are actually used in the literature.

An example of an approach to capture a quite different concept of diversity is to use the Gini coefficient. The Gini coefficient is a measure of distributional inequality and is, for example, used in economics to capture income inequality. In the context of recommender systems, the coefficient can be used to measure how often individual items are recommended (and subsequently more likely purchased). If a recommender system makes the same item suggestions to everyone, the distribution will be very skewed, leading to a high Gini coefficient and a *concentration bias* on a few items (Fleder and Hosanagar, 2007)[2]. Differently from the ILS metric, diversity assessments based on the Gini index are not based on the analysis of individual lists, but on how often individual items appear in recommendation lists across users, which is why it is called "aggregate diversity" in Adomavicius and Kwon (2012). Questions of aggregate diversity, concentration biases, and related concepts of coverage—see Jannach et al (2015) for an in-depth analysis—are not the focus of our present work, which aims to study human diversity perceptions at the level of individual lists.

In many research works on recommendation diversity, the goal is to balance the typical trade-off between accuracy (relevance) and diversity metrics, see Zheng and Wang (2022); Jannach (2022) for related surveys on multi-objective recommender systems. An alternative to such approaches is to design evaluation metrics that combine various aspects, including accuracy, diversity, or novelty in a single metric. In the area of information retrieval, Clarke et al (2008) for example proposed to consider various such aspects when computing the nDCG (*Normalized Discounted Cumulative Gain*) of an item ranking. Later on, similar ideas were proposed for recommendation problems—e.g., in Vargas (2011) and Vargas and Castells (2011)—to design *relevance-aware* beyond-accuracy metrics. Technically, such metrics are however again often based on the ILS metric. In our present work, we focus on human diversity perceptions independent of the individual relevance of the items for a given user.

Overall, while the review by Kunaver and Požrl (2017) shows that there are several alternative approaches to computationally assessing the different notions of diversity, the ILS is the most frequently used metric in the literature, and we therefore use it as the basis for our research.

[2] The Herfindahl index is used as an alternative measure of concentration biases, for instance, in Adomavicius and Kwon (2012).

2.2 Assessing the Similarity Perception of Item Pairs

In the context of recommender systems, similarity functions play a central role in different ways. In content-based recommendation approaches in general, similarity functions serve as the basis to assess the match between a given item and a user's past preferences (de Gemmis et al, 2015). Also, similarity functions are commonly a main component that determines the item ranking for the problems of similar-item recommendation (Brovman et al, 2016) and next-item recommendation (Zeng et al, 2019). Moreover, similarity functions are typically the foundation of diversification approaches (Kunaver and Požrl, 2017; Ziegler et al, 2005; Vargas and Castells, 2011; Chen et al, 2013), where the goal is to match the users' diversity needs or to support their exploration efforts (Tsai and Brusilovsky, 2018).

For all such purposes, it is important that the chosen similarity function reflects user perception, an aspect that usually has to be validated through corresponding user studies, see, e.g., Ekstrand et al (2014). In the following, we review works that studied user perceptions based on *pairwise* item similarity judgments.

In Colucci et al (2016), participants judged the similarity of *movies* in pairwise comparisons using a binary judgment (yes–no). For 62 % of the evaluation pairs there was complete consensus among the participants. Yet, these human judgments were only partly aligned with the output of three algorithmic similarity functions (with the highest precision value being only .55). Building on this dataset, Wang et al (2017) designed two content-based recommendation approaches, where one considered human perceptions in the recommendation process whereas the other did not. Their experiments showed that users indeed preferred the recommendations that considered the human similarity judgments. The results by Colucci et al (2016) and Wang et al (2017) indicate that humans largely agree in their similarity perceptions of item pairs, while these perceptions are not aligned with algorithmic similarity functions. Further, their works demonstrate that different similarity functions result in discrepancies between objective similarity measures and human perception. Differently from their works, which focus on pairwise item similarities, our work focuses on the perception of entire lists and the correspondence with the ILS measure.

The movie domain was also the focus of the work by Yao and Harper (2018). In their work, the authors evaluated similarity scoring algorithms in terms of how well they reflect the users' perceptions. To this end, study participants had to rate the similarity of movie pairs which were selected with six different similarity scoring algorithms spanning a range of activity- and content-based approaches. The results suggest that content-based approaches to defining the similarity of movies best reflect the users' perception of similarity.

Trattner and Jannach (2019) studied in more depth how specific item features (e.g., title, plot, movie poster)—when used in a content-based similarity function—correlate with the similarity of items as perceived by users. Differently from Yao and Harper (2018), their study design required participants to compare a reference movie with a list of movies. The similarity measure based on tags reflected user perceptions well, which was also shown in Yao and Harper (2018). Moreover, capturing similarity in the latent space using matrix factorization proved to be particularly powerful as it did not only reflect user perceptions well but was also the approach that led to the highest usefulness scores in terms of the participants' interest in trying out a movie recommendation. In our present work, we therefore rely on latent item representations when comparing items in two of our studies. Moreover, as this approach is domain-independent, it allows us to make a cross-domain comparison.

Human similarity judgments were also central to the work by Lee (2010) in the *music* domain. In their work, the authors collected human judgments on how "musically similar" pairs of songs were via Amazon's Mechanical Turk platform, and they then compared those judgments with a ground truth of expert judgments. One main finding of their work was that crowdsourcing—as also done in our present study—can be considered a reliable source for music similarity judgments. Some earlier work in the music domain (Ellis et al, 2002), however, indicated that finding a computational metric that gives "reasonable agreement" with the human judgments can be challenging. And Downie et al (2007) found that providing participants with a similarity definition (here: "musically similar" or "melodically similar") influences the similarity judgments.

In the *food* and *recipe* domains, van Pinxteren et al (2011) employed a card-sorting approach to identify ingredients, preparation techniques, cuisine, meal type, and preparation time as the most relevant characteristics that determine the similarity of recipes. In Trattner and Jannach (2019), recommendations based on the recipe instructions, title, and ingredient lists, as well as a combined model, led to the highest similarity perception. In one of our studies in the recipe domain we, therefore, also rely on a combined approach where we consider ingredients and cooking instructions as item meta-data.

Finally, in the *news* domain, Starke et al (2021) compared several similarity functions (based on title, body text, image features) and concluded that using the articles' body text for capturing similarity between news articles comes closest to the human similarity perception. Yet, in this domain, the similarity functions were overall shown to be much weaker than in the recipe and the movie domain. Thus, they suggest to use other features than body text only in case the used similarity function is specifically adapted the news domain. Their cross-domain comparison suggests that in terms of capturing similarity, the news domain is closer to the movie than the recipe domain; although in general the news domain may require similarity functions that are less 'taste-related' than in the recipe and movie domains.

2.3 Assessing the Diversity Perception of Item Lists

While several works have addressed similarity perception, only a few have specifically addressed the diversity perception of lists. In the music domain (specifically, electronic music), for example, Porcaro et al (2022) found that instrument and samples, sub-genre or sub-style, tempo, and mood strongly influence what track lists were considered diverse. On the artist level, the artists' origin and nationality, gender, and skin tone were considered key factors for diverse music lists. They also found that the used metrics reflect the diversity perceptions particularly well for participants coming from Western and educated societies and in the age range between 18 and 35 years.

Although primarily studying similarity perception, Trattner and Jannach (2019) also address how similarity perception relates to diversity perception—specifically, the perception of list diversity. For both the movie and recipe domain, the item features determining perceived list diversity were found to be similar to those determining perceived similarity. For movies, the matrix factorization based approach was a particularly useful approach to capture users' perception of list diversity. Similarly, title, movie poster, plot, and genre were useful features in a content-based approach. In the recipe domain, the image of the dish as the basis for the similarity function reflected the users' list diversity perception best, followed by title, instructions, and the ingredients list. While the work of Trattner and Jannach (2019) and our present work address related topics and adopt similar research methodology, there are stark differences between the two works. Trattner and Jannach (2019) aimed to find ways to construct similar-item recommendations in a reliable way. Specifically, they investigate which item features determine the perceived similarity of a given pair of items. Our work, in contrast, aims to validate if commonly used metrics for lists of items are suitable proxies for human perceptions.

Overall, the examined prior work indicates that not all similarity functions correlate equally well with the similarity perceptions of users. Thus, the findings in the literature support the hypothesis investigated in our present work that the specifics of how a similarity metric is implemented matter.

3 Experimental Design

Goals of Studies We recall that the goal of our research is to assess and validate to what extent different ILS metrics correlate with the users' perception of diversity in two popular application domains of recommender systems, movies and recipes. Further, based on the observations by Ge et al (2011), we investigate whether the items' order within a list impacts the users' perceptions.

Overview of Studies Overall, we conducted four complementary studies to address these aspects. Essentially, in each of the studies, the participants were shown recommendation lists with different levels of ILS and asked to report their diversity perceptions. Specifically, we studied the users' perceptions when *(i)* latent-item representations were used to compare items and when *(ii)* application-specific similarity measures were applied. For both, we executed the studies in the domains of movies and recipes. Figure 1 shows an overview of our four studies: *Study-1*$_{movies}$ and *Study-1*$_{recipes}$ rely on latent item vectors and involve nine different list types with varying ILS levels and item orders as our manipulated variables. *Study-2*$_{movies}$ and *Study-2*$_{recipes}$ use domain-specific ILS metrics. In this second set of studies, we focus only on varying the ILS levels (low, mid, high), which leads to three list types per domain. In all studies, each list shown to users contained *seven* items. We considered larger list sizes to be cognitively too challenging for study participants, cf. Miller (1956); Jensen and Lisman (1996).

