Understanding the role of latent feature diversification on choice difficulty and satisfaction

Martijn C. Willemsen¹ · Mark P. Graus² ·
Bart P. Knijnenburg³

Received: 14 March 2014 / Accepted in revised form: 8 August 2016 /

Published online: 12 September 2016

© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract People like variety and often prefer to choose from large item sets. However, large sets can cause a phenomenon called “choice overload”: they are more difficult to choose from, and as a result decision makers are less satisfied with their choices. It has been argued that choice overload occurs because large sets contain more similar items. To overcome this effect, the present paper proposes that increasing the diversity of item sets might make them more attractive and satisfactory, without making them much more difficult to choose from. To this purpose, by using structural equation model methodology, we study diversification based on the latent features of a matrix factorization recommender model. Study 1 diversifies a set of recommended items while controlling for the overall quality of the set, and tests it in two online user experiments with a movie recommender system. Study 1a tests the effectiveness of the latent feature diversification, and shows that diversification increases the perceived diversity and attractiveness of the item set, while at the same time reducing the perceived difficulty of choosing from the set. Study 1b subsequently shows that diversification can increase users’ satisfaction with the chosen option, especially when they are choosing from small, diverse item sets. Study 2 extends these results by testing our diversification algorithm against traditional Top-N recommendations, and finds that diverse,

The acting editor for the review of this article has been Dietmar Jannach, Technical University of Dortmund, Germany.

✉ Martijn C. Willemsen
M.C.Willemsen@tue.nl

¹ Department IE&IS, Human-Technology Interaction Group, Eindhoven University of Technology, IPO 0.17, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

² Department IE&IS, Human-Technology Interaction Group, Eindhoven University of Technology, IPO 0.20, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

³ School of Computing, Clemson University, 215 McAdams Hall, Clemson, SC 29631, USA

small item sets are just as satisfying and less effortful to choose from than Top-N recommendations. Our results suggest that, at least for the movie domain, diverse small sets may be the best thing one could offer a user of a recommender system.

Keywords Recommender systems · Diversification · Matrix factorization · User experience · User-centric evaluation · Decision-making · Choice overload · Human-computer interaction

1 Introduction

Every day we are confronted with an abundance of decisions. The large assortments found in supermarkets and online stores ensure that all tastes are catered for, allowing each individual to maximize his or her utility. However, some people argue that these large assortments result in a Tyranny of Choice ([Schwartz 2004](#)) and that in the end they do not contribute to our overall happiness. Indeed, research in consumer and social psychology has shown that there are major drawbacks in offering large lists of attractive items, especially when customers do not have strong preferences, resulting in what has been labeled Choice Overload ([Iyengar and Lepper 2000](#); [Chernev 2003](#); [Scheibehenne et al. 2010](#)).

[Scheibehenne et al. \(2010\)](#) indicate that item similarity constitutes an important potential moderator of choice overload: Choice overload is more likely to occur for item sets with many items that are equally attractive. This might especially apply to recommender systems, which explicitly try to provide the user with highly attractive recommendations ([Bollen et al. 2010](#)). Large item sets are often inherently less varied, because most real-life assortments have limits to the potential variety they can offer. This, too, makes choosing from large item sets more difficult.

These results suggest that the diversity of the set of items under consideration is an important determinant of choice difficulty and satisfaction, which might have important implications in the domain of recommender systems ([Bollen et al. 2010](#)). In pioneering work on case-based recommender systems, [Smyth and McClave \(2001\)](#) already argued that there is a tradeoff between attaining the highest accuracy and providing sufficient diversity. Similarly, [Ziegler et al. \(2005\)](#) showed that topic diversification can enhance users' satisfaction with a list of book recommendations. [Knijnenburg et al. \(2012, Sect. 4.8\)](#) reproduced [Ziegler et al.](#)'s findings in a movie recommender and, using their user-experience framework, showed that the positive effect of diversification on user satisfaction could be explained by an increase in perceived recommendation quality, a decrease in choice difficulty, and an increase in perceived system effectiveness.

In the present paper, we will expand on this earlier work by applying insights from the psychological literature on choice overload to the study of diversification in recommender systems. Specifically, the net result of choice overload is typically reduced choice satisfaction, because of two counteracting forces: larger items sets increase the attractiveness but also increase the difficulty of choosing from the set. Arguably, this problem can be solved by increasing the diversity of the set, all else remaining equal. Therefore, we will discuss and test *latent feature diversification*, a

diversification method that integrates insights from psychological research into the core mechanism of a matrix factorization recommendation algorithm. Because latent feature diversification provides maximum control over item quality and item set variety on an individual level it can increase the diversity (and thus reduce the choice difficulty) of an item set while maintaining perceived attractiveness and satisfaction. We will test this algorithm in two studies, for a total of three user experiments, employing the user-centric framework of [Knijnenburg et al. \(2012\)](#)

Our primary research question asks whether latent feature diversification applied in a recommender system can help users to find appropriate items in a satisfactory way, with minimal choice difficulty. Diversification might reduce the average predicted quality of recommendation lists compared to standard Top-N lists, but we expect that the increased diversity might still result in higher satisfaction because of the reduced difficulty. We also consider the length of the recommendation list, and investigate the effect of list length as well as the interplay of diversification and list length on choice satisfaction. Finally, the user-centric framework allows us to investigate both how and why diversification influences choice difficulty and satisfaction, by providing better insights into the psychological mechanisms driving these effects.

2 Theory and existing work

Choice overload as described in the literature ([Iyengar and Lepper 2000](#); [Scheibehenne et al. 2010](#)) is the phenomenon of choice difficulty reducing choice satisfaction because the difficulty of choosing from a larger set of options outweighs the increased attractiveness of the larger set. In the current paper we focus not so much on the phenomenon of choice overload but on its underlying mechanisms: how do choice difficulty and perceived attractiveness influence satisfaction and what is the role of diversity? We will first discuss existing literature on attractiveness and choice difficulty (Sect. 2.1) and diversification (Sect. 2.2) before introducing our latent feature diversification method in Sect. 2.3 and an overview of the studies in Sect. 2.4.

2.1 Attractiveness and choice difficulty in large item sets

In the psychological literature it is argued that larger assortments offer important advantages for consumers ([Chernev 2003](#); [Scheibehenne et al. 2010](#)). They can serve a larger target group, satisfying the needs of a diverse set of customers. They also allow customers to seek variety and maintain flexibility during the purchase process ([Simonson 1990](#)). Moreover, the variety of products offered by a specific brand is perceived as a cue for the quality of the brand ([Berger et al. 2007](#)). In other words, larger item sets are often perceived to be more attractive.

However, these beneficial aspects of large sets only seem to occur when people have stable and explicit preferences. [Chernev \(2003\)](#) labels this as having an “ideal point”: Only when people hold clear preferences for particular attribute values (an “ideal point”) they can assign specific weights to the different attributes that define a choice. Indeed, under such circumstances the likelihood of finding an option that matches one’s preferences is greater for larger item sets, because larger sets often contain a

larger variety of attributes and span a wider range of attribute values. However, when people lack knowledge of the underlying attribute structures of a decision domain, they instead construct their preferences on the fly, i.e., while making the decision (Bettman et al. 1998). In such cases they are much less capable of articulating their preferences, and a larger item set might thus only hinder their decision process by making the preference construction process more difficult. In a series of studies Chernev (2003) showed that people without ideal points are indeed more likely to change their minds after a decision (i.e. showed less confidence in their decision) when confronted with a large assortment compared to a small assortment.

Recommender systems could potentially reduce this choice difficulty as such systems help users to find a limited set of items that fit their personal preferences from a much larger set of potential alternatives. However, choice overload effects also occur in item lists generated by recommender systems. Bollen et al. (2010) performed a user experiment with a matrix factorization movie recommender, comparing three conditions: a small Top-5 list, a large Top-20 list, and a large lower quality 20-item list, composed of the Top-5 plus 15 lower-ranked movies. Users experienced significantly more choice difficulty when presented with the high quality Top-20 item list, compared to the other two lists. The increased difficulty counteracted the increased perceived attractiveness of the Top-20 list, showing that in the end, choice satisfaction in all three conditions was about the same. Behavioral data corroborated these findings, as users spent more effort evaluating the items of the Top-20 list compared to the other two lists.

Effort may provide one possible explanation for the increased choice difficulty of larger item sets. The amount of presented information and the required number of comparisons is larger for larger sets (the number of pairs of items to compare increases with $O(n^2)$). Therefore, more effort will be required to construct a preference when the number of items increases, especially for cases where the decision maker has no ideal point (Chernev 2003). But besides this *objective effort*, the primary driver of choice difficulty seems to be the *cognitive effort* due to considerations underlying the comparisons to be made. One important determinant of cognitive effort is the similarity between the items (Scheibehenne et al. 2010). Fasolo et al. (2009) studied real world assortments and showed that as the number of items in a set increase, the *density*¹ of an item sets grows, i.e., the differences between the items on their underlying attributes become smaller. This increases the similarity of the items and likewise the required cognitive effort to make a decision, as the number of potential candidates that are close to each other increases. For example, in such a dense set, the second best option is typically also very attractive, causing potential *regret* with the chosen option (Schwartz 2004). Moreover, people prefer to make decisions they can easily justify (Shafir et al. 1993), especially when item sets become larger (Sela et al. 2009). Fasolo et al. (2009) argue that in large sets attribute levels are so close to one another that it is hard to decide which option is better, and the final decision is subsequently harder to justify. This difficulty due to the increase in cognitive effort is what we expect to drive the perceived *choice difficulty*.

¹ Fasolo et al. (2009) define density as the distance (on each attribute) between products and their closest neighbors.

Scheibehenne et al. (2010) observed in their meta-analysis of the choice overload phenomenon that previous research has not controlled for similarity/diversity and the number of tradeoffs between options in the set, thereby making it difficult to disentangle the effects of item set *diversity* and *size* on the ease of comparison and thus choice overload. In the experimental paradigms employed in existing work such control is indeed difficult, because there is typically no knowledge of individual-level preferences of the participants to be able to individually control for levels of diversity and tradeoffs. To wit: increasing diversity without accounting for individual-level preferences may simply render some of the choice options irrelevant to the participant's taste.

In the current paper we will use a matrix factorization recommender algorithm to *diversify* based on the latent features calculated by the algorithm and used for predicting ratings. These latent features allow us to manipulate the diversity of a set of items while controlling for attractiveness (by keeping the quality of the item set constant). For example, a set of items with the same predicted ratings that have very similar scores along these latent features would create a uniform set with a high density, potentially causing much choice difficulty. A different set—again with the same predicted ratings—that maximizes the distance between the features would create a more varied set with a lower density. Without reducing the overall quality of the item set, this diversification method can potentially reduce the cognitive effort in choosing from an item set (i.e. choice difficulty) for the same number of items.

However, there may be a negative side to diversification, one that has received little attention in the literature: As options become more varied, they may encompass more difficult *tradeoffs*. Such tradeoffs generate conflicts that require a lot of effort to resolve, as they require one to sacrifice something when choosing one item over another. In other words, there might be a limit to how much diversification can reduce choice difficulty. Scholten and Sherman (2006) proposed a double mediation model, which shows a U-shaped relation between the size of the tradeoffs in a set and the amount of conflict that is generated by the set. They suggest that not only very uniform sets (i.e., sets with high density) are difficult to choose from (as we argued earlier), but also very diverse sets are difficult because of tradeoff difficulty: tradeoffs are larger, which makes it more difficult to make the decision because greater sacrifices need to be incurred. In other words, there might be an optimal (medium) level of diversification that has the lowest choice difficulty.

In the next section we will review relevant literature on diversification and explain why for our particular research goal we employed diversification on the latent features of a matrix factorization algorithm.

2.2 Diversifying recommendations

The ultimate goal of a recommender system is to produce a personalized set of items that are the most relevant to a user. However, only relying on highest predicted relevance can result in ignoring other factors that influence user satisfaction. In line with this Smyth and McClave (2001) already argued that there is a tradeoff between attaining the highest accuracy and providing sufficient diversity. Their (bounded) greedy

diversity algorithm tried to find the items that are the most relevant to the current user but that are maximally diverse from the other candidate items. McGinty and Smyth (2003) subsequently argued that the optimal level of diversity depends on the users' phase in the decision process. In the initial stage of interaction with a recommender system, diversity helps a user to find a set of relevant items sooner, thereby speeding up the recommendation process. Later in the process too much diversity reduces the recommendation efficiency, as relevant cases might be lost due to the diversification process. However, case-based recommenders studied in these papers differ from collaborative filtering algorithms we employ in the current paper, so these findings cannot readily be generalized to our work.