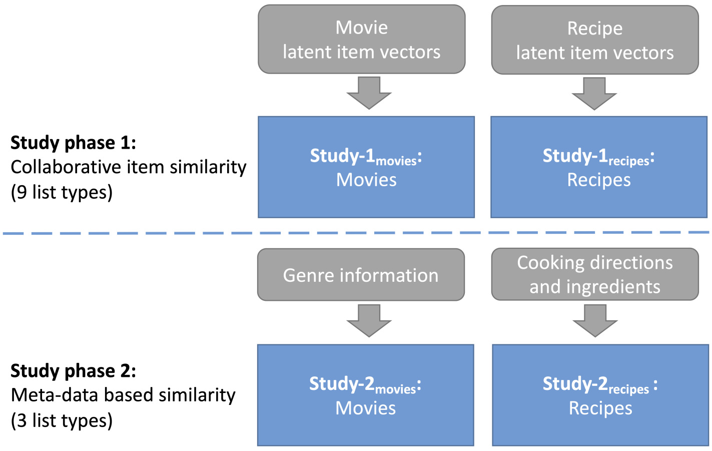

Fig. 1 Overview of the studies

In our experiment, we employ a *mixed design*. Each participant rated three lists, which constitutes the within-subject element of the design. While all participants rated three lists, the selection of the lists was randomized and, thus, the set of lists varied across participants, which constitutes a between-subjects element of our design.

Subsidiary Research Questions While the main focus of our work is validating the ILS metric, we also designed the studies to answer relevant subsidiary questions. First, we addressed whether the familiarity of items influences the perception of a list. Similar to Porcaro et al (2022), where domain knowledge played a role in diversity perception, participants without background knowledge regarding a specific set of items might perceive the list differently (e.g., more diverse) than participants with more knowledge. Second, we addressed whether different diversity levels impacted the participants' decision processes, e.g., in terms of their perceived choice difficulty or choice confidence, cf. Pu et al (2011); Ekstrand et al (2014). Third, we investigated if the popularity of the items had any impact on the perception of the lists. As observed in other works (e.g., Abdollahpouri et al, 2019), recommender systems may face the problem of popularity bias, which could eventually alter the way users perceive the quality of the recommendations. Fourth, we checked for any gender-specific differences in the responses of the participants. Lastly, we also explored what would happen if participants were presented with only one pair of very dissimilar items, compared to presenting an entire list with high diversity.

3.1 Creating Diverse Recommendation Lists

Creating Lists with Varied Diversity Levels To create recommendation lists of different diversity levels, we followed an approach which is both repeatable and as free as possible from potential researcher bias. The general idea of our approach is to first create a set of k recommendation lists for a *randomly* selected set of users from a given dataset using a matrix factorization approach.[3] Then, we compute the ILS values of these recommendation lists to obtain an estimate of the distribution of the ILS values for a given dataset.[4]

Based on this approximated distribution, we then select three lists of the k sampled ones, which we then consider as representations of being of *low*, *medium* and *high* diversity. In our case, we simply used the lists with the lowest ILS and the highest ILS to serve as representatives for a low-ILS and a high-ILS list, respectively. To determine a mid-ILS representative list, we computed the mean of the just discussed highest and lowest ILS values. Then, we picked the one recommendation list from the ten samples, which was closest to this mean ILS value.

In addition to these three lists, we designed two more special lists for our experiments for each domain.

- The first of them, which we call *popsim*, was created by randomly picking one of the 20 most popular items in each dataset and then determining their most similar items among the 100 most popular items. We created this list to study the diversity perception of lists that contain popular similar items, which—at least in the movie domain—are also often items that users are familiar with. We note that popular item recommendations are commonly used as a simple baseline in research works. A validation of this baseline approach is therefore a particularly useful reference point for works on recommender systems algorithms.
- The second list, which we named *upp*, serves as a proxy for an upper bound for ILS values that we may realistically observe in practice. In the movie domain, we selected a collection of *sequels* for this purpose (*"Batman"*); in the recipe domain, we manually picked a set of highly similar recipes for muffins. With *upp*, we created lists with a set of items with a close-to-maximum ILS value.

[3] We note that matrix factorization still is a highly effective method for this purpose (Rendle et al, 2020). Technically, we used the *svd* method implemented in the *numpy* Python library. We determined the optimal number of latent factors for each dataset using grid search. We share all code online at `https://github.com/Belzex/ILS_Study`.

[4] In our experiments, we used $k = 10$ sampled lists. In future works, a larger value for k may be explored in order to obtain a closer approximation of the distribution.

This process gives us *five* lists with different levels for the ILS values. An overview of these lists is given in Table 1.

Abbr.	**List title**	**Description**
$rec_{low\text{-}ILS}$	Recommendation-based (low-ILS)	This list is created by choosing ten random users and creating top-N recommendations using matrix factorization. We then select the list with the **lowest** ILS.
$rec_{mid\text{-}ILS}$	Recommendation-based (mid-ILS)	Created like $rec_{low\text{-}ILS}$, but we select the list with a **medium** ILS.
$rec_{high\text{-}ILS}$	Recommendation-based (high-ILS)	Created like $rec_{low\text{-}ILS}$, but we select the list with the **highest** ILS.
popsim	Popular & Similar	We create this list by randomly selecting one of the top 20 items in the dataset and finding the six most similar items in the top-N items. The order of the resulting list was randomized.
upp	Upper Bound	A list of related films in a franchise in the movie domain (*Batman*); a set of manually selected very similar muffin recipes in the food domain.

Table 1 Overview of how recommendation lists of increasing similarity were created.

In terms of technical details, we relied on two public datasets in our experiments that contain user-item ratings and metadata, *MovieLens-25M*[5] for the movie domain and the *FoodRecSys* dataset[6] for recipes. We considered two ways of computing ILS values for the created lists:

- *A domain-independent approach:* Given the insights from earlier research (Trattner and Jannach, 2019; Yao and Harper, 2018), we measured the ILS of a set of recommendations based on the cosine similarity of latent item vectors (or: embeddings) of the list elements. We obtained the latent vectors directly from the used matrix factorization model.
- *A domain-specific approach:* In this case, we used item meta-data—genre for movies and recipe descriptions for the recipes—to compute the similarity between items. More details for each domain are given later below.

Creating Lists with Varied Item Orderings To study potential effects of the item order, we created *four* additional lists. Two of these lists were permutations of $rec_{mid\text{-}ILS}$ and two were permutations of *popsim*. Our rationale for creating the permutations was to add *(a)* a list variation where the neighboring items are as similar as possible (rec_{max} and $popsim_{max}$), and *(b)*, a variation where the neighboring items are dissimilar (rec_{min} and $popsim_{min}$). Our hypothesis is that clusters of similar items (e.g., at the beginning of a list) may make a set of recommendations appear more similar than when the items are dispersed over the list; see also Ge et al (2012). Technically, to build

[5] `https://grouplens.org/datasets/movielens/25m`

[6] `https://www.kaggle.com/datasets/elisaxxygao/foodrecsysv1`

these lists that either *maximize the similarity of neighbors* or *minimize the similarity of neighbors*, we randomly picked one item of the base list and then added the remaining items iteratively in a way that either the most similar or least similar item was appended to the list. In the beginning, the randomly selected item is used as a reference and the similarity to all the other items is calculated. This process is repeated with the latest appended item as the new reference until all items have been placed in the list. Table 2 provides an overview of these four additional lists with different item orders.

Abbr.	**List title**	**Description**
rec_{min}	Recommendation-based (mid-ILS, **minimize** similarity of neighbors)	From the *Recommendation-based (mid-ILS)* list, we randomly select one item. Based on this reference item, we choose the **least** similar item and put it next. We continue this process until all seven items are placed.
rec_{max}	Recommendation-based (mid-ILS, **maximize** similarity of neighbors)	Similar to rec_{min}, but this time we place the items with the **highest** similarity next to each other.
$popsim_{min}$	Popular & Similar (**minimize** similarity of neighbors)	Identical to rec_{min} with the difference that we use *popsim* as the base list.
$popsim_{max}$	Popular & Similar (**maximize** similarity of neighbors)	Identical to rec_{min}, except that we use *popsim* as the base list.

Table 2 Overview of how recommendation lists with different orders were created.

Overall, this process left us with nine different lists, where the first five were designed to represent lists of different ILS levels, and where the last four where variations of two of the lists for which we changed the item orderings.

3.2 Experiment Flow and Details

Tasks for Participants The tasks for the participants—as outlined in Figure 2—were identical in all four studies.[7] After reading the instructions and after informed consent, each participant was shown three lists of recommendations, all from the same domain. For each of these lists, participants had then to select one item to watch (movies) or try out (recipes) next. In addition, they had to specify their familiarity level for each of the shown items[8], which allows us to analyze if item familiarity has an effect on user perception. Moreover, six questions regarding the similarity/diversity perception and the

[7] The studies can be accessed at `http://moviestudy.eu-west-2.elasticbeanstalk.com/` for movies and `http://recipestudy.eu-west-2.elasticbeanstalk.com/` for recipes.

[8] For the movie domain, the options were “never heard of”, “have heard of”, “have seen it”. The options for the recipe domain were similar: “the type of recipe is new to me”, “I have prepared a similar dish”, “I have prepared this dish”.

choice experience were asked before the participants could move on to the next list. The exact questions for the movie domain are shown in Table 3 and for the recipe domain in the Appendix.

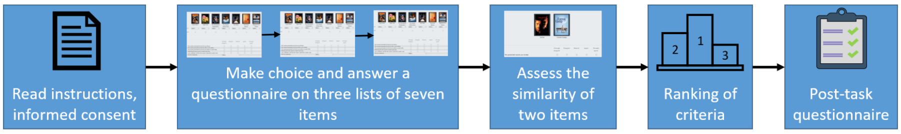
Read instructions, informed consent
Make choice and answer a questionnaire on three lists of seven items
Assess the similarity of two items
2
1
3
Ranking of criteria
Post-task questionnaire

Fig. 2 Experiment process

Question		Focus	Type
Q1 (diversity)	The movies presented in this list are diverse	Diversity	5-point Likert scale
Q2 (variety)	The movies presented in this list offer a rich variety	Diversity	5-point Likert scale
Q3 (similarity)	The movies presented in this list are similar to each other	Diversity	5-point Likert scale
Q4 (choice easiness)	Selecting the next movie to watch was easy	Choice Easiness	5-point Likert scale
Q5 (choice confidence)	I am confident I will like the movie I selected to watch next	Choice Confidence	5-point Likert scale
Q6 (no. of good options)	The list contained . . .	No. of good options	Predefined answers

Table 3 Questionnaire items for each list. We note that the questions were on user *perceptions*, e.g., regarding perceived diversity. For brevity, we only use the names of the aspects here and in other tables, e.g., diversity. The options for Q6 were "more than one good option", "exactly one good option", "no option that I liked".

After providing feedback on the three lists, participants were shown a pair of "extreme" items from one of the recommendation lists, i.e., two items with the lowest pairwise similarity. Again, participants had to express their similarity/diversity perception for this pair. With this additional question, our goal is to assess to what extent the diversity perception of the pair of extremes is correlated with that of an entire list.

Next, to understand what makes a list diverse for participants in a given domain, we asked them to explain in free text what criteria they used to assess the diversity of a list of movies and recipes, respectively. After this step, participants had to rank *predefined* criteria[9] as done in Trattner and Jannach (2019). Finally, we asked demographic questions in the post-task questionnaire (age group, gender, domain expertise).

[9] In the movie domain, the seven criteria were *genre*, *actors*, *cover image*, *plot*, *runtime*, *title*, and *release year*. In the recipe domain, the eight criteria were *title*, *nutrients*, *recipe image*, *ingredients*, *cuisine*, *cooking duration*, *reviews*, and *instructions*. Participants were not forced to rank all criteria.

Setup of the Individual Studies The studies in the first phase were identical except for the domain. The same is true for the second phase. The difference between the two study phases lie *(i)* in *the type* of the used ILS metric and *(ii)* in the set of the examined lists from Table 1 and Table 2.

- **Phase 1:** In Phase 1, the similarity of items was based on latent item vectors as mentioned above. Participants were assigned three lists from all nine types as presented in Table 1 and Table 2 in a randomized process. To avoid that too similar lists were presented to one participant, which may bias their perceptions, we ensured that every participant received exactly one *recommendation-based* list ($rec_{mid\text{-}ILS}$, rec_{min}, or rec_{max}) and exactly one *homogeneous* list (*popsim*, $popsim_{min}$, or $popsim_{max}$). For each of these two list types (i.e., recommendation-based and homogeneous), one list is selected randomly for each participant. Therefore, we are able to compare, across multiple participants, if the order of items, on average, makes a difference on the human perception.
- **Phase 2:** In Phase 2, the ILS metric was based on meta-data: we used genre overlap for the movie domain, as done, e.g., by Vargas et al (2014), and cosine similarity of the Latent Dirichlet Allocation (LDA) embeddings of ingredients and cooking directions as done by Hauptmann et al (2021). We recall that we purposely select different types of meta-data for each domain. Using these different metric implementations, our goal is to assess the diversity perceptions when the ILS metric is based on commonly used meta-data. As we studied potential order effects already in Phase 1, we only consider the three recommendation-based lists with varying diversity levels in Phase 2.

3.3 Participants

We recruited crowdworkers as participants through Amazon's Mechanical Turk platform. To ensure high-quality responses, we only invited crowdworkers who had an approval rate of at least 99 % and who had already successfully completed more than 500 tasks. As another measure to increase the reliability of the responses, we included two attention checks in the questionnaires, and we filtered out those crowdworkers who did not work carefully and missed these checks. On average, the participants needed 9 minutes for the task and received a payment of 1.5$–2$ on successful completion.

Overall, 791 participants completed the study, of which 531 were considered reliable after the attention check. The number of participants per study were 223 for $Study\text{-}1_{movies}$, 195 for $Study\text{-}1_{recipes}$, 55 for $Study\text{-}2_{movies}$ and 58 for $Study\text{-}2_{recipes}$. We recruited a smaller number of participants for Phase 2 because we considered fewer lists in this phase. From all study participants, around 60 % were male and almost 40 % were female. Three participants identified themselves with a different gender. The majority of the participants were between 26–35 years old (41 %).

Participants who were involved in the studies in the movie domain (*Study-1*$_{movies}$, *Study-2*$_{movies}$) an average reported themselves to be quite engaged in movies, and we found no difference between the study populations in that respect. Specifically, when participants were asked about their interest in movies, their average responses where 3.86 ($SD = 0.98$) in *Study-1*$_{movies}$ and 3.87 ($SD = 1.00$) in *Study-2*$_{movies}$.

Looking at the participants' cooking behavior and preferences, we also observe that participants in both study phases were very similar. Around two thirds of the participants in each phase reported cooking at least four times per week. Almost 70 % of participants in both phases had no restrictions in eating behavior, around 17 % were vegetarians, 6 % were vegans, and about 8 % mentioned restrictions for other reasons including religious ones, gluten intolerance, and allergies. We note here that none of the participants with such restrictions provided negative feedback regarding our selection of recipes; and all of them stated that they found at least one recipe in the list that they liked.

4 Results

In this section, we discuss the results of our studies with respect to the given research questions.

4.1 RQ1: Correspondence of ILS and Human Diversity Perception

4.1.1 Results for Phase 1—Using Latent Item Representations

In Phase 1, we used latent item representations as the basis for comparing items with the cosine similarity measure. The computed ILS metric values for the movies and recipe domains are shown in Table 4. Generally, we observe that the range of obtained ILS values is not too large (from 0.35 to 0.51). For the movie domain, the ILS value of the $rec_{mid\text{-}ILS}$ and *popsim* lists were incidentally very similar (around 0.41). Those lists in which only the order was changed (i.e., rec_{min} and rec_{max}; $popsim_{min}$ and $popsim_{max}$) have of course identical ILS values.

The highest ILS value, as expected, is obtained for the *upp* lists, which consist of hand-picked items that are assumed to be very similar. The lowest similarity in the movie domain is found for the $rec_{low\text{-}ILS}$ list. In the recipe domain, the lowest similarity—and thus highest diversity—is observed for the *popsim* list consisting of popular items. This may appear a bit surprising because the list was constructed by adding items similar to a randomly selected popular reference item. However, as we limited the set of candidate items for this list to the 100 most popular items, the overall ILS value remained comparably low. This confirms previous findings that the recommendation of the most popular items in some domains can lead to quite diversified lists (e.g., Ribeiro et al, 2015; Vargas et al, 2014).

Table 4 ILS values for Study Phase 1, using latent item vectors

List	Movies	Recipes
$rec_{low\text{-}ILS}$	0.37	0.38
$rec_{mid\text{-}ILS}$	0.41	0.41
$rec_{high\text{-}ILS}$	0.46	0.44
popsim	0.41	0.35
upp	0.51	0.46
rec_{min}	0.41	0.41
rec_{max}	0.41	0.41
$popsim_{min}$	0.41	0.35
$popsim_{max}$	0.41	0.35

Table 5 and Figure 3 show the average responses for the questions on the subjective perceptions for $Study\text{-}1_{movies}$, ordered by the values for Q1 on diversity. Looking at the responses for Q1, we observe that the ranking of the lists roughly follows the patterns for the ILS metric reported in Table 4. On top of the list, we find $rec_{low\text{-}ILS}$, $rec_{mid\text{-}ILS}$, and rec_{max}, where the latter is a reordered version of $rec_{mid\text{-}ILS}$ to maximize the similarity of neighboring elements. Thus, lists with low ILS values lead to high diversity perceptions. At the bottom end of the list, we find *upp*, as expected, and also $rec_{high\text{-}ILS}$, which has a high ILS value. Generally, we therefore found that the ILS value of a list and the users' perceptions are well aligned in this study. The Spearman correlation coefficients between the ILS with subjective perceptions were -0.39 for diversity, 0.37 for variety, and 0.44 for similarity, which corresponds to a *medium* correlation. All correlations were significant with $p < .001$. The answers to questions Q2 (variety) and Q3 (similarity) are also positively and negatively correlated with diversity, which is also expected.

Table 5 Results for $Study\text{-}1_{movies}$ (movies, 9 lists); numbers show average responses and standard deviations.

List	ILS	Q1	Q2	Q3	Q4	Q5
		Diversity	Variety	Similarity	Choice	Choice
					easiness	confidence
rec_{max}	0.41	4.14 (0.90)	4.11 (0.90)	2.61 (1.31)	3.76 (1.10)	4.26 (0.88)
$rec_{mid\text{-}ILS}$	0.41	4.12 (0.83)	4.03 (1.05)	2.84 (1.24)	4.14 (0.88)	4.27 (0.71)
$rec_{low\text{-}ILS}$	0.37	4.05 (1.00)	4.14 (1.00)	2.59 (1.35)	3.78 (1.04)	4.18 (0.84)
rec_{min}	0.41	3.94 (0.79)	4.01 (0.74)	2.97 (1.20)	4.03 (0.96)	4.12 (0.99)
popsim	0.41	3.86 (0.96)	3.82 (1.11)	3.15 (1.27)	3.65 (1.18)	4.04 (0.99)
$popsim_{max}$	0.41	3.76 (0.91)	3.82 (0.95)	2.91 (1.09)	3.75 (1.07)	4.14 (0.86)
$popsim_{min}$	0.41	3.75 (0.88)	3.75 (0.97)	3.04 (1.12)	3.78 (1.13)	4.13 (0.82)
$rec_{high\text{-}ILS}$	0.46	3.17 (1.14)	3.24 (1.22)	3.68 (1.02)	3.96 (0.96)	4.32 (0.90)
upp	0.51	2.37 (1.32)	2.60 (1.49)	4.43 (0.93)	3.77 (1.12)	4.15 (1.10)