Bridge and Kelly (2006) similarly investigated the role of diversification and showed that different methods of diversifying collaborative filtering recommendations makes simulated users reach their target items in a conversational recommender system more quickly. Their diversification was based on item-item similarity calculated from rating patterns. Recently Ribeiro et al. (2014) also proposed a diversification method using item similarity based on rating patterns. A simulation showed that their diversification method results in recommendations that are simultaneously accurate, diverse and novel.

Existing research such as discussed above typically evaluates diversity by performing simulations to show that enhancing diversity improves the accuracy or efficiency with which simulated users interact with a recommender system. Ge et al. (2010) propose a number of metrics that are more suitable for the evaluation of diversity: Coverage, describing the number of items a recommender system can and does recommend; and Serendipity, describing to what extent the recommended items are unexpected and satisfactory. However, to better understand how real users perceive and evaluate diversity, and to be able to answer our question of whether diversification can reduce choice difficulty, we need to go beyond these simulations and study actual users' diversity perceptions, choice difficulty, and satisfaction with diversified item sets. We will briefly review relevant work that did employ user experiments to study the effect of item set diversity.

Ziegler et al. (2005) showed that topic diversification (using an ontology acquired separately from the data used to calculate recommendations) can enhance the *perceived attractiveness*² of a list of book recommendations. In a user experiment they demonstrated that despite the lower precision and recall of the diversified recommendations, diversification had a positive effect on users' perception of the quality of item sets produced by their recommender algorithm. The effects as established in their study were small, and overall perceived quality reached a plateau at some level of diversification, after which it actually decreased (cf. Scholten and Sherman 2006). As Ziegler et al. did not measure any potential moderating or mediating variables, their study provides no insight into the psychological mechanisms underlying this interesting effect. In addition, they acknowledge that the use of an external ontology may have led to a mismatch between the diversity calculated by the algorithm and the diversity perceived by their users.

² Ziegler et al. actually measured 'overall list value' and labeled this as satisfaction, but we prefer to use the more descriptive term 'perceived attractiveness' for this measure.

Knijnenburg et al. (2012, Sect. 4.8) attempted to reproduce Ziegler et al.'s findings in a movie recommender. In this study, recommendation sets were diversified based on movie genre information. Going beyond the original study by Ziegler et al., this study made use of the user-centric evaluation framework by Knijnenburg et al. to provide insight into the underlying mechanisms that allow diversification to increase choice satisfaction. The study confirmed Ziegler et al.'s positive effect of diversification on perceived quality, and showed that this in turn decreased choice difficulty and increased perceived system effectiveness, ultimately leading to a higher choice satisfaction.

In the present paper, we will expand on this earlier work to gain more insight in what factors affect choice difficulty and satisfaction in recommender systems. For this purpose we will test a diversification method that does not require external sources such as an external ontology (Ziegler et al. 2005) or a genre list (Knijnenburg et al. 2012), but that provides direct control over item quality and item set diversity on the individual level, using latent feature diversification.

2.3 Latent feature diversification

Matrix factorization algorithms (Koren et al. 2009) are widely used in recommender systems. These algorithms are based on singular value decomposition that reduces a high-dimensional user/item rating matrix into two lower dimensional matrices that describe users and items as vectors in a latent feature space in such a way that the relative positions of a user and item vector can be used to calculate the predicted ratings. In essence, the rating a user is predicted to assign to a specific item is equal to the inner product of the corresponding user- and item-vectors.

This approach is mathematically analogous to the multi-attribute utility theory (MAUT) framework (Bettman et al. 1998) used in decision making psychology. Matrix factorization models share with existing theories of MAUT the idea that the utility of an option is a sum of the utility of its attributes weighted by an individual decision maker's attribute weights. The difference is that the dimensions in matrix factorization models describe abstract (latent) features, while in MAUT they describe concrete (interpretable) features.

The simplified choice of a camera can serve as an example of how decisions are described in MAUT and how we can diversify options while controlling for attractiveness. Imagine Peter, who is considering buying a camera, described along two attribute dimensions: zoom and resolution. Peter thinks that resolution and zoom are equally important, so his user vector is $\bar{w}_{peter} = (0.5, 0.5)$. If we assume a linear utility function for attribute values (i.e. doubling an attribute value doubles the utility for the corresponding attribute), the utility for a 10 MP/10 $\times$ zoom camera $u_{peter, 10 MP/10 \times} = \bar{w}_{peter} \cdot \bar{x}_{10 MP/10 \times} = (0.5 * 10) + (0.5 * 10) = 10$ is higher than that of a 12 MP/7 $\times$ zoom camera $u_{peter, 12 MP/7 \times} = \bar{w}_{peter} \cdot \bar{x}_{12 MP/7 \times} = (0.5 * 12) + (0.5 * 7) = 9.5$, so Peter would prefer the first option. Similarly, we can calculate the utility for a 8 MP/12 $\times$ zoom and 15 MP/5 $\times$ zoom and see that Peter would like these equally much as the 10 MP/10 $\times$ alternative.

This example illustrates that there exist equipreference hyperplanes of items—orthogonal to the user vector—that are all equally attractive for that user (their inner

Fig. 1 Camera example to illustrate diversity in MAUT

products are the same) but that might differ a lot in terms of their features. The three cameras (10 MP/10 $\times$, 8 MP/12 $\times$ and 15 MP/5 $\times$) are in such a hyperplane (or more precisely on a line in this 2-dimensional example), see Fig. 1. To test the effect of diversification independent of attractiveness, our goal is to choose items on this equipreference line that are either close to each other (low diversity, e.g. 11 MP/9 $\times$, 10 MP/10 $\times$ and 9 MP/11 $\times$) or far apart from each other (high diversity, e.g. like the 8 MP/12 $\times$, 10 MP/10 $\times$ and 15 MP/5 $\times$ cameras in the figure). In terms of (Fasolo et al. 2009) the first set has a higher density than the second set because the inter-attribute differences are smaller, and we expect the first set of similar items therefore to require more cognitive effort (and thus cause more choice difficulty). However, if the items are too far apart then the increased trade-off difficulty may increase the choice difficulty again (cf. Scholten and Sherman's (2006) U-shaped relationship).

2.3.1 Latent features in matrix factorization

Matrix factorization models are similar to MAUT in the way that items have higher predicted ratings if they score high on (latent) features that an individual user has higher weights for. A difference is that in MAUT attributes describe concrete, identifiable properties of choice alternatives, while in matrix factorization they describe abstract, 'latent' features. Though it is hard to ascribe a simple unidimensional meaning to these features (cf. Graus 2011), they are related to the dimensionality of users' preferences (Koren et al. 2009 suggest they might encompass composite attributes such as 'escapist' movies). Using these latent features in a way similar to how attributes are treated in MAUT allows us to construct sets of items of the same quality (i.e. predicted rating) that are either highly diverse (with alternatives that have larger differences on latent feature scores, i.e. low density) or more similar (with alternatives that have smaller differences on latent feature scores, i.e. high density).

Fig. 2 Latent feature diversification

More specifically, we will construct such sets in the movie domain, for which large existing datasets allow us to construct a good matrix factorization model (Bollen et al. 2010). Our ideal goal is to construct sets of movies that differ in diversity, while controlling for quality. This means we would like to select a diverse (or non-diverse) set of movies from a hyperplane (orthogonal on the user vector) of movies with similar predicted ratings, analog to our camera example above. However, as we are bound by existing sets of movies and their vectors, it is in practice impossible to select a sufficiently large and diverse set that will have all the movies on *exactly* the same hyperplane. Our requirement to manipulate recommendation set diversity while controlling for quality thus requires us to tolerate some variation in predicted ratings. This implies that we have to find a small range (ΔR) in predicted ratings that allows us to extract both high and low diversity sets. This type of sets is illustrated in Fig. 2 for two dimensions.

As the factors in a matrix factorization model tend to be normally distributed, the latent feature space represents a multidimensional hypersphere, or simply a circle in the two-dimensional case. The Top-N predicted items of a user are those that are farthest in the direction of the user vector. We are interested in selecting a set of recommendations with high predicted ratings, but with sufficient diversity. To achieve this, we select our items to diversify from a Top-N set of highest predicted ratings for a user (light-grey area). From that area we select both low-diversity and high-diversity subsets of items (darker grey areas). Our constraints are that the width of the set (or the maximum distance perpendicular to the user vector) is maximal to achieve high diversity, and the height of the set (or the maximum distance parallel to the user vector, or ΔR : the differences in ratings) is minimal to achieve low variation in item ratings (i.e. attractiveness), while still providing enough items to allow extracting both types of sets.

2.3.2 Diversification algorithm

The algorithm from Ziegler et al. (2005) was adopted and altered to meet the requirements of the current study. The algorithm greedily adds items to the set of

```

Candidates  $\leftarrow$  Top  $N$  predictions by MF
 $R \leftarrow \{\}$ 
FirstItem  $\leftarrow$  the  $i \in \text{Candidates}$  for which  $d(i, \text{Centroid}(\text{Candidates}))$  is minimal
Insert FirstItem into  $R$ 
Remove FirstItem from Candidates
while  $|R| < k$  do
    best  $\leftarrow$  the  $i \in \text{Candidates}$  for which  $\sum_{l \in R} d(i, l)$  is maximal
    Insert best into  $R$ 
    Remove best from Candidates
end while
return  $R$ 

```

Fig. 3 Latent feature diversification algorithm

recommendations, maximizing at each addition the total distance between all items in the set. This section will elaborate on the algorithm and will check the validity of the algorithm through simulation.

The applied algorithm (see Fig. 3) performs a greedy selection from an initial set (*Candidates*) to extract a maximally diverse recommendation set (R) of size k . Implementing the algorithm requires deciding on two parameters: the distance metric to use, and what set to take as initial set of candidate items. In this specific application, we used the first order Minkowski distance ($(d(a, b) = \sum_k |a_k - b_k|$, with a, b as items and k iterating over the latent features), also known as Manhattan, City Block, Taxi Cab or L_1 distance as the distance metric $d(i, l)$. This was done to ensure that differences along different latent features are considered in an additive way, and large distances along one feature cannot be compensated by shorter distances along other features. This means that two items differing one unit on two dimensions are considered as different as two items differing two steps along one dimension (using Euclidean distance, these differences would be $\sqrt{2}$ and 2 respectively). This is more in line with how people perceive differences between choice alternatives with real attribute dimensions. Additionally, initial analyses showed that we obtained about 15 % higher diversity for City Block distance than for common Euclidian distances.

To ensure that quality is kept equal for different levels of diversity, the diversification starts from the centroid of the initial recommendation set. This allows for selecting items with both higher and lower predicted ratings, and as a result every set will have the same average predicted rating. Note that selecting the centroid is for the sole purpose of experimental control. From a more practical perspective this approach might be suboptimal as in many cases it will exclude the highest predicted items to be part of the recommended set. We will address this issue in study 2 through a slight adaptation of our algorithm.

Diversity can be manipulated by restricting the sets from which items are chosen in the diversification algorithm. For our high level diversity, the algorithm selected the N most diverse items among the 200 items with the highest predicted rating. For medium diversity, the algorithm selected the N most diverse items from the 100 items closest to the centroid of these 200 items. For low diversity, the algorithm simply selected the N items closest to the centroid. To test our algorithm we ran several simulations and tests using the 10M movielens (Harper and Konstan 2015) dataset. As an initial starting set, the Top-200, or the 200 items with highest predicted rating was found to provide a good balance between maximum range in predicted rating and maximum

Table 1 Mean scores for AFSR and predicted rating (and their standard deviations) per Recommendation Set Size and Level of Diversification, based on a simulation with 200 random users from the movielens dataset. SE indicates the standard error

Set size	Level of diversification	Average feature score range (AFSR) mean (SE)	Predicted rating mean (SE)	SD predicted rating mean (SE)
5	Low	0.295 (0.005)	4.505 (0.001)	0.088 (0.004)
	Medium	0.634 (0.008)	4.529 (0.001)	0.115 (0.004)
	High	1.210 (0.012)	4.561 (0.002)	0.143 (0.006)
20	Low	0.586 (0.008)	4.537 (0.001)	0.134 (0.004)
	Medium	1.005 (0.011)	4.558 (0.001)	0.164 (0.004)
	High	1.615 (0.013)	4.604 (0.002)	0.212 (0.007)

diversity. The range in predicted rating was lower than the mean average error (MAE) of the predictions of our matrix factorization model, implying that even the difference in rating between the highest and lowest prediction in the set fall within the error margin of predictions for a particular user. The attractiveness differences (ΔR) would therefore most likely not be perceived by the user.