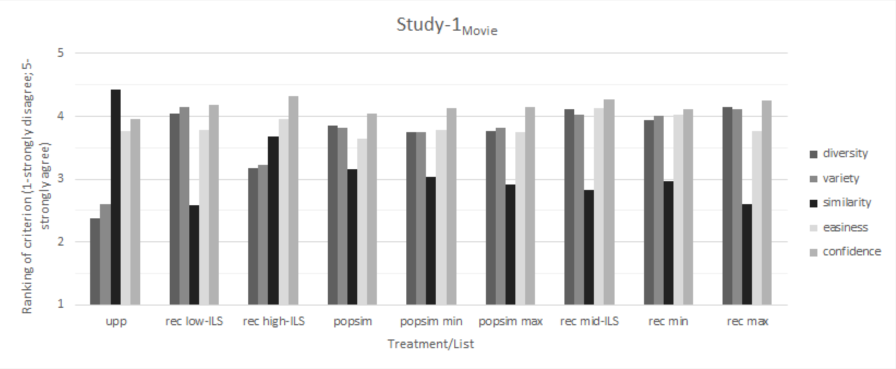

Fig. 3 Average responses in *Study-1$_{movies}$* (movies, 9 lists)

A Kruskal-Wallis test[10] revealed significant differences for the variables *diversity*, *variety*, and *similarity* with all p-values smaller than .001. No significant difference was observed at the significance level $\alpha = .05$ for *choice easiness* and *choice confidence*. For choice easiness, we however observe some tendency that selecting from the lists with popular items (e.g., *popsim*) and with high similarity (e.g., *rec$_{high\text{-}ILS}$*) was perceived to be more difficult; with $p = .06$, the overall differences according to the Kruskal-Wallis test were only significant with an α level of .1. [11]

We then performed a post-hoc analysis with Wilcoxon's signed rank test, in which we made pairwise comparisons. Bonferroni correction was applied to the alpha level to account appropriately for these multiple comparisons. The observations for *diversity*, *variety* and *similarity* are often similar and we only report selected results for *diversity* here.[12]

Looking at the obtained numbers, we found that the diversity perception of a list of movie sequels (*upp*) and the recommendation list with the highest ILS value (*rec$_{high\text{-}ILS}$*) was significantly lower than for all other lists. Interestingly, while we found that the difference between *rec$_{high\text{-}ILS}$* and *rec$_{mid\text{-}ILS}$* was significant, the difference between *rec$_{mid\text{-}ILS}$* and *rec$_{low\text{-}ILS}$* was apparently not noticed by the participants. The responses obtained for the *popsim* list were

[10] As every participant gave feedback to three lists, these three assessments per participant might possibly be correlated, which would not meet the test assumption that the observations are independent. To assess if this problem exists, we fit two types of regression models for each variable. One regression assumed that all observations were independent; the other was a mixed-effects model that considered the hierarchical (user-level) structure in the data. In the mixed-effects model, we did not observe a high interclass correlation coefficient. Moreover, the obtained regression coefficients for both models were almost identical. We therefore conclude that the participant-related effects are minimal, and using the Kruskal-Wallis test as well as Wilcoxon tests for the post-hoc analyses is appropriate. The same analysis was performed for the other three studies.

[11] We applied Kruskal-Wallis test because the prerequisites for applying ANOVA were not fulfilled.

[12] The detailed statistics can be found in the Appendix.

not significantly different from the other lists (except for the sequels in the *upp* list).

Table 6 shows the result for the recipe domain in Phase 1, i.e., again using latent item vectors as a basis for the similarity computations and ordered by the values for Q1 on diversity [13]. In line with the movie domain, a Kruskal-Wallis test indicated significant differences for *diversity*, *variety* and *similarity*, and no difference for *choice easiness* and *choice confidence*. Looking at the results of the post-hoc tests, however, the differences between the various treatment groups, i.e., lists of different diversity levels, are quite small and statistically *not* significant. An exception is the manually constructed *upp* list, which is however not a "natural" recommendation list but serves as an upper bound in our study. Notably, the difference between the natural list $rec_{low\text{-}ILS}$, $rec_{mid\text{-}ILS}$, and $rec_{high\text{-}ILS}$ are not significant for diversity, variety and similarity. In other words, differently from *Study-1*$_{movies}$, the ILS metric in *Study-1*$_{recipes}$ *cannot be considered a reliable proxy for user perceptions.* This suggests that the particular choice of how the ILS metric is implemented matters, and what works in one application might not work well in another.

Table 6 Results for *Study-1*$_{recipes}$ (recipes, 9 lists); numbers show average responses and standard deviations.

List	ILS	Q1 Diversity	Q2 Variety	Q3 Similarity	Q4 Choice easiness	Q5 Choice confidence
$rec_{low\text{-}ILS}$	0.38	4.31 (0.63)	4.33 (0.68)	2.87 (1.44)	3.96 (1.07)	4.33 (0.74)
$rec_{high\text{-}ILS}$	0.44	4.22 (0.57)	4.26 (0.66)	2.82 (1.32)	3.96 (1.14)	4.21 (0.66)
rec_{min}	0.41	4.16 (0.76)	4.27 (0.80)	2.98 (1.28)	4.17 (1.14)	4.33 (0.60)
$popsim_{min}$	0.35	4.16 (0.76)	4.27 (0.80)	2.98 (1.28)	4.17 (1.14)	4.33 (0.60)
$rec_{mid\text{-}ILS}$	0.41	4.16 (0.72)	4.23 (0.77)	2.53 (1.21)	3.86 (0.81)	4.27 (0.65)
rec_{max}	0.41	4.05 (0.62)	3.95 (0.81)	2.88 (1.21)	3.79 (1.07)	4.25 (0.70)
$popsim_{max}$	0.35	4.05 (0.62)	3.95 (0.81)	2.88 (1.21)	3.79 (1.07)	4.25 (0.70)
popsim	0.35	3.80 (0.92)	3.77 (0.96)	3.28 (1.13)	4.02 (0.96)	4.17 (0.98)
upp	0.48	2.37 (1.22)	2.68 (1.46)	4.19 (0.90)	3.65 (1.04)	3.95 (0.89)

4.1.2 Results for Phase 2—Using Domain-Specific Meta-Data

Next, we look at the results obtained in Phase 2, where we used application-specific meta-data to determine the similarity between two items and thus compute the ILS metric. Table 7 and Figure 4 show the results for the movie domain, where we used genre information when comparing movies. Recall that we only investigated three lists ($rec_{low\text{-}ILS}$, $rec_{mid\text{-}ILS}$, and $rec_{high\text{-}ILS}$) in this phase, as the other research questions regarding the item ordering, can be answered with the results obtained in Phase 1. The results in Table 7 show a trend in terms of *diversity*, *variety*, and *similarity*, but the differences are

[13] See Figure 5 in the Appendix for a visual representation of the data.

too small to reach statistical significance according to a Kruskal-Wallis test with $\alpha = .05$.[14] A similar trend can be observed for *choice easiness*, where the difficulty seems to increase when the list is less diverse. But again, the differences are not statistically significant at the .05 threshold level ($p = .15$). Overall, the results for diversity, variety, and similarity in *Study-2*$_{movies}$ are *not* aligned with the results in *Study-1*$_{movies}$, where we could observe a difference in user perceptions.

Table 7 Results for *Study-2*$_{movies}$ (movies, 3 Lists); numbers show average responses and standard deviations.

List	Q1 Diversity	Q2 Variety	Q3 Similarity	Q4 Choice easiness	Q5 Choice confidence
$rec_{low\text{-}ILS}$	4.09 (0.61)	4.20 (0.68)	2.64 (1.12)	3.95 (0.92)	4.35 (0.75)
$rec_{mid\text{-}ILS}$	4.11 (0.57)	4.07 (0.77)	2.75 (1.09)	3.98 (0.87)	4.31 (0.74)
$rec_{high\text{-}ILS}$	3.82 (0.89)	3.84 (0.90)	2.96 (1.12)	3.69 (0.98)	4.20 (0.80)

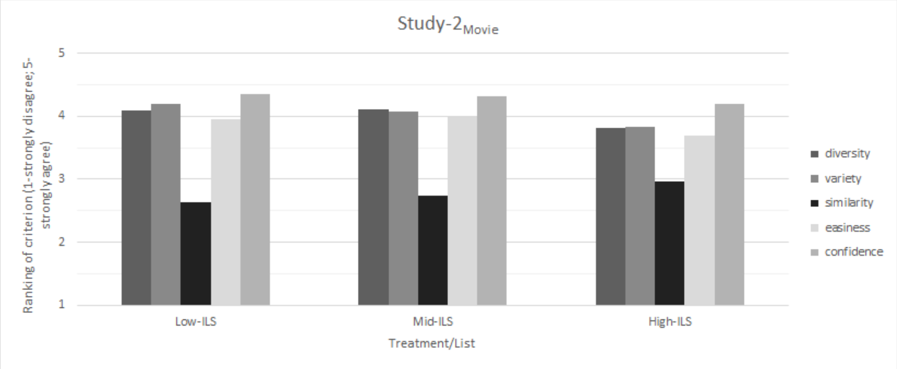

Fig. 4 Average responses in *Study-2*$_{movies}$ (movies, 3 lists)

Table 8 finally shows the outcomes for *Study-2*$_{recipes}$, where the similarity between recipes was based on comparing embeddings of the recipe ingredients and cooking directions[15]. A Kruskal-Wallis test indicated significant differences for *diversity* ($p = .001$), *variety* ($p = .003$), and *similarity* ($p < .001$). Post-hoc tests then revealed significant differences (with $p < .002$ after Bonferroni correction) for all three variables when comparing (a) $rec_{low\text{-}ILS}$ vs. $rec_{high\text{-}ILS}$ and (b) $rec_{mid\text{-}ILS}$ vs. $rec_{high\text{-}ILS}$.[16] This stands in contrast to *Study-1*$_{recipes}$,

[14] The p-values are at about .12 for diversity, .1 for variety, and .29 for similarity.

[15] See Figure 6 in the Appendix for a visual representation of the data

[16] Again, however, the user perceptions in these three dimensions were *not* significant when comparing the $rec_{low\text{-}ILS}$ and $rec_{mid\text{-}ILS}$ conditions, similar to what was observed in *Study-1*$_{movies}$.

where we could not observe statistically significant differences. Again, therefore, these results confirm that the particularities of how the ILS metric is implemented can make a difference.

Table 8 Results for *Study-2$_{recipes}$* (recipes, 3 Lists); numbers show average responses and standard deviations.

List	Q1 Diversity	Q2 Variety	Q3 Similarity	Q4 Choice easiness	Q5 Choice confidence
$rec_{low\text{-}ILS}$	4.02 (0.64)	4.07 (0.78)	3.27 (1.26)	3.73 (0.86)	4.10 (0.62)
$rec_{mid\text{-}ILS}$	4.03 (0.65)	4.07 (0.76)	3.14 (1.23)	3.73 (0.94)	4.15 (0.70)
$rec_{high\text{-}ILS}$	3.46 (1.00)	3.49 (1.05)	4.00 (0.74)	3.95 (0.77)	4.10 (0.63)

Table 9 summarizes our study outcomes in the context of RQ1. Using latent item vectors for similarity assessments worked well for the movie domain but not for recipes. The similarity metrics based on application-specific features, in contrast, only worked well for the recipe domain but not for movies.

Table 9 Summary of main study outcomes (RQ1)

Study	*Similarity based on ...*	*Diversity perceptions significantly different?*
Study-1$_{movies}$	Latent item vectors	Yes
Study-1$_{recipes}$	Latent item vectors	No
Study-2$_{movies}$	Application-specific features	No
Study-2$_{recipes}$	Application-specific features	Yes

Next, we analyzed our subsidiary research questions. First, as the items' popularity could potentially bias the users' diversity perception, we explored if there is a corresponding correlation. Second, we explored potential gender differences in the participants' responses. For instance, Knijnenburg et al (2011) highlight a possible effect of gender on the perception of the quality and variety of recommendations.

To answer the first question, we calculated the popularity level of each list to analyze potential correlations between the popularity of the items in a list and the diversity perception of a list. To this end, we defined the popularity level of a list as the mean of the *number of ratings* of the items in a list; an overview of the popularity levels can be found in Table 21 for the movie domain and Table 22 for the recipe domain, both in the Appendix.

We then computed the Spearman rank correlation coefficient to assess the relationship between list popularity and diversity perception. For both domains, we found the correlation to be *not* statistically significant at the chosen threshold ($\alpha < 0.05$). The p-values for the movie domain were $p = .23$ and for the recipe domain $p = .95$. We therefore have no evidence that the popularity level of a list strongly influences the users' diversity perceptions in our study.

To address the second side question, we investigated if there are gender-related differences in the participants' responses[17] As the prerequisites for a t-test were not fulfilled[18], we used a Wilcoxon test. All results of the Wilcoxon test are summarized in Table 20 in the Appendix. The Wilcoxon test revealed no statistically significant difference in the means for male and female respondents for any of our variables (with $\alpha = .05$). Therefore, we found no indication for significant differences in how different genders perceived the recommendations that were provided to them.

4.2 RQ2: Impact of Item Order on User Perceptions

A previous study (Ge et al, 2012) suggested that the order of the items in a (diversified) list may impact user perceptions. As described above, our study therefore involves permuted versions of the *popsim* and $rec_{mid\text{-}ILS}$ lists in which the similarity between neighboring elements was either minimized or maximized. Minimization leads to clusters of similar items, whereas maximization makes sure that similar items are more dispersed across the list.

The post-hoc analysis of the relevant pairs in the sets $\{popsim, popsim_{min}, popsim_{max}\}$ and $\{rec_{mid\text{-}ILS}, rec_{min}, rec_{max}\}$ revealed *no* significant differences in terms of *diversity*, *variety*, and *similarity*, both in $Study\text{-}1_{movies}$ and in $Study\text{-}1_{recipes}$. All p-values in these pairwise comparison were actually close to 1. Thus, our studies do not provide any evidence that the order of the elements impacted the users' diversity, variety, and list similarity perceptions.

However, we cannot rule out that order effects, as reported in Ge et al (2012), may exist in general. Compared to the study by Ge et al (2012), where participants had to assess the diversity of a list of twelve items, the number of items in the recommendation list was limited to seven items in our case. This may have impacted the outcomes of our study. Also, Ge et al (2012) relied on manually diversified lists (e.g., including several animation movies into a list that was labeled to be action movies), which were probably more "extreme" in terms of the diversity of the list elements. Moreover, in some of their treatments, these diverse elements (e.g., all animation movies in the action movie list) were manually bulked together at the top or bottom of the lists, or manually dispersed across the list. This may have also led to stronger effects of the orderings than we observed in our study. Overall, we therefore believe that more studies are needed to assess the effects of item placements on user perceptions.

Regarding the questions of *choice easiness* and *choice confidence*, *no* significant differences were found—neither in $Study\text{-}1_{movies}$ nor in $Study\text{-}1_{recipes}$, as discussed above. This is in some way aligned with the results of Ge et al

[17] For this analysis, we only considered male and female participants because our sample included only one participant who identified as being non-binary and two who preferred to not to indicate their gender.

[18] Shapiro-Wilk tests indicated that the responses for all variables deviated significantly from a normal distribution

(2012), where no significant impact on user satisfaction was found when the order of the items was changed. From the results shown in Table 5 and in Table 6, we can observe a slight trend in terms of *choice easiness* for the natural recommendation lists. In both studies, choice easiness is consistently higher when similar items are not clustered together, i.e., choices are easier for rec_{min} than for rec_{max}. These differences were however not found to be significant. For the permutations of the *popsim* lists, which consist of rather popular items, this trend was not observed.

As for the last analyses in this context, remember that we asked study participants to rate the similarity of the two most dissimilar items (in terms of the given ILS metric) after they had assessed an entire list of recommendations. This additional measurement serves two purposes. First, it helps us assess if participants were consistent in their assessments. If they are consistent, the similarity assessment for the *extreme* item pair should be lower than the one for the entire list. Second, if this is the case, it would also support that the design principle of the ILS metric, which is based on pairwise comparisons of all items, is valid. If the similarity assessment of the extreme pair would be very similar to the assessment for the entire list, then the similarity perception of the participants might be mainly guided by the extremes and not by the list in its entirety.

As an example for an extreme pair, we asked participants who saw and assessed the $rec_{mid\text{-}ILS}$ list to judge the similarity of the list's extreme items "Avatar" and "Harry Potter and the Order of the Phoenix". In the recipe domain, an example of an extreme pair is "Award Winning Soft Chocolate Chip Cookies" and "Orzo with Parmesan and Basil". Overall, there were four such pairs in the movie domain and four pairs in the recipe domain.[19] The results were consistent for all eight pairs in that the similarity assessments of the individual pairs were lower than those of the lists in which they were contained. The numbers obtained for pairs of movies and the lists that contained these pairs are shown in Table 10. The differences between the pair and list assessments were even stronger for the recipe domain.[20] We assume that this is the case because the provided lists often contained recipes for largely different types of dishes, e.g., main dishes and desserts. Overall, however, we find that (a) the participants are consistent in their assessment and that (b) the users actually consider more than the extreme points in a list when they judge the diversity (or: similarity) of a list.

4.3 RQ3: Criteria that Determine Similarity Assessments

With this research question, our goal is to obtain a better understanding of how users assess the similarities of items in the two examined domains. In a practical application, such an understanding is important to appropriately design

[19] Remember that there were three recommendation-based lists and one consisting of popular movies.

[20] The detailed data for the recipe domain are provided in the Appendix.

Table 10 Assessment of the most dissimilar items and the entire list (i.e., $rec_{mid\text{-}ILS}$) in the movie domain.

Pair assessment			**Containing list**
Movie 1	*Movie 2*	*Sim. (SD)*	*Sim. (SD)*
Avatar	Harry Potter and ...	2.70 (1.20)	2.84 (1.24)
Inception	Lord of the Rings...	2.35 (0.97)	2.84 (1.24)
The Avengers	Monsters Inc.	2.25 (1.23)	2.84 (1.24)
Lord of the Rings ...	2001: A Space Odyssey	2.14 (1.15)	2.84 (1.24)

a suitable similarity function that may then be used in the ILS computation. In our experiment flow described in Figure 2, participants had to accomplish two additional tasks after they had assessed the diversity of a given list and the similarity of the extreme pair in this list. First, they were asked to state, in free text, according to which criteria they have assessed the diversity of the given lists. Second, they were asked to rank up to 7 (movie domain) or 8 (recipe domain) pre-defined ranking criteria.

In a first step, we performed a qualitative analysis of the free text responses. We applied a bottom-up coding approach (Saldana, 2015), where we first identified a larger set of codes/tags in the participants' responses and where we then merged related tags. Overall, 498 words were tagged in the free-text responses in the movie studies. The most frequent aggregated codes for the movie domain are shown in Table 11.[21]

Table 11 Most frequent codes in free-text responses regarding the criteria that determine movie similarities.

Code	*Percentage of appearance*
Genre (including, e.g., *topics*)	28 %
Plot (including, e.g., *story*)	15 %
Theme (including, e.g., *setting*, *time period of movie*)	9 %
Target group (including, e.g., *age*)	6 %
People (including, e.g., *actors*, *characters*)	6 %
Style (including, e.g., *animation*, *visual effects*)	6 %
Production (including, e.g., *director*, *budget*, *awards*)	6 %
Mood (including, e.g., *pacing*, *emotional impact*)	4 %

The genre of the movie was the by far most frequently mentioned decision criterion by the participants, and it was mentioned almost twice as often as the second-ranked aspect, the plot. Interestingly, aspects related to movie poster, the title, music and sound effects or production year were only very rarely mentioned.

The pre-defined list of criteria which participants had to rank after their free-text responses included a number of meta-data elements that are commonly found in research datasets in the movie domain. To aggregate the ob-

[21] The aggregated codes for the recipe domain can be found in the Appendix.

tained incomplete rankings, we used the Borda count rank aggregation method Black (1958). In this method, an item at the first place of a list of n elements receives $n-1$ points, the second place gives $n-2$ points, etc. The obtained ranking is shown in Table 12.