To measure the success of the diversification algorithm, we measured two properties: the average rating of the recommendation set (which should be similar across diversification levels to ensure equal quality of the item sets), and the diversity of the recommendation set. There are many measures of diversity available, see for example the recent overview by [Castells et al. \(2015\)](#) that provides a number of metrics. However, as we aim for a specific type of diversity related to distances between options on the features, we base our measure on the density measure of [Fasolo et al. \(2009\)](#) that was used in earlier work on choice overload. We define the *AFSR* (Average Feature Score Range) of a recommendation set X as the average difference per feature (i_k) between the highest and lowest scoring items along that feature (Eq. 1).

$$AFSR(X) = \sum_{k=1}^D \frac{\max(i_k) - \min(i_k)}{D}, \quad (1)$$

where $i \in X$, i_k is the score of item i on feature k , and D is the number of dimensions.

We ran a simulation to verify that these recommendation sets meet the requirements. For 200 users randomly selected from the original MovieLens dataset, the recommendation sets for three different diversification levels were calculated and compared in terms of AFSR and average predicted rating. The results for recommendation sets with 5 (representing smaller recommendation sets) and 20 (large item sets) items can be found in Table 1. The numbers show that while the average predicted rating only differed minimally (about 0.06 stars between low and high, with more variance for higher diversity) over the diversity levels, the AFSR does differ a lot. This shows that the algorithm succeeds in manipulating diversity while maintaining equal levels of quality.

2.4 Overview of our user studies

In what follows, we present two studies that test latent feature diversification. In the first study, we test the user perceptions and experiences with recommendation lists for different levels of diversification and different number of items in the list. This study has two parts. The first part, study 1a, has been presented (and published) in an earlier version as workshop paper (Willemsen et al. 2011) and encompasses a basic test of our diversification algorithm. It asks users to inspect 3 lists with different levels of diversification, and measures subjective perceptions of attractiveness, diversity and difficulty. In this study users only reported their perceptions and did not choose an item from the recommendations. The second part, study 1b, performs a study in which users also choose an item, thus allowing us to directly test the impact of diversification on actual choice difficulty and choice satisfaction. Combined, the results of study 1 show that users indeed perceive the diversification generated by the algorithm, and that it is beneficial for reducing choice difficulty and improving satisfaction, especially for smaller item sets.

In study 1 the diversification starts from the centroid of the top-200. This way we controlled the average predicted rating of the lists (as described in Sect. 2.3). This means, though, that the lists used in study 1 are different from standard Top-N lists. The second study tests how the lists produced by our diversification algorithm compare against Top-N lists (i.e., lists that are optimized for predicted ratings). For this we modified our diversification algorithm slightly, starting the diversification from the top-predicted item (rather than the centroid), and manipulating diversity by varying the balance between predicted rating and diversity. Study 2 replicates the main result of study 1, showing that there is a benefit in diversifying recommendations based on the latent features to reduce choice difficulty and improve satisfaction, especially for small item sets.

3 Study 1

3.1 Goals of study 1

The first study aims to test how our latent feature diversification affects the user perceptions and experiences with recommendation lists. We test this in two steps. Study 1a ($N = 97$) tested if our diversification is at all perceived by the user and whether diversity can be linked to tradeoff and choice difficulty. To do this, the level of diversification was manipulated within subjects, (i.e. all subjects were asked to assess three lists with different diversification levels). List length was manipulated between subjects, i.e. each user only saw one list length. After each list, users' perceptions of the list were queried. Users were not asked to choose an item from the lists, and no choice satisfaction was measured. In study 1b ($N = 78$) participants were given only one set of personalized recommendations (list length and diversification were both manipulated between subjects) but this time they were asked to make a choice, and they also reported their choice satisfaction (along with choice difficulty, diversity and

attractiveness). Study 1b thus allowed us to see if diversification can indeed reduce choice difficulty and increase satisfaction.

3.2 System used in study 1

Both studies 1a and 1b used a movie recommender with a web-interface used previously in the MyMedia project. The software is currently being developed as MyMediaLite.³ A standard Matrix Factorization algorithm was used for the calculation of the recommendations. The 10M MovieLens dataset was used for the experiment, which, after removing movies from before 1994, contained 5.6 million ratings by 69820 users on 5402 movies. We further enriched the MovieLens dataset with a short synopsis, cast, director, and a thumbnail image of the movie cover taken from the Internet Movie Database.

The Matrix Factorization algorithm used 10 latent features, a maximum iteration count of 100, a regularization constant of 0.0001 and a learning rate of 0.01. Using a 5-fold cross validation on the used dataset, this specific combination of data and algorithm resulted in an RMSE of 0.854 and an MAE of 0.656, which is up to standards. An overview of performance metrics is given by (Herlocker et al. 2004).

Below we will discuss the setup and results of study 1a and 1b separately.

3.3 Study 1a: setup and expectations

The diversification algorithm manipulated the density of the set of recommendations while keeping the overall attractiveness of the set (in terms of the predicted ratings) constant. By using three levels of diversification, we were able to investigate whether the relation between diversity and difficulty is linear (higher diversity always reduces choice difficulty because it makes it easier to find reasons to choose one item over another; the predominant view in the literature) or U-shaped (diversity only helps up to a certain level, but very high diversity might result in large tradeoffs between items that are difficult to resolve due to the sacrifices that need to be made when choosing one item over the other; the view proposed by (Scholten and Sherman 2006)). As our main interest was in assessing differences in perceptions of diversity between the lists, we employed a within-subject design in which each participant is presented (sequentially) with a low, medium and high diversification set. This increases statistical power by allowing us to test the effect of diversification within subjects. To prevent possible order effects, the order of these sets was randomized over participants. We explicitly did not ask people to choose an item from each list, as we considered it to be unnatural to have to choose a movie three times, and spill-over effects might occur from one choice to another (i.e., the perception of diversity might be influenced by the preferences constructed for the movie chosen in the previous task).

We also varied the number of items in the set on 5 levels (5, 10, 15, 20 or 25; a between-subjects manipulation, each user saw all three lists in one of these lengths), as

³ See <http://www.mymedialite.net>.

the literature suggests that diversification might have a stronger impact on larger sets. However, given that recommender systems output personalized and highly attractive sets of items, we might find that even for small sets diversification has a strong impact on experienced difficulty.

To measure the subjective perceptions and experiences of these recommendations after the presentation of each set, we employed the user-centric framework for user experience of recommender systems (Knijnenburg et al. 2012). This framework models how objective systems aspects (OSA) can influence user experience (EXP) and user interactions (INT) and how these relations are mediated by users' perceptions as measured by Subjective System Aspects (SSA). The framework also models the role of situational characteristics (SC) and personal characteristics (PC).

Based on this framework we expected that the effect of diversification (OSA) on users' subjective evaluation of the set (EXP) (i.e. how difficult it is to make tradeoffs and choose from the set), is mediated by subjective perceptions (SSA) of the diversity and attractiveness of the set. In particular, we expect that item sets that are more diverse (i.e., that have a lower density on the attributes) are perceived as more varied and potentially also as more attractive, and that these two factors affect the experience of tradeoff difficulty and expected choice difficulty (again, as we did not ask to actually choose an item, so we asked how difficult they would *expect* it to be to choose from the list).

3.3.1 Research design and procedure

Participants for this study were gathered using an online participant database that primarily consists of (former) students who occasionally participate in online studies and in lab studies of the HTI group. Participants were compensated with 3 euro (about 4 US dollars) for participating. 97 participants completed the study (mean age: 29.2 years, sd: 10.3; 52 females and 45 males).

The study consisted of three parts. In the first part, participants answered a set of questions to measure a number of personal characteristics (PC). In their meta-analysis, Scheibehenne et al. (2010) show that the characteristics expertise and prior preferences are important moderators of choice overload. Therefore, we constructed a set of items to measure movie expertise and strength of preferences. We also measured maximizing tendency of our participants, using the short 6-item version (Nenkov et al. 2008) of the maximization questionnaire by (Schwartz et al. 2002). Schwarz defines people who always try to make the best possible choice as maximizers, and people who settle for "good enough" as satisficers. Maximizers consider more options whereas satisficers stop looking when they have found an item that meets their standards. Therefore the search costs of maximizers are higher and consequently it is suggested that they are more prone to experience choice difficulty.

After these questions, the second part of the study was used to gather rating information from the participant to be able to calculate and provide personalized recommendations. In this phase the participants were asked to rate a total of ten movies, which provided adequate recommendations without too much effort in earlier experiments (Bollen et al. 2010; Knijnenburg et al. 2012). They were presented with ten randomly selected movies at a time, with the instruction to rate only the movies they

Fig. 4 Screen shot of the recommendation part of interface, with a set of recommendations (titles and year) with predicted ratings. The user hovers the first title and is shown the information of this item

were familiar with (ratings were entered on a scale from 1 to 5 stars). After inspecting and (possibly) rating some of the ten movies shown, users could get a new list of movies by pressing a button. When the participant had entered ten ratings (or more if on this particular page a user would cross the 10 ratings threshold) in total, this button would guide them to the third part.

In the third part the participant sequentially (and in randomized order) received three sets of recommendations, each time with a different level of diversification (OSA). List length (OSA) was manipulated between subjects: each participant was either shown a rank-ordered list of 5 ($N = 19$), 10 ($N = 22$), 15 ($N = 22$), 20 ($N = 16$) or 25 ($N = 18$) movies; each movie represented by its title (see Fig. 4). The predicted rating (in stars and one-point decimal value) was displayed next to the title. If the participant hovered over one of the titles, additional information would appear in a separate preview panel. This additional information consisted of the movie cover, the title of the movie, a synopsis, the name of the director(s) and part of the cast. Before moving to the next list, participants were presented with a short questionnaire of 16 items, measuring perceived choice difficulty and tradeoff difficulty (EXP) and their perceived diversity and perceived attractiveness of the presented list (SSA). Participants thus answered these questions for each of the three lists.

3.3.2 Measures

The items in the questionnaires were submitted to a confirmatory factor analysis (CFA, as suggested by Knijnenburg et al. 2012; Knijnenburg and Willemssen 2015). The CFA used repeated ordinal dependent variables and a weighted least squares estimator, estimating 5 factors. Items with low factor loadings, high cross-loadings, or high residual correlations were removed from the analysis. Factor loadings

of included items are shown in Table 2, as well as Cronbach's alpha and average variance extracted (AVE) for each factor. The values of AVE and Cronbach's alpha are good, indicating convergent validity. The square root of the AVEs are higher than any of the factor correlations, indicating good discriminant validity. For expected choice difficulty, a single indicator was used. Based on the CFA, in which no stable construct for tradeoff difficulty could be fitted due to high cross loadings with other constructs, we selected a single indicator for tradeoff difficulty as well.

3.3.3 Results

Before we construct a structural equation model (SEM) that shows how the measured constructs relate to each other and to the diversification manipulation, we first check if our diversification algorithm indeed results in different levels of attribute diversity.

3.3.3.1 Manipulation checks To check our diversification algorithm, we compared the resulting diversity, predicted ratings and variance of the predicted ratings in our data (analyzed across list lengths). Table 3 shows that our diversification algorithm indeed increases the average range of the scores on the 10 matrix factorization features (AFSR, see Sect. 2.3). At the same time, the predicted average rating does not differ significantly between the three levels of diversification (differences are smaller than the standard error), showing that we manipulated diversity independent of (predicted) quality. The standard deviation of the predicted ratings for the three sets does increase slightly with increasing diversity.