We observe that the top two positions (genre, plot) are identical to those extracted from the free-text responses by the participants. Several of the subsequent codes from the free-text input (e.g., theme, target group, mood) were not part of our pre-defined list, which is based on common meta-data from movie rating datasets. This suggests that it might be helpful to include alternative information sources when computing similarities in this domain. In particular, user provided content in the form of *tags* appeared to be promising in the past (Vig et al, 2009; Trattner and Jannach, 2019).

Table 12 Borda count pre-defined criteria for assessing *movie* similarities.

Criterion	*Borda count*
Genre	3800
Plot	3437
Actors	3261
Title	3033
Image	2920
Release Year	2871
Runtime	2652

Interestingly, while the genre was considered as the main criterion to assess item similarity in the movie domain, building the ILS metric on genre information *alone* may be too limited, as indicated by the findings from *Study-2*$_{movies}$. In our genre-based approach, we used the Jaccard index as a distance measure. While we cannot rule out other ways of computing similarities based on genres, the study by Trattner and Jannach (2019) suggested that using an LDA-based comparison may not be more effective either.

The Borda count ranking of the pre-defined criteria in the recipe domain is shown in Table 13[22]. The most important criterion, according to the participants, are the ingredients, which are also one of the two pieces of information that we used in our application-specific similarity measure in the recipe domain. This is also in line with the observations in Trattner and Jannach (2019) where ingredients (as well as the image of the dish) were highly ranked. Note however that Trattner and Jannach (2019) found that cooking directions were also very highly correlated with human similarity perceptions and that images—although they were self-reported by participants to be very decisive–did not correlate to a similar extent with the perceptions in reality. Therefore, we relied both on ingredients and directions when designing a similarity function in our study, which in the end turned out to be indicative of the similarity as perceived by users.

[22] The absolute values are higher in the recipe domain compared to the movie domain because there were more pre-defined criteria.

Table 13 Borda count pre-defined criteria for assessing *recipe* similarities.

Criterion	*Borda count*
Ingredients	7185
Cuisine	6447
Image	6364
Nutrients	5778
Instructions	5650
Cooking duration	5520
Reviews	4937

Overall, we argue that it is important to first investigate, for each application domain, what the particular factors are that determine the diversity perception of users. In the case of recipes, we built in prior research by Trattner and Jannach (2019) and combined those features that were found to be good predictors of perceived similarities in a large-scale user study. For the movie domain, in contrast, we relied on genres as the only feature in the similarity function, as this is a common approach in the literature. Considering the importance ranking in Table 12, it seems advisable to combine genre information with other features, in particular with plot summaries. An investigation of which combination of features is most predictive of the diversity perception by users in the movie domain is however beyond the scope of our present work, which aims to validate typical similarity measures from the literature.

Lastly, we also investigated the effects of domain knowledge on the diversity perception of a list. Based on prior research conducted by Porcaro et al (2022) we might assume that participants with differing levels of background knowledge may perceive a list of items differently. On the one hand, users with more background knowledge might find more aspects in which items are different. On the other hand, if users know nothing about, for example, a movie, they might think all movies of that same genre are similar at a higher level.

To study if such effects exists, we analyzed if there is a correlation between the expressed diversity perception of participants and their level of familiarity with the items in a list. Technically, to determine the familiarity level of a list, we counted for how many items in the list the participant expressed at least some level of familiarity, i.e., when the item was not entirely unknown. Interestingly, the analysis indicated no correlation between the diversity perception of a list and the participants' familiarity with the items ($r < 0.1$).

5 Conclusions and Future Work

Overall, our research indicates that the ILS measure can indeed be a good proxy for human diversity perceptions. The details of how the metric is actually implemented can however matter, and a specific metric therefore has to be validated in a given domain and application, e.g., before algorithmic diversification techniques are applied. With these results, our work narrows an important research gap in the literature, where researchers often implicitly assume, without validation, that their used ILS metric—which is likely designed based only on intuition—would be a reliable proxy for human diversity perceptions.

One limitation of our research, so far, is that we cannot provide general guidelines regarding how to design a similarity metric for a given domain. Our findings seem to suggest that in the case of meta-data based approaches, it may be helpful to rely on more than one item feature (e.g., genre and plot or ingredients and cooking directions). As our studies were limited to two domains and on two particular ways of creating similarity functions based on meta-data, this indication is not too strong yet and needs more research in the future.

A potential threat to the validity of our studies may be seen in the representatives and the reliability of the participants. In particular, in the movie domain, many participants declared themselves to be almost movie enthusiasts. It therefore remains open to investigation in the future if the findings obtained in our present study would generalize to a participant population that is less engaged in the application domain. Regarding the reliability of the participants, we implemented a number of measures to ensure that we can trust that the responses are reliable, e.g., by only admitting crowdworkers with good past performance and by implementing attention checks in the online studies.

In terms of future works, it may be interesting to further investigate the effects of gender on the diversity perception of recommendation lists in more depth. While our work provided no evidence for differences in how different genders perceive the recommendation lists, other works suggest that there, in fact, may be a difference (Knijnenburg et al, 2011). Furthermore, it is important to expand our research to other frequently used computational metrics, which have not been validated to a sufficient extent yet. First of all, this includes common “beyond-accuracy” quality metrics, in particular *novelty* or *serendipity*. For instance, like for diversity, a number of research works introduce novelty metrics. Often, such novelty metrics are based on item popularity (Vargas and Castells, 2011). However, for most of these metrics, it remains to be shown that they are suitable proxies for user perceptions. Different from diversity—as investigated in our present work—the assessment of the novelty of a set of recommended items has to be done relative to what a particular user already knows. This adds additional complexity to the design of corresponding user studies.

In the last few years, also questions of *fairness* have gained increased research interest in the area of recommender systems and in machine learning in general. In that context, it was recently observed by Deldjoo et al (2022) that the majority of currently published research is based on non-validated assumptions and metrics. For many of these works, it therefore remains open that the described computational fairness goals correspond to what users would consider fair or unfair.

Acknowledgments

We thank Martijn C. Willemsen for his feedback on the statistical analyses.

References

Abdollahpouri H, Burke R, Mobasher B (2019) Managing popularity bias in recommender systems with personalized re-ranking. ArXiv pp 1–6, URL https://doi.org/abs/1901.07555

Adomavicius G, Kwon Y (2012) Improving aggregate recommendation diversity using ranking-based techniques. IEEE Transactions on Knowledge and Data Engineering 24(5):896–911, URL https://doi.org/10.1109/TKDE.2011.15

Black D (1958) The Theory of Committees and Elections. Springer

Bradley K, Smyth B (2001) Improving recommendation diversity. In: Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, pp 85–94

Brovman YM, Jacob M, Srinivasan N, Neola S, Galron D, Snyder R, Wang P (2016) Optimizing similar item recommendations in a semi-structured marketplace to maximize conversion. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp 199–202, URL https://doi.org/10.1145/2959100.2959166

Chen L, Wu W, He L (2013) How personality influences users' needs for recommendation diversity? In: CHI '13 Extended Abstracts on Human Factors in Computing Systems, pp 829–834, URL https://doi.org/10.1145/2468356.2468505

Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 659–666, URL https://doi.org/10.1145/1390334.1390446

Colucci L, Doshi P, Lee KL, Liang J, Lin Y, Vashishtha I, Zhang J, Jude A (2016) Evaluating item-item similarity algorithms for movies. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp 2141–2147, URL https://doi.org/10.1145/2851581.2892362

Deldjoo Y, Jannach D, Bellogin A, Difonzo A, Zanzonelli D (2022) A survey of research on fair recommender systems. URL https://doi.org/10.48550/ARXIV.2205.11127

Downie JS, Lee JH, Gruzd AA, Jones MC (2007) Toward an understanding of similarity judgments for music digital library evaluation. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, pp 307–308, URL https://doi.org/10.1145/1255175.1255235

Du Y, Ranwez S, Sutton-Charani N, Ranwez V (2021) Is diversity optimization always suitable? Toward a better understanding of diversity within recommendation approaches. Information Processing & Management 58(6):102721, URL https://doi.org/10.1016/j.ipm.2021.102721

Ekstrand MD, Harper FM, Willemsen MC, Konstan JA (2014) User perception of differences in recommender algorithms. In: Proceedings of the 8th ACM Conference on Recommender Systems, pp 161–168, URL https://doi.org/10.1145/2645710.2645737

Ellis DP, Whitman B, Berenzweig A, Lawrence S (2002) The quest for ground truth in musical artist similarity. In: Proceedings of the 3rd International Conference on Music Information Retrieval, URL https://ismir2002.ismir.net/proceedings/02-FP05-4.pdf

Fleder DM, Hosanagar K (2007) Recommender systems and their impact on sales diversity. In: Proceedings of the 8th ACM Conference on Electronic Commerce, p 192–199, URL https://doi.org/10.1145/1250910.1250939

Ge M, Gedikli F, Jannach D (2011) Placing high-diversity items in top-n recommendation lists. In: Proceedings of the Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP 2011 at IJCAI 2011)

Ge M, Jannach D, Gedikli F, Hepp M (2012) Effects of the placement of diverse items in recommendation lists. In: 14th International Conference on Enterprise Information Systems, pp 201–208, URL https://doi.org/10.5220/0003974802010208

de Gemmis M, Lops P, Musto C, Narducci F, Semeraro G (2015) Semantics-aware content-based recommender systems. In: Recommender Systems Handbook, pp 119–159, URL https://doi.org/10.1007/978-1-4899-7637-6_4

Hauptmann H, Leipold N, Madenach M, Wintergerst M, Lurz M, Groh G, Böhm M, Gedrich K, Krcmar H (2021) Effects and challenges of using a nutrition assistance system: results of a long-term mixed-method study. User Modeling and User-Adapted Interaction URL https://doi.org/10.1007/s11257-021-09301-y

Jannach D (2022) Multi-objective recommendation: Overview and challenges. In: Proceedings of the 2nd Workshop on Multi-Objective Recommender Systems co-located with 16th ACM Conference on Recommender Systems (RecSys 2022), URL https://arxiv.org/abs/2210.10309