3.3.3.2 Structural model The subjective constructs from the CFA were organized into a path model using a confirmatory structural equation modeling (SEM, as suggested by Knijnenburg et al. 2012; Knijnenburg and Willemsen 2015) approach with repeated ordinal dependent variables and a weighted least squares estimator. In the resulting model, the subjective constructs are structurally related to each other and to the experimental manipulations (i.e. the OSAs list length and diversification level). The model was constructed based on the user-centric framework for user experience of recommender systems (Knijnenburg et al. 2012). For analysis, we followed the recommendations by Knijnenburg and Willemsen (2015), creating a saturated model (following the framework's core pathway linking OSAs to EXP variables via SSAs) and then pruning non-significant effects from this model.

In the final model, the maximizer scale did not relate to any other variable, and was therefore removed from the analysis. The manipulation "list length" (whether participants were shown 5, 10, 15, 20 or 25 recommendations) also did not have a significant influence on the other variables, nor did it interact with diversification. The results are therefore collapsed over these conditions. We also did not observe any effect of the order in which the three sets were presented.

Table 2 Items presented in the questionnaires. Items without a factor loading were excluded from the analysis

Considered aspects	Items	Factor loading
Maximizing tendency (PC) Alpha: 0.86 AVE: 0.823	When I am in the car listening to the radio, I often check other stations to see if something better is playing, even if I am relatively satisfied with what I'm listening to	
	No matter how satisfied I am with my job, it's only right for me to be on the lookout for better opportunities	
	I often find it difficult to shop for a gift for a friend	
	Renting videos is really difficult. I'm always struggling to pick the best one	
	No matter what I do, I have the highest standards for myself	0.903
Strength of preference (PC) Alpha: 0.77 AVE: 0.628	I never settle for the second best	0.911
	I have clearly defined preferences concerning movies	0.890
	I know what kind of movies I like	0.924
	Most of the time I let someone else pick a movie for me	-0.491
	I trust the opinions of others when choosing a movie	
Movie expertise (PC) Alpha: 0.89 AVE: 0.746	I am a movie lover	0.826
	Compared to my peers I watch a lot of movies	0.951
	Compared to my peers I am an expert on movies	0.842
	I only know a few movies	-0.830
Expected choice difficulty (EXP, single item)	I would find it difficult to choose a movie from this list	1.000
Tradeoff difficulty (EXP, single item)	The list contained movies that were good on some aspects, and other movies that were good on other aspects	
	I had to put a lot of effort into comparing the different aspects of the movies	1.000
	Some movies were clearly better than others	
Perceived recommendation diversity (SSA) Alpha: 0.87 AVE: 0.617	Many movies had comparable good aspects	
	The list of movies was varied	0.932
	All the movies were similar to each other	-0.811
	Most movies were from the same genre	-0.744
	Many of the movies in the list differed from other movies in the list	0.794
	The movies differed a lot from each other on different aspects ^a	0.612

Table 2 continued

Considered aspects	Items	Factor loading
Perceived recommendation attractiveness (SSA) Alpha: 0.93 AVE: 0.794	I would give the recommended movies a high rating	0.905
	The list of movies showed too many bad items	−0.838
	The list of movies was attractive	0.965
	I didn't like any of the recommended items	−0.773
	The list of recommendations matched my preferences	0.960
	The list of movies had at least one movie that I liked	

^a This item was originally asked as the third item of the tradeoff difficulty scale, but in hindsight fitted much better with the diversity scale

Table 3 Mean diversity (measured by AFSR) and mean predicted ratings (and their standard deviations) for the presented items in study 1a. SE indicates standard error of the mean

Diversity	Average feature score range (AFSR) mean (SE)	Predicted rating mean (SE)	SD predicted rating mean (SE)
Low	0.959 (0.015)	4.486 (0.042)	0.163 (0.010)
Medium	1.273 (0.016)	4.486 (0.041)	0.184 (0.011)
High	1.744 (0.024)	4.527 (0.039)	0.206 (0.013)

The final model had a good model fit⁴ ($\chi^2(179) = 256.5$, $p < 0.001$, CFI = 0.989, TLI = 0.987, RMSEA = 0.041, 90 % CI [0.030, 0.051]). Figure 5 displays the effects found in this model. Path coefficients in the final model are standardized; the numbers on the arrows (A → B) denote the estimated mean difference in B, measured in standard deviations, between the medium or high diversification list and the low diversification list. For all other arrows, the numbers denote the standardized increase or decrease in B, given a one standard deviation increase or decrease in A. The number in parentheses denotes the standard error of this estimate, and the p-value below these two numbers denotes the statistical significance of the effect. As per convention, only effects with $p < .05$ are included in the model. The medium and high diversification conditions are compared to the low diversification baseline condition; numbers on the arrows originating in the conditions denote the mean differences between participants in medium or high diversification condition and participants in the low diversification condition.

To better understand the effects, we plotted the marginal effects of the medium and high diversification condition relative to low diversification condition on the subjective constructs in Fig. 6. Our diversification algorithm affects the perceived diversity in a seemingly linear fashion (see Fig. 6), with medium and high diversification resulting in

⁴ Theoretically, a good model is not statistically different from the fully specified model ($p > 0.05$). However, this statistic is commonly regarded as too sensitive (Bentler and Bonett 1980). Based on extensive simulations, Hu and Bentler (1999) propose cut-off values for other fit indices to be: CFI > 0.96, TLI > 0.95, and RMSEA < 0.05, with the upper bound of its 90 % CI falling below 0.10.

Fig. 5 The SEM fitted for study 1a

significantly higher perceived diversity than the low diversification condition. Higher perceived diversity subsequently increases the perceived attractiveness of the recommendations. The medium diversification condition also has a direct positive effect on attractiveness, making medium diversification as attractive as the high diversification (and both are significantly more attractive than low diversification; see Fig. 6). There is also a direct effect of expertise, a personal characteristic, on attractiveness, showing that expert participants report higher perceived attractiveness.

In terms of tradeoff difficulty, we observe that this is significantly (and negatively) influenced by the high diversification condition, as well as a main effect of strength of preferences (consistent with the work of Chernev (2003) on ideal point availability). So both high diversification and a high self-reported strength of preferences cause people to experience lower tradeoff difficulty. The negative effect of diversification on tradeoff difficulty goes against the expectation (Scholten and Sherman 2006) that higher diversification leads to options that encompass larger tradeoffs between the attributes. Potentially, the high diversification setting of our algorithm does not generate items that encompass difficult tradeoffs, at least not in the specific domain of movies.

These constructs together influence the expected choice difficulty experienced by the user, which goes up with increased tradeoff difficulty, but goes down with increased diversity and attractiveness. The net result of our diversification manipulation on expected choice difficulty is negative: the higher the diversity of the set, the more attractive and diverse, and the less difficult participants expect it to be to choose from the set (the marginal effects in Fig. 6 suggest that choice difficulty decreases almost linearly with diversification level).

3.3.4 Discussion of study 1a results

The results of study 1a show how expected choice difficulty and tradeoff difficulty associated with a set of recommendations are influenced by the diversity of the items

Fig. 6 Marginal effects on perceived diversity, attractiveness, tradeoff difficulty, and expected choice difficulty for mid and high diversification relative to low diversification. *Error bars* are one standard error of the mean

on the underlying latent features. This is a noteworthy result, as it shows that these latent features have a psychological meaning to the participant, providing support for our basic assumption that a parallel can be drawn between attribute spaces in regular decision making domains and the latent features in a matrix factorization space. By diversifying the items on these latent features, we increase the perceived diversity and attractiveness of the set, and subsequently reduce expected choice difficulty. Our net result thus is not a U-shaped relation between diversity and choice difficulty, but rather a simple downward trend.

Though intuitively one would expect an effect of list length on the perceived diversity or the experienced difficulty, we do not observe such an effect. Given that our diversification algorithm finds items that are maximally spaced out from each other on the latent features within a set of options with the same quality (predicted ratings), this might be not very surprising: when all items are good, diversification helps for both small and large sets. We would thus expect the effect of diversification to be roughly equal for different list lengths. Moreover, as the 5 different list lengths were manipulated between subjects, we have limited statistical power

to detect small differences.⁵ Finally, in study 1a we did not ask participants to make a choice from the item sets. This might be one reason why no effect of list length on choice difficulty was observed as the participants did not have to commit to any of the options.

Study 1a has established an effect of our diversification algorithm on perceived diversity and *expected* choice difficulty. In study 1b, we further investigated our diversification algorithm, but now we explicitly asked participants to choose an item from the list of recommendations so we can measure their *actual* choice difficulty and also measure their satisfaction with the chosen item (as in [Bollen et al. 2010](#)).

3.4 Study 1b

3.4.1 Setup and expectations

The goal of the second study was to measure the effect of diversification on users' satisfaction with the chosen option by using an experimental design in which the participants choose an item from the presented set. Study 1b therefore allows us to measure to what extent actual (rather than perceived) choice difficulty might be reduced by diversifying a large recommendation set. Furthermore, we expected that even for small sets diversification might be beneficial: In Study 1a, the number of items did not seem to affect the perceived diversity, showing that diversification enhances the attractiveness of both large and small item sets. In Study 1b we wanted to find out whether such diversified small item sets can render large item sets obsolete. For non-personalized item sets, this would be a daunting task: it seems impossible to create a small set of items that would fit everyone's needs. However, a *personalized* diversified small item set might be just as satisfying as a large set, since this set is tailored to the user. In that case, the additional benefits of more items may not weigh up against the increased difficulty.

To accomplish these goals, we conducted a study in which we manipulate both diversity and item set size. Study 1a showed no detrimental effects of the high level of diversification due to large tradeoffs; this condition just showed stronger effects than medium diversification. In the current study we thus only include low and high levels of diversification, as this will show us the most pronounced effect on choice satisfaction. Besides varying the level of diversification, we also vary the number of items in the set on three levels, 5, 10 and 20 items. The extreme levels (5 and 20) are identical to those used by [Bollen et al. \(2010\)](#). As most literature on choice overload (e.g., [Reutskaja and Hogarth 2009](#)) predicts an inverted U-shaped relation between satisfaction and number of items, we explicitly included an intermediate level but we did not use 5 levels (as in study 1a) as this would require two additional (between-subjects) conditions and thus many additional participants. Different from study 1a, and to prevent spill-over effects from participants' previous decisions, participants

⁵ In our analysis we also attempted to group conditions (e.g. contrast 5 and 10 items against 15–25 items) to increase statistical power, but again did not observe statistically significant differences in experienced difficulty between these two groups.

Table 4 Items presented in the questionnaires. Items without a factor loading were excluded from the analysis

Considered aspects	Items	Factor loading
Perceived recommendation diversity (SSA) Alpha: 0.82 AVE: 0.594	The list of movies was varied	0.890
	All the movies were similar to each other.	−0.729
	The list of recommendations contained movies from many different genres	0.809
	Many of the movies in the list differed from other movies in the list	
	The movies differed a lot from each other on different aspects	0.632
Perceived recommendation attractiveness (SSA) Alpha: 0.92 AVE: 0.792	I would give the recommended movies a high rating	0.946
	The list of movies showed too many bad items	−0.739
	The list of movies was attractive	0.933
	I didn't like any of the recommended items	
	The list of recommendations matched my preferences	0.927
Choice satisfaction (EXP) Alpha: 0.95 AVE: 0.882	How satisfied are you with the chosen movie?	0.904
	My chosen movie could become one of my favorites	0.930
	I would recommend the chosen movie to others	0.966
	I think I would enjoy watching the chosen movie	0.953
	I would rather rent a different movie from the one I chose.	
Choice difficulty (EXP) Alpha: 0.73 AVE: 0.555	I think I chose the best movie from the options	
	I was in doubt between several movies on the list	
	I changed my mind several times before making a decision	
	The task of making a decision was overwhelming	0.903
	It was easy to select a movie	−0.662
	Comparing the movies took a lot of effort	0.640
	I have a good justification for my decision	

only received one list of recommendations and we manipulate both diversification (2 levels) and list length (3 levels) between subjects.

3.4.2 Research design and procedure

Participants for this study were gathered using the same online participant database as study 1a, but participants from study 1a were excluded from participation. Participants were compensated with 3 euro (about 4 US dollars) for participating. 87 participants completed the study (mean age: 29.0 years, sd: 8.91; 41 females and 46 males).