Jannach D, Lerche L, Kamehkhosh I, Jugovac M (2015) What recommenders recommend: an analysis of recommendation biases and possible countermeasures. User Modeling and User-Adapted Interaction 25(5):427–491, URL https://doi.org/10.1007/s11257-015-9165-3

Jannach D, Kamehkhosh I, Lerche L (2017) Leveraging multi-dimensional user models for personalized next-track music recommendation. In: Proceedings of the Symposium on Applied Computing, p 1635–1642, URL https://doi.org/10.1145/3019612.3019756

Jensen O, Lisman JE (1996) Novel lists of 7+/-2 known items can be reliably stored in an oscillatory short-term memory network: interaction with long-term memory. Learning & Memory 3(2-3):257–263, URL https://doi.org/10.1101/lm.3.2-3.257

Kaminskas M, Bridge D (2016) Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Trans Interact Intell Syst 7(1), URL https://doi.org/10.1145/2926720

Knijnenburg BP, Willemsen MC, Kobsa A (2011) A pragmatic procedure to support the user-centric evaluation of recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems, p 321–324, URL https://doi.org/10.1145/2043932.2043993

Kunaver M, Požrl T (2017) Diversity in recommender systems – a survey. Knowledge-Based Systems 123:154–162, URL https://doi.org/10.1016/j.knosys.2017.02.009

Lee JH (2010) Crowdsourcing music similarity judgments using mechanical turk. In: Proceedings of the 11th International Society for Music Information Retrieval Conference, pp 183–188

Lin K, Sonboli N, Mobasher B, Burke R (2020) Calibration in collaborative filtering recommender systems: A user-centered analysis. In: Proceedings of the 31st ACM Conference on Hypertext and Social Media, pp 197–206, URL https://doi.org/10.1145/3372923.3404793

Mauro N, Ardissono L (2019) Extending a tag-based collaborative recommender with co-occurring information interests. In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, pp 181–190, URL https://doi.org/10.1145/3320435.3320458

McGinty L, Smyth B (2003) On the role of diversity in conversational recommender systems. In: Case-Based Reasoning Research and Development, pp 276–290, URL https://doi.org/10.1007/3-540-45006-8_23

McNee SM, Riedl J, Konstan JA (2006) Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: CHI '06 Extended Abstracts on Human Factors in Computing Systems, pp 1097–1101, URL https://doi.org/10.1145/1125451.1125659

Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review 63(2):81–97, URL https://doi.org/10.1037/h0043158

Nilashi M, Jannach D, bin Ibrahim O, Esfahani MD, Ahmadi H (2016) Recommendation quality, transparency, and website quality for trust-building in recommendation agents. Electronic Commerce Research and Applications 19:70–84, URL https://doi.org/10.1016/j.elerap.2016.09.003

van Pinxteren Y, Geleijnse G, Kamsteeg P (2011) Deriving a recipe similarity measure for recommending healthful meals. In: Proceedings of the 16th International Conference on Intelligent User Interfaces, pp 105–114, URL https://doi.org/10.1145/1943403.1943422

Porcaro L, Gómez E, Castillo C (2022) Perceptions of diversity in electronic music: The impact of listener, artist, and track characteristics. Proceedings of the ACM on Human-Computer Interaction 6, URL https://doi.org/10.1145/3512956

Pu P, Chen L, Hu R (2011) A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems, pp 157–164, URL https://doi.org/10.1145/2043932.2043962

Rendle S, Krichene W, Zhang L, Anderson J (2020) Neural collaborative filtering vs. matrix factorization revisited. In: Proceedings of the 14th ACM Conference on Recommender Systems, RecSys '20, p 240–248, URL https://doi.org/10.1145/3383313.3412488

Ribeiro MT, Ziviani N, Moura ESD, Hata I, Lacerda A, Veloso A (2015) Multiobjective pareto-efficient approaches for recommender systems. ACM Trans Intell Syst Technol 5(4), URL https://doi.org/10.1145/2629350

Saldana J (2015) The Coding Manual for Qualitative Researchers, 3rd edn. Sage Publications

Shi Y, Zhao X, Wang J, Larson M, Hanjalic A (2012) Adaptive diversification of recommendation results via latent factor portfolio. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, p 175–184, URL https://doi.org/10.1145/2348283.2348310

Starke AD, Øverhaug S, Trattner C (2021) Predicting feature-based similarity in the news domain using human judgments. In: Proceedings of the 9th International Workshop on News Recommendation and Analytics

Trattner C, Jannach D (2019) Learning to recommend similar items from human judgements. User Modeling and User-Adapted Interaction 30:1–49, URL https://doi.org/10.1007/s11257-019-09245-4

Tsai CH, Brusilovsky P (2018) Beyond the ranked list: User-driven exploration and diversification of social recommendation. In: 23rd International Conference on Intelligent User Interfaces, pp 239–250, URL https://doi.org/10.1145/3172944.3172959

Vargas S (2011) New approaches to diversity and novelty in recommender systems. In: Fourth BCS-IRSG Symposium on Future Directions in Information Access, p 8–13, URL https://doi.org/10.5555/2227322.2227324

Vargas S, Castells P (2011) Rank and relevance in novelty and diversity metrics for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems, pp 109–116, URL https://doi.org/10.1145/2043932.2043955

Vargas S, Castells P, Vallet D (2012) Explicit relevance models in intent-oriented information retrieval diversification. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 75–84, URL https://doi.org/10.1145/2348283.2348297

Vargas S, Baltrunas L, Karatzoglou A, Castells P (2014) Coverage, redundancy and size-awareness in genre diversity for recommender systems. In: Proceedings of the 8th ACM Conference on Recommender Systems, pp 209–216, URL https://doi.org/10.1145/2645710.2645743

Vig J, Sen S, Riedl J (2009) Tagsplanations: Explaining recommendations using tags. In: Proceedings of the 14th International Conference on Intelligent User Interfaces, p 47–56, URL https://doi.org/10.1145/1502650.1502661

Wang C, Agrawal A, Li X, Makkad T, Veljee E, Mengshoel O, Jude A (2017) Content-based top-n recommendations with perceived similarity. In: Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics, pp 1052–1057, URL https://doi.org/10.1109/SMC.2017.8122750

Willemsen MC, Graus MP, Knijnenburg BP (2016) Understanding the role of latent feature diversification on choice difficulty and satisfaction. User Modeling and User-Adapted Interaction 26(4):347–389, URL https://doi.org/10.1007/s11257-016-9178-6

Yao Y, Harper FM (2018) Judging similarity: A user-centric study of related item recommendations. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp 288–296, URL https://doi.org/10.1145/3240323.3240351

Zeng Z, Lin J, Li L, Pan W, Ming Z (2019) Next-item recommendation via collaborative filtering with bidirectional item similarity. ACM Transactions on Information Systems 38(1), URL https://doi.org/10.1145/3366172

Zheng Y, Wang DX (2022) A survey of recommender systems with multi-objective optimization. Neurocomputing 474:141–153, URL https://doi.org/10.1016/j.neucom.2021.11.041

Ziegler CN, McNee SM, Konstan JA, Lausen G (2005) Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web, pp 22–32, URL https://doi.org/10.1145/1060745.1060754

Mathias Jesse Mathias Jesse is a doctoral candidate in Computer Science in the Digital Age Research Center (D!ARC) at University of Klagenfurt, Austria. He received his MSc degree from the same university in Information Management. His research is focused on persuasive recommender systems.

Christine Bauer is an Assistant Professor at Utrecht University, The Netherlands. Her research activities center on interactive intelligent systems. Central themes in her research are context and context-adaptivity. Core interests in her current research activities are fairness and multi-method evaluations. Christine has co-authored more than 100 publications, three of them awarded as best research paper and one received an award of excellence. She holds five awards as best or outstanding reviewer.

Dietmar Jannach is a Professor of Computer Science at University of Klagenfurt, Austria. He has worked on different areas of artificial intelligence, including recommender systems, model-based diagnosis, and knowledge-based systems. He is the leading author of a textbook on recommender systems and has authored more than hundred research papers, focusing on the application of artificial intelligence technology to practical problems.

A Additional Material

In this appendix, we provide additional detailed material, including the questionnaire items for the recipe domain (Table 14), the participants' assessments of the most dissimilar items in the recipe domain (Table 15), coding results for the recipe domain (Table 16), the results of the statistical significance tests (pairwise comparisons) for the movie study in Phase 1 (Table 17 to Table 19), the results on the gender differences in all studies (Table 20), the popularity levels of items for both domains (Table 21 and Table 22), and the distribution of human responses for all studies (Table 23 to Table 26). Furthermore, we provide additional figures for the average responses of *Study-1*$_{recipes}$ (Table 5 and *Study-2*$_{recipes}$ (Table 6).

Table 14 Questionnaire items for each list in the recipe domains. The options for Q6 were "more than one good option", "exactly one good option", "no option that I liked".

Question		Focus	Type
Q1 (diversity)	The recipes presented in this list are diverse	Diversity	5-point Likert scale
Q2 (variety)	The recipes presented in this list offer a rich variety	Diversity	5-point Likert scale
Q3 (similarity)	The recipes presented in this list are similar to each other	Diversity	5-point Likert scale
Q4 (choice easiness)	Selecting a recipe was easy	Choice Easiness	5-point Likert scale
Q5 (choice confidence)	I am confident I will like the recipe I selected to watch next	Choice Confidence	5-point Likert scale
Q6 (no. of good options)	The list contained . . .	No. of good options	Predefined answers

Table 15 Assessment of most dissimilar recipes and the entire list (i.e., *rec*$_{mid\text{-}ILS}$).

Pair assessment			**Containing list**
Recipe 1	*Recipe 2*	*Sim. (SD)*	*Sim. (SD)*
Award Winning Soft Chocolate Chip Cookies	Orzo with Parmesan and Basil	2.00 (1.46)	2.53 (1.21)
Banana Crumb Muffins	Broiled Tilapia Parmesan	1.77 (1.45)	2.53 (1.21)
Delicious Ham and Potato Soup	Broiled Tilapia Parmesan	2.31 (1.23)	2.53 (1.21)
Fluffy Pancakes	Cranberry Pistachio Biscotti	2.26 (1.16)	2.53 (1.21)

Table 16 Most frequent codes in free-text responses regarding the criteria that determine recipe similarities.

Code	*Percentage of appearance*
Ingredients (including, e.g., *main ingredient*)	29 %
Category (including, e.g., *course*, *type*)	27 %
Flavor (including, e.g., *taste*, *spiciness*)	11 %
Cooking process (including, e.g., *directions*, *duration*)	10 %
Cuisine (including, e.g., *origin*)	9 %
Familiarity (including, e.g., *knowledge*)	5 %
Nutrition (including, e.g., *proteins*, *healthiness*)	4 %
First impression (including, e.g., *image*, *title*)	4 %

Table 17 Pairwise Wilcoxon test with Bonferroni correction for *Study-1*$_{movies}$ (movies, 9 lists); numbers show the p-value for each pairwise comparison on *diversity*. Significant results ($p < .05$) are printed in bold face.

List	*upp*	$rec_{low\text{-}ILS}$	$rec_{mid\text{-}ILS}$	$rec_{high\text{-}ILS}$	*popsim*	rec_{min}	rec_{max}	$popsim_{min}$
$rec_{low\text{-}ILS}$	**< .001**							
$rec_{mid\text{-}ILS}$	**< .001**	1						
$rec_{high\text{-}ILS}$	**< .01**	**< .001**	**< .001**					
popsim	**< .001**	.826	1	**< .01**				
rec_{min}	**< .001**	.88	1	**.001**	1			
rec_{max}	**< .001**	1	1	**< .001**	1	1		
$popsim_{min}$	**< .001**	**.02**	.31	**.046**	1	1	.21	
$popsim_{max}$	**< .001**	**.048**	.54	**.04**	1	1	.40	1

Table 18 Pairwise Wilcoxon test with Bonferroni correction for *Study-1*$_{movies}$ (movies, 9 lists); numbers show the p-value for each pairwise comparison on *variety*. Significant results ($p < .05$) are printed in bold face.

List	*upp*	$rec_{low\text{-}ILS}$	$rec_{mid\text{-}ILS}$	$rec_{high\text{-}ILS}$	*popsim*	rec_{min}	rec_{max}	$popsim_{min}$
$rec_{low\text{-}ILS}$	**< .001**							
$rec_{mid\text{-}ILS}$	**< .001**	1						
$rec_{high\text{-}ILS}$	.149	**< .001**	**.002**					
popsim	**< .001**	.43	1	.14				
rec_{min}	**< .001**	.44	1	**.003**	1			
rec_{max}	**< .001**	1	1	**< .001**	1	1		
$popsim_{min}$	**< .001**	**.008**	1	.33	1	1	1	
$popsim_{max}$	**< .001**	**.033**	1	.12	1	1	1	1

Table 19 Pairwise Wilcoxon test with Bonferroni correction for *Study-1*$_{movies}$ (movies, 9 lists); numbers show the p-value for each pairwise comparison on *similarity*. Significant results ($p < .05$) are printed in bold face.

List	*upp*	$rec_{low\text{-}ILS}$	$rec_{mid\text{-}ILS}$	$rec_{high\text{-}ILS}$	*popsim*	rec_{min}	rec_{max}	$popsim_{min}$
$rec_{low\text{-}ILS}$	**< .001**							
$rec_{mid\text{-}ILS}$	**< .001**	.23						
$rec_{high\text{-}ILS}$	**< .001**	**< .001**	**< .001**					
popsim	**< .001**	**.003**	1	.44				
rec_{min}	**< .001**	**.03**	1	**.01**	1			
rec_{max}	**< .001**	1	1	**< .001**	1	1		
$popsim_{min}$	**< .001**	**.004**	1	**.02**	1	1	1	
$popsim_{max}$	**< .001**	**.02**	1	**< .001**	1	1	1	1

Table 20 Pairwise Wilcoxon test with Bonferroni correction for all studies with $\alpha = .05$; numbers show the p-value for each pairwise comparison on the gender.

Study	Diversity	Variety	Similarity	Choice easiness	Choice confidence
Study-1$_{movies}$	.08	.22	.07	.90	.73
Study-1$_{recipes}$	.86	.12	.56	.08	.21
Study-2$_{movies}$	.34	.29	.46	.10	.29
Study-2$_{recipes}$	.24	.43	.87	.53	.11

Table 21 Average number of ratings for the lists in *Study-1*$_{movies}$ and the corresponding diversity perception. We used the mean of *popsim*, $popsim_{max}$, and $popsim_{min}$ for the final diversity perception rating. The same was done for $rec_{mid\text{-}ILS}$.

List	Average number of ratings for list	Diversity perception
upp	15978	2.37
popsim	45174	3.82
$rec_{low\text{-}ILS}$	26939	4.05
$rec_{mid\text{-}ILS}$	29984	3.92
$rec_{high\text{-}ILS}$	46622	3.17

Table 22 Average number of ratings for the lists in *Study-1*$_{recipes}$ and the corresponding diversity perception. We used the mean of *popsim*, $popsim_{max}$, and $popsim_{min}$ for the final diversity perception rating. The same was done for $rec_{mid\text{-}ILS}$.

List	Average number of ratings for list	Diversity perception
upp	106	2.37
popsim	58	3.82
$rec_{low\text{-}ILS}$	5183	4.31
$rec_{mid\text{-}ILS}$	6769	4.12
$rec_{high\text{-}ILS}$	5385	4.22

Table 23 Distribution of human responses for each list in *Study-1*$_{movies}$. Each participant received three different lists, hence the different number of responses per list; the responses are split by gender.

List	Male	Female	Other (non-binary/undisclosed)
upp	39	36	0
$rec_{low\text{-}ILS}$	43	29	0
$rec_{mid\text{-}ILS}$	47	27	0
$rec_{high\text{-}ILS}$	50	26	0
popsim	44	27	0
rec_{min}	38	30	0
rec_{max}	47	34	0
$popsim_{min}$	39	37	0
$popsim_{max}$	49	27	0

Table 24 Distribution of human responses for each list in *Study-1*$_{recipes}$. Each participant received three different lists, hence the different number of responses per list; the responses are split by gender.

List	Male	Female	Other (non-binary/undisclosed)
upp	33	24	2
$rec_{low\text{-}ILS}$	38	30	0
$rec_{mid\text{-}ILS}$	32	28	0
$rec_{high\text{-}ILS}$	44	24	0
popsim	40	24	1
rec_{min}	39	22	1
rec_{max}	44	28	1
$popsim_{min}$	39	32	0
$popsim_{max}$	36	22	1

Table 25 Distribution of human responses for each list in *Study-2*$_{movies}$. Each participant received three different lists, hence all lists have the same amount of responses; the responses are split by gender.

List	Male	Female	Other (non-binary/undisclosed)
$rec_{low\text{-}ILS}$	33	21	1
$rec_{mid\text{-}ILS}$	33	21	1
$rec_{high\text{-}ILS}$	33	21	1

Table 26 Distribution of human responses for each list in *Study-2*$_{recipes}$. Each participant received three different lists, hence all lists have the same amount of responses; the responses are split by gender.

List	Male	Female	Other (non-binary/undisclosed)
$rec_{low\text{-}ILS}$	38	20	0
$rec_{mid\text{-}ILS}$	38	20	0
$rec_{high\text{-}ILS}$	38	20	0

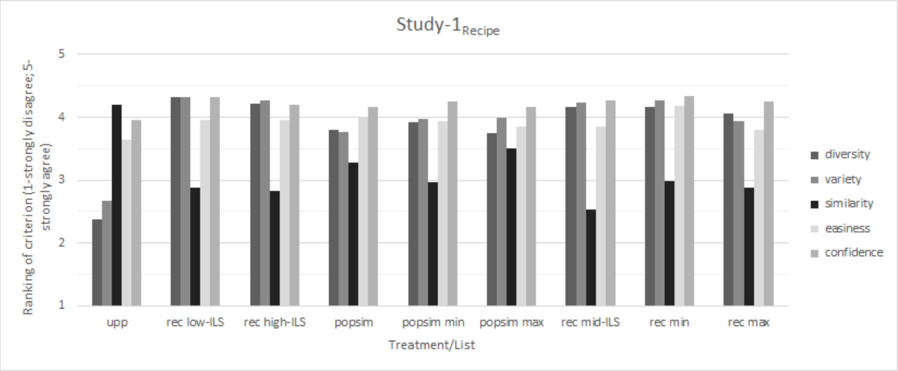

Fig. 5 Average responses in *Study-1$_{recipes}$* (recipes, 9 lists)

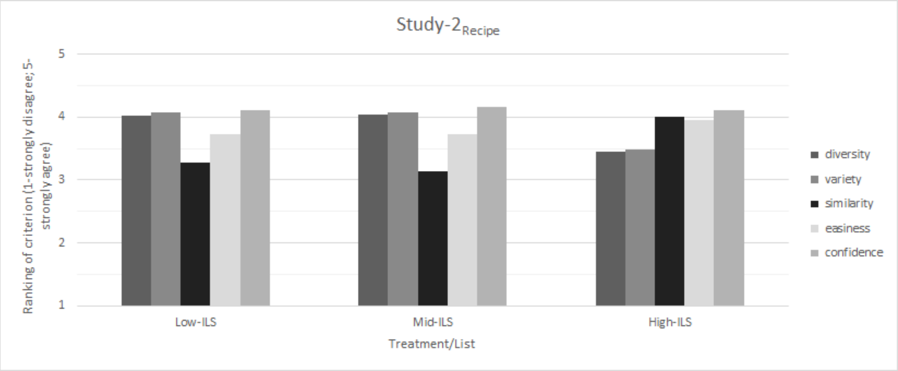

Fig. 6 Average responses in *Study-2$_{recipes}$* (recipes, 3 lists)