Like study 1a, study 1b consisted of three parts. Part 1 (measuring personal characteristics) and part 2 (asking for ratings to train the system) were identical to study 1a. In the third part participants received one set of recommendations, which, depending on the experimental condition, consisted of 5, 10 or 20 items that were of high or low diversification. See Table 5 for the number of participants in each condition. Recommended movie titles were presented list-wise, with the predicted rating (in stars

Table 5 Mean diversity (measured by AFSR) and mean predicted ratings (and their standard deviations) for the presented items in our study

List length	Diversity	N (number of participants)	Average feature score range (AFSR) mean (SE)	Predicted rating mean (SE)	SD predicted rating mean (SE)
5	Low	13	0.695 (0.041)	4.473 (0.140)	0.130 (0.026)
	High	13	1.228 (0.044)	4.436 (0.122)	0.232 (0.044)
10	Low	14	0.800 (0.033)	4.568 (0.105)	0.166 (0.026)
	High	14	1.510 (0.053)	4.515 (0.087)	0.186 (0.019)
20	Low	12	0.961 (0.044)	4.635 (0.113)	0.141 (0.024)
	High	12	1.810 (0.039)	4.599 (0.094)	0.242 (0.048)

SE indicates standard error of the mean

and one-point decimal value) displayed next to the title. If the participant hovered over one of the titles, additional information appeared in a separate preview panel, identical to study 1a. In study 1b, however, we asked participants to choose one item from the list (the movie they would most like to watch) before proceeding to the final questionnaire. The final questionnaire presented 22 items, measuring the perceived diversity and attractiveness of the recommendations, the choice difficulty and participants' satisfaction with the chosen item. Some items of these questionnaires were slightly reworded from study 1a (e.g., we reverse coded one item of the diversity scale to have a more balanced set of items). We did not measure tradeoff difficulty, because we did not require this aspect after finding in study 1a that the hypothesized U-shaped effect of diversity (which was hypothesized to result from a positive effect of diversity on tradeoff difficulty) did not hold (a negative effect of diversity on tradeoff difficulty was found instead).

3.4.3 Measures

After initial inspection of the process data, we excluded 9 participants that clearly put little effort in the experiment. These participants went over the entire experiment unrealistically quickly and only inspected 1 or 2 movies for a very short time (less than 500 ms) during the decision. The remaining analysis thus contains 78 participants.

The items in the questionnaires were submitted to a confirmatory factor analysis (CFA). The CFA used ordinal dependent variables and a weighted least squares estimator, estimating 7 factors. Items with low factor loadings, high cross-loadings, or high residual correlations were removed from the analysis. Factor loadings of included items are shown in Table 4, as well as Cronbach's alpha and average variance extracted (AVE) for each factor. The values of AVE and Cronbach's alpha are good, indicating convergent validity. The square roots of the AVEs are higher than any of the factor correlations, indicating good discriminant validity. In Table 4 we only report the 4 factors that we used in the final SEM. Maximization tendency, strength of preference and expertise showed loadings similar to study 1a. However, in the resulting SEM, none of these constructs significantly contributed to the model, except for a small effect of

preference strength on perceived difficulty that did not affect the other relations in the model. This effect is left out because it does not contribute to our overall argument.

Besides these subjective measures we also included a behavioral variable representing the log-transformed total number of hovers users made on the movie titles. Hovering on a movie title was required to read the description of that movie. A large number of hovers means that participants switch back and forth⁶ to compare different movies (to exclude accidental or transitional hovers, only hovers longer than 1 second were counted). This measure is a behavioral indicator of the amount of effort users put into making the decision. As discussed in the introduction, this measure would project the *objective effort* a user puts into choosing, whereas the choice difficulty questions measure the *cognitive effort* of the user.

3.4.4 Results

3.4.4.1 Manipulation check For completeness, we checked again if the diversification algorithm impacts the actual diversity of the latent features without affecting the predicted ratings. Higher diversity indeed causes a significantly higher average feature score range (AFSR; see Table 5, differences much larger than the standard error of the mean), while the average predicted rating of the items remains equal between the two diversity levels (differences smaller than standard error of the mean). As in study 1a, we do observe a slight increase in the standard deviation of the predicted ratings with increasing diversity.

3.4.4.2 Structural model The subjective constructs from the CFA (see Table 4) were organized into a path model using a confirmatory SEM approach with ordinal dependent variables and a weighted least squares estimator. In the resulting model, the subjective constructs are structurally related to each other, to the hover measure and to the experimental manipulations (list length and diversification level). As we already noted, the maximizer scale, strength of preference, and expertise did not relate to any other variable, and were therefore removed from the analysis. The final model had a good model fit ($\chi^2(167) = 198.9$, $p = 0.046$, CFI = 0.99, TLI = 0.99, RMSEA = 0.050, 90 % CI [0.007, 0.074]). Figure 7 displays the effects found in this model.

The model shows that choice satisfaction increases with perceived attractiveness and perceived diversity, but decreases with the experienced choice difficulty and the number of hovers made. The two manipulations (diversity and list length) are represented by the two purple OSA boxes at the top of the display. The three levels of list length are tested simultaneously as well as separately.⁷ As expected, longer lists (the 10- and 20-item set) increase the number of hovers, as well as the choice difficulty. They also have a direct positive effect on satisfaction compared to the 5-item

⁶ As we were interested in switching behavior we prefer the hover count over hover time and we indeed found that hover time was not a significant factor in our model.

⁷ The χ^2 values test the effect of the 10- and 20-item sets simultaneously, thus testing the null hypothesis that there is no difference between the three conditions ($\beta_{5\text{-item-set}} = \beta_{10\text{-item-set}} = \beta_{20\text{-item-set}}$); the β coefficients below the χ^2 values represent the effect (in β SD difference) of each set size (10 and 20) tested against the baseline of 5 items.

For perceived attractiveness we see similar results. Attractiveness is higher when diversity is perceived to be higher, and therefore only the highly diverse 5-item set has a significantly higher attractiveness (see Fig. 8) compared to the low diverse 5-item set.

⁸ As in study 1a, factors are scaled to have an SD of 1.

Fig. 8 Marginal effects of our manipulations on the subjective constructs in the SEM model. Error bars represent 1 standard error from the mean

to be less difficult to choose from. In total, for both low and high diversity, the difficulty increases monotonically with list length, and we see that difficulty is on its lowest level for the highly diverse 5-item set (see Fig. 8).

Finally, we investigate the resulting choice satisfaction. There is a direct effect of set size on satisfaction, showing higher satisfaction with longer lists. However, satisfaction is also influenced positively by perceived attractiveness and diversity, and negatively by higher choice difficulty and increased hovers. Note that the effect of hovers is independent of choice difficulty: the number of hovers increase with list length, and the more hovers people make the lower their satisfaction. So both the objective effort (hovers) and cognitive effort (choice difficulty) incurred by longer lists reduce satisfaction.

The marginal effects on satisfaction (Fig. 8) reveal that 10- and 20-item sets are perceived as more satisfying than the non-diversified 5-item set, but that there are no significant differences between the 10- and 20-item sets. The most interesting result is that the highly diverse 5-item set stands out, because it is perceived to be as satisfying as the 10- and 20-item sets. Our model also explains why: the diversified 5-item list excels in terms of perceived diversity and attractiveness, while at the same time being

less difficult to choose from. These three effects completely offset the direct negative effect of short list length on satisfaction. Note that the study does not reveal a choice overload effect, as we do not observe that longer items lists are less satisfying than shorter ones (despite the increased choice difficulty).

3.4.5 Discussion of study 1b results

Study 1b reveals an important role of diversity and list length on the choice satisfaction of decision makers using a recommender system. Consistent with Study 1a we see that diversification increases the attractiveness of an item set, while at the same time reducing the choice difficulty. Together these effects can increase the satisfaction with the chosen item, but we observe this effect for small item sets only; diversification does not seem to matter much for larger item sets, as these sets are satisfying even with low diversification.

Unlike study 1a where diversification increased perceived diversity regardless of set size, we now observe that diversification increases perceived diversity in *small sets only*. This limited effect in study 1b is not entirely surprising, because study 1b differs from study 1a in many aspects. Most importantly, in study 1a we asked participants to carefully *inspect* the entire set of items, before asking them their perceptions of the list. In study 1b, on the other hand, we asked them to *choose* from the set, without asking them explicitly to inspect the entire list. This means that participants might have stopped inspecting the set earlier (i.e. they may have ignored some items altogether), and that they are therefore less able to assess its overall diversity, especially for the larger sets. Inspecting the hovers corroborates our expectation that participants behave differently: in study 1a the percentages of unique acquisitions are relatively high; between 80–90 % of the items are actually looked at. These numbers do not differ much between 5, 10 and 20 items. In study 1b the percentages of unique acquisitions drop substantially with increasing list length—from 80 % in the 5-item set to only 45 % in the 20-item set. In other words, if the participant is asked to choose (rather than to inspect), she is much more likely to inspect fewer items, and might therefore have less opportunity to get a good impression of the overall diversity of the set. The resulting assessment on the diversity scale might therefore regress towards the middle, showing lower values for the 10- and 20-item sets compared to the 5-item set.

Consistent with previous research (Bollen et al. 2010), our model shows that satisfaction with the chosen item is an interplay of choice difficulty and attractiveness. However, different from Bollen et al., we also observe a direct effect from perceived diversity on satisfaction, and a direct effect of list length on satisfaction. These effects might explain why we do not observe a choice overload effect in this study, as people seem to derive inherent satisfaction from choosing from longer lists that overrules the difficulty of choosing from such longer lists.

In terms of attractiveness and satisfaction, item sets with a length of at least 10 seem to be optimal when creating a recommender system. However, this study convincingly shows that small but diverse items sets can be just as satisfying as larger sets, and more importantly, that they result in lower choice difficulty. Diverse small sets may thus be the best thing one could offer a user of a recommender system.

3.5 Overall discussion of study 1

The goal of study 1 was primarily to show that a diversification based on psychological principles can be effective in positively enhancing subjective experiences with a recommender system. The two experiments of study 1 clearly show that latent feature diversification can reduce choice difficulty and—especially for short item lists—increase choice satisfaction. However, the precise experimental control we enforced in study 1 has a drawback for more practical applications. Most importantly, in study 1 we diversified our item set from the centroid of the Top-200 latent feature space. This allowed us to attain maximal diversity while controlling for the average quality in each list, but it also implied that our lists did not excel in terms of predicted accuracy, as they did not necessarily include items with the highest predicted ratings. So despite the fact that we show that diversified lists are more attractive, less difficult and in some cases more satisfying, we do not know yet if such diversified lists can hold up against standard lists such as the Top-N lists that exclusively present the items with the highest predicted rating, as typically provided by recommender algorithms.

For a practical test of our diversification algorithm, we will need to adapt our diversification algorithm to allow us to tradeoff the most accurate/relevant items with the most diverse items. This allows for a realistic comparison against the pure Top-N lists that are pervasive in today's recommender systems research and applications. In the next section we discuss this adaptation to our diversification approach in detail, and compare it against the centroid diversification approach used in study 1. We then present study 2, a study similar to study 1b, in which we hypothesize that latent feature diversification can hold up against Top-N recommendations—by means of reducing choice difficulty and increasing attractiveness and satisfaction—even though the recommended items have lower predicted ratings than undiversified Top-N lists.

4 Diversification versus Top-N

Because the diversification in study 1 was started from the centroid of a Top-N set, there was no guarantee that the items with the highest predicted rating were included in the actual recommendation set. This begs the question how diversification compares to a 'standard' Top-N recommendation list.

To make this comparison, the diversification algorithm needs to be altered slightly. For a direct comparison to Top-N the diversification process needs to start from the item with highest predicted rating instead of the item closest to the centroid of the Top-200. To make sure that this alteration did not reduce the effectiveness of the diversification process, its consequences in terms of AFSR were studied by performing a simulation. For 100 randomly selected users from the MovieLens dataset, AFSR was calculated for recommendation sets that started either from the item with highest predicted rating (Top-1) or the centroid. Rather than a decreased AFSR, Table 6 shows even a marginal increase in AFSR when starting from the top predicted item rather than the centroid, indicating that starting from the top does not result in lower diversity than the diversification method starting from the centroid employed in study 1a and 1b. Thus, we can safely assume that the high diversification condition created in study

Table 6 AFSR scores for high diversity sets of 5/20 items comparing new (starting from Top-1) and old (starting from Centroid) algorithms

AFSR	Recommendation set size			
	5		20	
	Mean	SE	Mean	SE
Top-1	1.218	0.005	1.583	0.006
Centroid	1.185	0.005	1.574	0.006

2 (where we start from Top-1) is similar to the high diversification condition used in study 1 (where we started from the centroid).

As study 2 aims to investigate the effect of different levels of diversification compared to ‘standard’ Top-N recommendations, we needed a method to vary the level of diversification along a scale ranging from ‘standard’ Top-N recommendations (no diversification) to recommendations with high diversification, which, for a better comparison, start from the item with highest predicted rating (Top-1). This was done by introducing a weighting parameter α (similar to Ziegler et al. 2005), that controls the relative importance of Top-N ranking versus diversity.

At every iteration in the diversification, all candidate items are ranked on two criteria: 1) the distance of this item with all other items in the recommendation set, and 2) the predicted rating. The weighting factor α defines the relative weight that is given to either ranking (α for the diversification ranking and $1 - \alpha$ for the predicted rating ranking), resulting in a weighted combined ranking. This allows for a continuous range of diversification, where $\alpha = 0$ results in the Top-N recommendation set (diversity is not considered at all), $\alpha = 1$ results in the high diversity recommendation set (predicted rating is not considered at all), and any value in between results in a mix of diversity and predicted rating.

Whereas in study 1b we only used two levels of diversification, in this study at least three levels of diversification are needed to test whether a pure Top-N ($\alpha = 0$), a pure diversification⁹ ($\alpha = 1$), or a mix of both is preferred. For this latter “medium diversity” condition, we try to balance average predicted rating and AFSR, as taking diverse items comes at the expense of items with high predicted ratings. In order to find the right point in this trade-off, AFSR and average rating were plotted for several levels of α in Fig. 9a–d. These figures show for AFSR an inflection point at $\alpha = 0.3$ (see Fig. 9a, b), indicating that the highest relative gain in AFSR occurs going from $\alpha = 0$ to $\alpha = 0.3$. Fig. 9c, d show linearly decreasing average ratings for increasing alpha. These findings hold for both short (left panels) and long (right panels) recommendation lists.

Based on these findings, the optimal α -level for the medium diversification condition was determined to be 0.3. The mean AFSR and average predicted rank (the position of the recommended item in a list ordered on decreasing predicted rating) of the resulting

⁹ Even for pure diversification, items are taken from the Top-200, so they are still of high personalized quality.

Fig. 9 **a** AFSR across alpha levels (5-item sets), **b** AFSR across alpha levels (20-item sets), **c** Average rating across alpha levels (5-item sets). **d** Average rating across alpha levels (20-item sets)

recommendation sets are displayed in Table 7 for recommendation lists of 5 and 20 items. The table shows that the medium diversification condition lies approximately halfway between the high and low diversification condition in terms of AFSR. We also inspect the average rank, which is more informative than the mean rating as we want to know how close the new medium lists are to the Top-N list. The average rank of the medium diversification list is still quite low (close to the Top-N), showing a nice balance between AFSR and rank (i.e. predicted rating) for this medium diversification.

5 Study 2

5.1 Setup and expectations

Study 2 investigates the trade-off between conventional Top-N sets and diversified recommendation sets for small and larger item sets. For this purpose, study 2 manipulated diversification on three levels and list length on two levels. For diversification, we compared non-diversified Top-N lists with highly diversified lists similar to the high diversification condition from Study 1, and one medium diversification list that trades off accurate Top-N lists with diverse lists, as described in Sect. 4. For item set size we used 5 and 20 items, the two outer levels of study 1b. We did not include the 10-item lists, since there were no significant differences between 10 versus 20

Table 7 Predicted AFSR and average Rank per condition

Diversification	Recommendation set size	
	5	20
AFSR		
High ($\alpha = 1$)	1.380	1.793
Medium ($\alpha = 0.3$)	1.096	1.486
Top-N ($\alpha = 0$)	0.774	1.270
Rank		
High ($\alpha = 1$)	78.284	89.380
Medium ($\alpha = 0.3$)	7.484	17.849
Top-N ($\alpha = 0$)	3.000	10.500

items in the previous studies. Additionally, this allowed us to limit the number of between-subjects conditions to 6 (3×2).

We expect that despite the fact that diversified lists in study 2 contain theoretically less attractive items than Top-N lists, such diverse lists will be perceived to be less difficult to choose from, as study 1 showed strong effects of diversification on choice difficulty. Both experiments in study 1 also showed that diversified lists were perceived to have a higher quality. However, in study 1 we controlled for equal attractiveness between diversified lists, whereas in study 2 we exert less control over the average predicted rating in order to compare against Top-N lists. By design, the Top-N lists will have items with higher predicted ratings than the diversified lists, which select items with a lower predicted rating (ranking) in order to improve diversity. So for study 2 we do not have strong predictions about the effect of diversity on perceived attractiveness, as on the one hand the Top-N lists should be theoretically the most attractive (highest predicted rating) but more diverse lists were perceived to be attractive in study 1. As choice satisfaction is an interplay between difficulty and perceived attractiveness, we expect that diversification will overcome the difficulty associated with choosing from high-quality item lists, so that in the end diversified lists, despite having lower predicted ratings, might be as satisfactory or even more satisfactory to choose from than traditional Top-N lists. Study 1b also showed that the effect of diversification was more pronounced for small item sets. Similarly, we expect to find a more positive effect of diversification for short lists in Study 2.

5.2 Research design and procedure

Participants for this study were gathered using the JFS participant database from our department, a different database with different participants from those in study 1. Participants were compensated by playing in a lottery that paid 15 euro for 1 out of every 5 participants (expected return was 3 euro). 165 participants completed the study (mean age: 25.6 years, $SD = 9.8$, 59 % males). After initial inspection of the data, 6 participants were excluded because they clearly put little effort in the experiment (short response times, inconsistent responses in the survey).

The procedure of Study 2 was similar to study 1b, with a few minor changes. Because Study 1b did not find any effects of the maximization scale, we dropped this scale from the experiment. The remaining items measuring personal characteristics (i.e., of preference strength and movie expertise) were subsequently moved to after the choice task as part of the larger survey evaluating diversity, attractiveness, difficulty and satisfaction. Study 2 thus consisted of two parts. Participants started with the task to rate at least 15¹⁰ items, using the same movie recommender system as used in study 1. In the second part, participants received a set of recommendations, which, depending on the experimental condition, consisted of 5 or 20 movie titles that were of high, medium or no (Top-N) diversification (the number of participants per condition are in Table 9). The predicted ratings were displayed next to the movie titles. If the participant hovered over one of the titles, additional information appeared in a separate preview panel, identical to study 1. Participants were asked to choose one item from the list before proceeding to the final questionnaire. The final questionnaire presented 36 questions measuring the perceived diversity and attractiveness of the recommendations, the choice difficulty, and participants' satisfaction with the chosen item, as well as preference strength and movie expertise, and a new set of questions that measured familiarity with the movies (which was included as an additional control, as Top-N lists might contain more popular/familiar movies).

5.3 Measures

The items in the questionnaires were submitted to a confirmatory factor analysis (CFA). The CFA used ordinal dependent variables and a weighted least squares estimator, estimating 7 factors. Items with low factor loadings, high cross-loadings, or high residual correlations were removed from the analysis. In the final SEM model, the familiarity construct correlated too much with attractiveness, and was therefore removed from the analysis. We also excluded the preference strength and expertise constructs from the final SEM. Different from study 1b, these constructs did load significantly¹¹ (with small effect sizes) on some of the other constructs. However, a model without these personal characteristics showed identical relations between the other constructs: in other words the preference and expertise constructs do not help our understanding of the interplay between perceived diversity and attractiveness, difficulty and satisfaction. Moreover, as these questions are asked after interaction, the responses might also be affected by the experiences with the recommender system not just by the users' general movie expertise or preference strength, as we measured it in study 1. Thus, for sake of brevity and for easier comparison with study 1b, we decided to exclude these constructs from the models.

Factor loadings of the included items are shown in Table 8, as well as Cronbach's alpha and average variance extracted (AVE) for each factor. The values of AVE and

¹⁰ We selected 15 rather than 10 items to get slightly better predictions for each user.

¹¹ These significant effects might be partly explained by the fact that now the expertise and preference strength questions were asked together with the other constructs, rather than in a separate questionnaire before the interaction with the system, as was done in study 1.

Table 8 Items presented in the questionnaires. Items without a factor loading were excluded from the analysis

Considered aspects	Items	Factor loading
Perceived recommendation diversity (SSA) Alpha: 0.80 AVE: 0.596	The list of movies was varied.	0.834
	All the movies were similar to each other.	−0.733
	The list of recommendations contained movies from many different genres	
	Many of the movies in the list differed from other movies in the list	0.792
	The movies differed a lot from each other on different aspects	0.724
Perceived recommendation attractiveness (SSA) Alpha: 0.89 AVE: 0.712	I would give the recommended movies a high rating	0.797
	The list of recommended movies showed too many bad items.	−0.791
	The list of recommended movies was attractive	0.943
	I didn't like any of the recommended items	
	The list of recommendations matched my preferences	0.837
Choice satisfaction (EXP) Alpha: 0.93 AVE: 0.847	I am satisfied with the movie I chose	0.916
	My chosen movie could become one of my favorites	0.910
	I would recommend the chosen movie to others	0.959
	I think I would enjoy watching the chosen movie	0.895
	I would rather rent a different movie from the one I chose	
Choice difficulty (EXP) Alpha: 0.71 AVE: 0.534	I think I chose the best movie from the options	
	I was in doubt between several movies on the list	
	I changed my mind several times before making a decision	
	The task of making a decision was overwhelming	0.516
	It was easy to select a movie	−0.911
	Comparing the movies took a lot of effort	0.711
	I have a good justification for my decision	

Cronbach's alpha are good, indicating convergent validity. The square roots of the AVEs are higher than any of the factor correlations, indicating good discriminant validity.

5.4 Results

5.4.1 Manipulation check

For completeness, we checked if the diversification algorithm impacts the actual diversity of the recommendation sets, and the mean and SD of the predicted ratings. Higher diversity indeed causes a significantly higher average feature score range (AFSR; see Table 9). Different from study 1b, the high diversification condition now results in

Table 9 Diversity and predicted ratings of the presented items in study 2

List length	Diversification	N (number of participants)	Average feature score range (AFSR) mean (SE)	Predicted rating mean (SE)	SD predicted rating mean (SE)
5 items	None (Top-N, $\alpha = 0$)	25	0.770 (0.025)	4.506 (0.062)	0.054 (0.008)
	Medium ($\alpha = 0.3$)	27	1.157 (0.040)	4.411(0.058)	0.145 (0.015)
	High ($\alpha = 1$)	27	1.351 (0.027)	4.339 (0.053)	0.269 (0.020)
20 items	None (Top-N, $\alpha = 0$)	26	1.366 (0.033)	4.407 (0.054)	0.109 (0.009)
	Medium ($\alpha = 0.3$)	27	1.643 (0.026)	4.347(0.066)	0.139 (0.013)
	High ($\alpha = 1$)	27	1.854 (0.031)	4.116 (0.056)	0.188 (0.013)

Fig. 10 The SEM fitted for study 2

significantly lower predicted ratings compared to the no diversification (Top-N) condition. The medium diversification condition is still close to the predicted ratings of the Top-N (about 1 SE lower). As in study 1, we do observe an increase in the standard deviation of the predicted ratings with increasing diversification.

5.4.2 Structural model

The subjective constructs from the CFA (see Table 8) were organized into a path model using a confirmatory structural equation modeling (SEM) approach with ordinal dependent variables and a weighted least squares estimator. In the resulting model, the subjective constructs are structurally related to each other and to the experimental manipulations (list length and diversification level, coded as a linear contrast between the three levels¹²). As in study 1b, we included a variable representing the log-transformed number of hovers users made. The final model had a good model fit ($\chi^2(127) = 165.263$, $p = 0.0127$, CFI = 0.992, TLI = 0.991, RMSEA = 0.044, 90 % CI [0.021, 0.061]). Figure 10 displays the effects found in this model.

Similar to study 1b, choice satisfaction increases with perceived attractiveness and list length and decreases with choice difficulty and the number of hovers. Different from study 1b, where hovers only affected satisfaction, the number of hovers also increases the difficulty and reduces attractiveness, indicating both direct and indirect negative effects of an increased effort on choice satisfaction. The number of hovers is (not surprisingly) impacted by list length: users inspect more items when confronted with a 20-item list. However, list length also directly influences satisfaction positively, indicating there are two opposing forces on choice satisfaction: a direct positive effect

¹² We also dummy-coded the levels of diversification with similar results, but found that a linear effect was a better predictor. We also did not find any significant interaction effects between list length and diversification levels.

Fig. 11 Marginal effects of our manipulations on the subjective constructs in the SEM model of study 2. Error bars represent 1 standard error from the mean

(users are more satisfied when choosing from larger lists) but several opposite negative effects due to increased effort by means of additional hovers. We calculated the total effect of list length on choice satisfaction and found that it is essentially zero, as the total indirect effect (via all indirect paths through hovers and accuracy and difficulty) is -0.332 ($SE = 0.111$, $p = 0.003$), which almost completely negates the direct positive effect.

As in study 1, we find that perceived diversity positively influences the perceived attractiveness. This is noteworthy, as the diversified lists have lower predicted ratings. Perceived diversity is only affected by the level of diversification, not by list length. It increases linearly with the level of diversification, as can be seen in the marginal effects plots in Fig. 11. Different from study 1, the perceived diversity does not reduce difficulty but there is a direct negative effect of the level of diversification on difficulty: more diversified lists are less difficult to choose from (and the plots in Fig. 11 show that this effect is mostly driven by the medium diversification condition).

The marginal effects of the model on satisfaction (see Fig. 11) reveal that 20-item sets are perceived as more satisfying than Top-5-item sets, but that for the 5-item set choice satisfaction increases with diversity. Consequently, the highly diverse 5-item set stands out, being just about as satisfactory to choose from as the 20-item set. These results are especially noteworthy because the Top-N sets in this study have an advantage in terms of their predicted ratings: they contain items that have a higher rank, i.e. they are likely more attractive for the user.

Table 10 Position, rating and rank of chosen items in study 2

List length	Diversity	Chosen option (mean and std. err)		
		List position	Rating	Rank
5 items	None (Top-5)	3.60 (0.27)	4.51 (0.07)	2.60 (0.27)
	Medium	4.41 (0.59)	4.41 (0.07)	13.52 (5.37)
	High	4.19 (0.27)	4.30 (0.07)	76.59 (12.76)
20 items	None (Top-20)	10.15 (0.92)	4.45 (0.05)	9.15 (0.92)
	Medium	10.33 (1.18)	4.40 (0.08)	16.7 (2.68)
	high	9.93 (1.07)	4.16 (0.07)	71.22 (11.84)

As lists in this study were diversified from the highest predicted item, the diversified lists still contain several items that have high predicted ratings. So an alternative explanation for the relative advantage of diversified list could be that users simply picked items that had the highest predicted ratings. In this case users would mainly pick items that would also have shown up in the non-diversified list—a result that would make it hard to claim the success of diversification. Note also that it would be easy and effortless for users to implement this strategy, because we sorted the list on predicted ratings, putting items with the highest predicted rating first. It is therefore important to also inspect what items users choose from the lists.

Analyzing their choices, it is clear that participants do not exclusively pick the highest-ranked items from the diversified lists. In fact, Table 10 (list position column) shows that participants in the diversified conditions did not choose items higher up the list than participants in the Top-N conditions. It follows that they picked items with lower predicted rating/rank (since items lower down the diversified lists have an even lower rating and rank than items lower down the regular Top-N lists), resulting in less ‘optimal’ choices in terms of predicted ratings. Note that these ‘suboptimal’ choices nevertheless come without any reduction in choice satisfaction, and for smaller lists there is even a clear increase in choice satisfaction for more diverse lists. At the same time, these diversified lists result in a reduction in choice difficulty. Therefore, these results show that, especially for short lists, our latent feature diversification can outperform standard Top-N lists in terms of user experience.

5.5 Discussion of study 2 results

Study 2 shows, in line with the results of study 1b, that diversification can be beneficial in terms of reducing effort and difficulty without any negative effects on the overall satisfaction with the item they choose from these lists. The results are substantially different for short (5) and long (20) item lists. For longer item lists, the differences in choice satisfaction are small, whereas for shorter lists the differences are substantial. The high diversity 5-item list equals the Top-5 list in terms of satisfaction, despite the fact that the average rank of the chosen option in this list is much lower, due to the

diversification. This provides a concrete demonstration of [McNee et al. \(2006\)](#) claim that “being accurate is not enough” for a recommender system to succeed.

As in study 1b, study 2 does not provide evidence for choice overload in the classical sense: Longer lists are not perceived to be less satisfying than short lists. Although these longer lists are associated with more choice difficulty (cognitive effort) and more effort (study 2 clearly shows the mediating effects of the number of hovers on choice difficulty and satisfaction), there is a direct effect that makes these longer lists more satisfying. Apparently, in the domain of movies, 20 items are not too overloading. In fact, in a classroom experiment similar to study 2 we tested a list length of 30 items and still did not find a reduction in overall satisfaction, compared to a 5-item list. The upper bound on the total number of recommendations that users are comfortable with may thus be even higher than that. This may be partially due to the convenient ordering of the list by predicted rating, as well as the fact that people are used to choosing from much larger assortments in real life (e.g. on commercial video services such as Netflix and YouTube).

That said, this study clearly demonstrates that diversification can reduce cognitive effort as measured by our choice difficulty construct, as we already observed in study 1b. Study 2 solidifies these results by demonstrating that this positive effect of diversification holds even when comparing the diversified lists to Top-N lists with higher predicted quality, and that especially the diversified 5-item set stands out. We can thus indeed reiterate our suggestion that diverse small sets may be the best thing one could offer a user of a recommender system.

6 General conclusion and discussion

6.1 Discussion of the results

We inspected the influence of latent feature diversification on choice difficulty and satisfaction in two studies containing three user experiments. Using the user-centric framework of [Knijnenburg et al. \(2012\)](#) we were able to demonstrate how our diversification manipulations are perceived and experienced by users, and thereby explain why they are successful. Our results suggest that diversification helps to reduce choice difficulty (the cognitive effort with making a decision), and that diversified smaller sets might be the key to prevent choice overload. Selecting from a small, diversified set of recommendations reduces objective effort and choice difficulty, but maintains similar or higher levels of satisfaction than choosing from larger, non-diverse sets. This holds even when we compared (in study 2) the diversified list against Top-N lists with higher predicted ratings.

A concern one might have is that diversification removes too many of the highest ranked items in a set (i.e. the list is suboptimal in terms of precision and recall), and that therefore such a diversification would not hold up against Top-N lists that per definition contain (what are predicted to be) the highest quality items. Our choice data provides counter-evidence to these concerns: If our users would just be focused on selecting the highest ranked movies, we should observe them to predominantly pick the movies at the top of the list (i.e. the items with the highest predicted ratings) and even more so

for the high diversification lists. The data of study 2 suggest otherwise: Users do not choose higher up the list with increasing diversification, but instead select items that have a lower rank and lower predicted rating than users in the Top-N condition. This takes away most of our concerns, and again indicates that there is more to satisfying recommendations than just prediction accuracy, as McNee et al. have argued before (McNee et al. 2006).

Our results have important implications for both the field of psychology and the field of recommender systems. From a theoretical perspective, our results seem to corroborate the idea that the latent features in matrix factorization algorithms have psychological meaning, as suggested before (Koren et al. 2009; Graus 2011). The ability to control psychological features of recommendations may turn out to be a very valuable asset for future research in psychology. Specifically, researchers have suggested that findings from decision-making research can be very relevant for the development of recommender systems (Jameson et al. 2011; Mandl et al. 2011). Our results suggest that manipulating the underlying latent feature space of matrix factorization algorithms can help to explore the role of decision-making phenomena in a controlled, systematic way, and therefore might also be interesting for psychologists to apply.

Specifically, in regard to choice overload, latent feature manipulations allowed us to exert more control over the stimuli presented to users, enabling us to diversify the recommendations while keeping the average predicted rating constant. This in turn enabled us to more thoroughly investigate the role of diversity on attractiveness and difficulty, the mediating factors of choice overload. Until now, these mediators of choice overload have not been given the attention they deserve (Chernev et al. 2010). Moreover, what constitutes choice difficulty is still part of a larger debate in decision making psychology and consumer research (Scholten and Sherman 2006; Fasolo et al. 2009; Reutskaja and Hogarth 2009). In our introduction we distinguished the cognitive effort from objective effort, two drivers of difficulty that until now could not easily be disentangled. Study 1b and 2 have shown that our latent feature diversification is able to directly reduce the cognitive effort, and that this factor could be distinguished as a different driving force on choice satisfaction than the objective effort as measured by the number of hovers. These results provide relevant insights for future psychological or consumer research.

For practitioners in the field of recommender systems, the results suggest that latent feature diversification might be a promising way to improve the user experience of recommender systems while at the same time reducing the complexity of the interface and reducing the cognitive effort expended by the user. Especially for mobile recommenders, where device restrictions require short lists of recommendations, diversification can provide a way to keep the overall satisfaction high.

6.2 Limitations

We need to be cautious in generalizing the present results, as we have only studied the domain of movies. Future work should study choice overload both in similar ‘leisure’ domains (i.e., music and books) but also in different domains such as finan-

cial decisions (e.g., pension plans) or consumer decisions (e.g., home appliances or car purchases) that encompass important tradeoffs among the underlying features. In those domains, tradeoff difficulty may play a more important role, so diversification may have a U-shaped effect on satisfaction after all. Moreover, in those other domains we might expect choice difficulty to be higher, and likewise to show more benefits from diversification. Research in such domains should potentially also investigate if diversification as we employed is only useful for people with low preference stability (as earlier research ([Chernev 2003](#)) suggested). In our studies we did measure preference strength, but it did not have strong effects on the structural relations in our model. Future work could empirically compare low and high preference stability groups and the role of diversification. It would also be worthwhile to directly compare (in a user study) our diversification method (that used an internal measure: the latent features from the Matrix factorization model) with other methods that used external sources such as an external ontology ([Ziegler et al. 2005](#)) or a genre list ([Knijnenburg et al. 2012](#)).

Our studies also had some limitations. One is the size of the samples used. SEM modeling typically requires large samples (ideally 200 or more participants), more than we used in the current setup. However, study 1a used a repeated design (questionnaires were asked 3 times for three different diversifications) which provides additional power. Furthermore, we used pre-validated questionnaires from earlier studies with relatively high factor loadings and many items per factor, our models did not assume indicator normality (a WLS estimator was used), regressive effects were moderately large, and there was no missing data. Using these properties to extrapolate findings presented by [Wolf et al. \(2013\)](#) and [Sideridis et al. \(2014\)](#), we argue our sample size is in the right ballpark. We also made sure that the different models for our studies were consistent in the items used for each latent construct and the structural model as much as possible, which makes the results more robust from a reproducibility standpoint. Nevertheless, study 1a could not reveal an effect of list length potentially because of low statistical power, and also study 1b had a limited number of participants given the 6 conditions. This was one reason why we chose to use a larger participant pool in study 2. As study 1b and 2 are mostly in agreement, this strengthens our confidence in the replicability of our results.

Some design decisions were made for our studies that might affect the overall perceptions of the lists. For example, we showed the predicted ratings next to the recommended movie titles. Showing these ratings can affect the perception of attractiveness, or they could enhance the difficulty (e.g., if all ratings are 5 stars, this might amplify the choice overload effect). However, all conditions showed predicted ratings, so it did not affect the differences we observed between conditions. Regardless, it would be interesting to test if showing ratings affects the overall perceptions of choice difficulty and attractiveness. Note, though, that [Bollen et al. \(2010\)](#) manipulated the display of predicted ratings and found no differences between these conditions.

Other design decisions we already discussed briefly before. Study 1 diversifies from the centroid to empirically separate item quality (in terms of predicted rating) from diversity. This poses some practical limitations, and the slightly modified version that diversifies from the highest predicted rating as used in study 2 overcomes these limitations. We did not compare the two algorithms side by side in study 2 as this would

have required many additional participants. However, our analysis in Sect. 4 showed that the algorithms achieve similar diversification levels and the results of study 2 are similar to study 1b with respect to the effect of diversification on perceived diversity and attractiveness. We therefore think that the new instantiation of the algorithm used in study 2 is a good candidate for future research.

6.3 Conclusion

Concluding, our work shows a promising direction for the field of recommender systems. Whereas long lists of Top-N recommendations seem to be the industry norm, we encourage recommender systems researchers to take choice difficulty into consideration, and adopt our method of latent feature diversification as a potential solution. Indeed, in an age where users fear for a ‘filter bubble’ (Pariser 2012) and are ruled by the tyranny of choice (Schwartz 2004), the diversification of a small set of recommendations may come as a welcome change to the endless lists that are, indeed, just more of the same.

Acknowledgments The authors would like to thank the participants of the workshop Decisions@RecSys 2011 in which study 1a was presented before for helpful comments, and Kai Fu and Linda Velter-Bremmers, for assisting in running the first study. We thank James Lomas for his assistance in running Study 2.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Bentler, P.M., Bonett, D.G.: Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* **88**(3), 588–606 (1980)
- Berger, J., Draganska, M., Simonson, I.: The influence of product variety on brand perception and choice. *Market. Sci.* **26**(4), 460–472 (2007). doi:[10.1287/mksc.1060.0253](https://doi.org/10.1287/mksc.1060.0253)
- Bettman, J.R., Luce, M.F., Payne, J.W.: Constructive consumer choice processes. *J. Consum. Res.* **25**(3), 187–217 (1998)
- Bollen, D., Knijnenburg, B.P., Willemsen, M.C., Graus, M.: Understanding choice overload in recommender systems. Proceedings of the Fourth ACM Conference on Recommender Systems-RecSys '10, pp. 63–70. Barcelona (2010). doi:[10.1145/1864708.1864724](https://doi.org/10.1145/1864708.1864724)
- Bridge, D., Kelly, J.P.: Ways of computing diverse collaborative recommendations. In: Proceedings of the 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, pp. 41–50. AH'06. Springer, Berlin (2006) doi:[10.1007/11768012_6](https://doi.org/10.1007/11768012_6)
- Castells, P., Hurley, N.J., Vargas, S.: Novelty and diversity in recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (eds.) *Recommender Systems Handbook*, pp. 881–918. Springer, New York (2015). doi:[10.1007/978-1-4899-7637-6_26](https://doi.org/10.1007/978-1-4899-7637-6_26)
- Chernev, A.: When more is less and less is more: the role of ideal point availability and assortment in consumer choice. *J. Consum. Res.* **30**(2), 170–183 (2003)
- Chernev, A., Böckenholt, U., Goodman, J.: Commentary on scheibehenne, greifeneder, and todd choice overload: is there anything to it? *J. Consum. Res.* **37**(3), 426–428 (2010). doi:[10.1086/655200](https://doi.org/10.1086/655200)
- Fasolo, B., Hertwig, R., Huber, M., Ludwig, M.: Size, Entropy, and density: what is the difference that makes the difference between small and large real-world assortments? *Psychol. Market.* **26**(3), 254–279 (2009). doi:[10.1002/mar.20272](https://doi.org/10.1002/mar.20272)

- Ge, M., Delgado-Battenfeld, C., Jannach, D.: Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 257–260. RecSys '10. ACM, New York (2010) doi:[10.1145/1864708.1864761](https://doi.org/10.1145/1864708.1864761)
- Graus, M.: Understanding the latent features of matrix factorization algorithms in movie recommender systems. Eindhoven University of Technology, Master Thesis (2011). <http://alexandria.tue.nl/extra2/afstversl/tm/Graus%202011.pdf>
- Harper, F.M., Konstan, J.A.: The MovieLens datasets: history and context. *ACM Trans. Interact. Intell. Syst.* **5**(4), 19 (2015). doi:[10.1145/2827872](https://doi.org/10.1145/2827872)
- Herlocker, J.L., Konstan, Joseph A., Terveen, Loren G., Riedl, John T.: Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* **22**(1), 5–53 (2004). doi:[10.1145/963770.963772](https://doi.org/10.1145/963770.963772)
- Hu, L., Bentler, P.M.: Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Eq. Model.* **6**(1), 1–55 (1999). doi:[10.1080/10705519909540118](https://doi.org/10.1080/10705519909540118)
- Iyengar, S.S., Lepper, M.R.: When choice is demotivating: can one desire too much of a good thing? *J. Personal. Soc. Psychol.* **79**(6), 995–1006 (2000). doi:[10.1037/0022-3514.79.6.995](https://doi.org/10.1037/0022-3514.79.6.995)
- Jameson, A., Gabrielli, S., Kristensson, P.O., Reinecke, K., Cena, F., Gena, C., Vernerio, F.: How can we support users' preferential choice? In: Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, pp. 409–418. CHI EA '11. ACM, Vancouver (2011) doi:[10.1145/1979742.1979620](https://doi.org/10.1145/1979742.1979620)
- Knijnenburg, B.P., Willemsen, M.C.: Evaluating recommender systems with user experiments. In: Ricci, F., Rokach, L., Shapira, B. (eds.) *Recommender Systems Handbook*, pp. 309–352. Springer, New York (2015). doi:[10.1007/978-1-4899-7637-6_9](https://doi.org/10.1007/978-1-4899-7637-6_9)
- Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, Chris: Explaining the user experience of recommender systems. *User Model. User-Adapt. Interact.* **22**(4–5), 441–504 (2012). doi:[10.1007/s11257-011-9118-4](https://doi.org/10.1007/s11257-011-9118-4)
- Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. *IEEE Comput.* **42**(8), 30–37 (2009)
- McGinty, L., Smyth, B.: On the role of diversity in conversational recommender systems. In: Ashley, K., Derek, B. (eds.) *Case-Based Reasoning Research and Development. Lecture Notes in Computer Science*, p. 1065. Springer, Berlin (2003)
- Mandl, M., Felfernig, A., Teppan, E., Schubert, M.: Consumer decision making in knowledge-based recommendation. *J. Intell. Inform. Syst.* **37**(1), 1–22 (2011). doi:[10.1007/s10844-010-0134-3](https://doi.org/10.1007/s10844-010-0134-3)
- McNee, S.M., Riedl, J., Konstan, J.A.: 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. CHI EA '06. ACM, Montréal, Québec (2006) doi:[10.1145/1125451.1125659](https://doi.org/10.1145/1125451.1125659)
- Nenkov, G.Y., Morrin, M., Ward, A., Schwartz, B., Hulland, J.: A short form of the maximization scale: factor structure, reliability and validity studies. *Judgm. Decis. Mak. J.* **3**(5), 371–388 (2008)
- Pariser, E.: *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Books, New York, NY (2012)
- Reutskaja, E., Hogarth, R.M.: Satisfaction in choice as a function of the number of alternatives: when 'goods satiate. *Psychol. Market.* **26**(3), 197–203 (2009). doi:[10.1002/mar.20268](https://doi.org/10.1002/mar.20268)
- Ribeiro, M.T., Ziviani, N., Moura, E.S.D., Hata, I., Lacerda, A., Veloso, A.: Multiobjective pareto-efficient approaches for recommender systems. *ACM Trans. Intell. Syst. Technol.* **5**(4), 53:1–53:20 (2014). doi:[10.1145/2629350](https://doi.org/10.1145/2629350)
- Scheibehenne, B., Greifeneder, R., Todd, P.M.: Can there ever be too many options? a meta-analytic review of choice overload. *J. Consum. Res.* **37**(3), 409–425 (2010). doi:[10.1086/651235](https://doi.org/10.1086/651235)
- Scholten, M., Sherman, S.J.: Tradeoffs and theory: the double-mediation model. *J. Exp. Psychol.* **135**(2), 237–261 (2006). doi:[10.1037/0996-3445.135.2.237](https://doi.org/10.1037/0996-3445.135.2.237)
- Schwartz, B.: The tyranny of choice. *Sci. Am.* **290**(4), 70–75 (2004)
- Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., Lehman, D.R.: Maximizing versus satisficing: happiness is a matter of choice. *J. Personal. Soc. Psychol.* **83**(5), 1178–1197 (2002). doi:[10.1037/0022-3514.83.5.1178](https://doi.org/10.1037/0022-3514.83.5.1178)
- Sela, A., Berger, J., Liu, W.: Variety, vice, and virtue: how assortment size influences option choice. *J. Consum. Res.* **35**(6), 941–951 (2009). doi:[10.1086/593692](https://doi.org/10.1086/593692)
- Shafir, E., Simonson, I., Tversky, A.: Reason-based choice. *Cognition* **49**(1–2), 11–36 (1993)
- Sideridis, G., Simos, P., Papanicolaou, A., Fletcher, J.: Using structural equation modeling to assess functional connectivity in the Brain: power and sample size considerations. *Educ. Psychol. measurement* **74**(5), 733–758 (2014). doi:[10.1177/0013164414525397](https://doi.org/10.1177/0013164414525397)

- Simonson, I.: The effect of purchase quantity and timing on variety-seeking behavior. *J. Market. Res.* **27**(2), 150–162 (1990). doi:[10.2307/3172842](https://doi.org/10.2307/3172842)
- Smyth, B., McClave, P.: Similarity versus diversity. In: Aha D., Watson I. (eds.) *Case-Based Reasoning Research and Development. Lecture Notes in Computer Science* 2080. pp. 347–61. Springer, Berlin (2001) http://link.springer.com/chapter/10.1007/3-540-44593-5_25
- Willemsen, M.C., Knijnenburg, B.P., Graus, M.P., Velter-Bremmers, L.C.M., Fu, K.: Using latent features diversification to reduce choice difficulty in recommendation lists. In: *RecSys'11 Workshop on Human Decision Making in Recommender Systems*, vol. 811, pp. 14–20. CEUR-WS, Chicago, IL (2011) <http://ceur-ws.org/Vol-811/paper3.pdf>
- Wolf, E.J., Harrington, K.M., Clark, S.L., Miller, M.W.: Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety. *Educ. Psychol. Measurement* **76**(6), 913–934 (2013). doi:[10.1177/0013164413495237](https://doi.org/10.1177/0013164413495237)
- Ziegler, C-N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: *Proceedings of the 14th International Conference on World Wide Web*, pp. 22–32. WWW '05, ACM, Chiba (2005). doi:[10.1145/1060745.1060754](https://doi.org/10.1145/1060745.1060754)

Martijn C. Willemsen is Professor of Human Decision Making in Interactive Systems in the Human-Technology Interaction Group at Eindhoven University of Technology. He received a M.Sc. in Technology and Society and a MSc. In Electrical Engineering from Eindhoven University of Technology and a Ph.D. on the topic of Human Decision Making from the same university. He specialized in process tracing methods during a one year post-doc at Columbia University, NY. His primary interests lie in the understanding of cognitive processes of decision making by means of process tracing and in the application of decision making theory in interactive systems such as recommender systems. Dr. Willemsen has published over 30 papers in high ranking journals and conferences in Psychology and Computer Science in the areas of decision making, process tracing and recommender systems.

Mark P. Graus is a Data Science Ph.D. candidate in the Human-Technology Interaction Group at Eindhoven University of Technology. He received his B.Sc. in Innovation Sciences from Eindhoven University of Technology and his M.Sc. degree in Human-Technology Interaction from the same university. His primary interests lie in the combination of psychological theory and data mining. Previous work includes papers published in the fields of recommender systems (RecSys) and website personalization and adaptation (UMAP).

Bart P. Knijnenburg is Professor of Human-Centered Computing at Clemson University. Dr. Knijnenburg received his B.S. degree in Innovation Sciences and his M.S. degree in Human-Technology Interaction from Eindhoven University of Technology, his M.A. in Human-Computer Interaction from Carnegie Mellon University, and his Ph.D. in Information and Computer Sciences from the University of California, Irvine. Dr. Knijnenburg works on privacy decision-making and user-centric evaluation of adaptive systems. He has published over 20 papers in high-ranking journals and conferences in the areas of privacy, recommender systems, and human-computer interaction.