How Users Ride the Carousel: Exploring the Design of Multi-List Recommender Interfaces From a User Perspective

Benedikt Loepp
University of Duisburg-Essen
Duisburg, Germany
benedikt.loepp@uni-due.de

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

Multi-list interfaces are widely used in recommender systems, especially in industry, showing collections of recommendations, one below the other, with items that have certain commonalities. The composition and order of these "carousels" are usually optimized by simulating user interaction based on probabilistic models learned from item click data. Research that actually involves users is rare, with only few studies investigating general user experience in comparison to conventional recommendation lists. Hence, it is largely unknown how specific design aspects such as carousel type and length influence the individual perception and usage of carousel-based interfaces. This paper seeks to fill this gap through an exploratory user study. The results confirm previous assumptions about user behavior and provide first insights into the differences in decision making in the presence of multiple recommendation carousels.

CCS CONCEPTS

• **Human-centered computing** → **User interface design**; • **Information systems** → **Recommender systems**.

KEYWORDS

Carousel interfaces, Choice overload, Multi-list recommendations, User Experience.

ACM Reference Format:
Benedikt Loepp and Jürgen Ziegler. 2023. How Users Ride the Carousel: Exploring the Design of Multi-List Recommender Interfaces From a User Perspective. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3604915.3610638

1 INTRODUCTION

While one-dimensional lists dominated recommender systems (RS) for a long time, it has now become the de-facto standard to show multiple collections of recommendations [10, 27]. Each collection is displayed as a single row, often referred to as a "carousel" [2] or "shelf" [22], containing items with certain commonalities. In this way, Netflix displays various types of personalized movie recommendations, featuring genres, popular themes, and curated content. Spotify recommends new releases, podcasts on certain topics, and songs similar users are listening to. Each list comes with a descriptive label, based on which the type of a carousel can be defined according to the scheme of explanation styles proposed by Kouki et al. [16]: 1) Carousels where the label has a *user-based* style contain collaborative filtering results. 2) The *item-based* style describes carousels with items similar to those the user has rated positively in the past. 3) The *content-based* style uses metadata to highlight that the items match personal preferences. 4) The *social* style refers to the preferences of peers or friends. 5) Global *item popularity* is often used as a label for non-personalized carousels.

All these types are widely used in industry, but the empirical basis for the design of multi-list recommender interfaces (MLRI) is weak. The few existing studies share several limitations, such as focusing on the carousel order based on probabilistic models learned from item click data and simulating the corresponding user behavior [2, 6, 27]. As a result, user interaction is often not fully captured, e.g., scrolling and navigation, and assumptions are made that have not been tested with actual users, although it is known from conventional lists that the individual decision making can play a significant role in the recommendation process [3, 28]. This is especially a problem since carousels are usually composed and ordered in a personalized manner, but without considering design criteria such as carousel type and visible length, even though they may have a similar influence from the user perspective. The few existing *user* studies likewise pay little attention to the specific characteristics of the user and the presentation format, focusing instead on general user experience in comparison to lists and grids [10, 31]. Against this background, we conducted an exploratory user study ($N = 113$) with a prototypical carousel-based movie RS to address the following questions:

- RQ1 How do carousel type and visible length affect the perception of the system and the recommendations?
- RQ2 How do carousel type and visible length affect the interaction with a MLRI?
- RQ3 How do individual decision-making traits affect the subjective perception and the user interaction?

2 BACKGROUND AND RELATED WORK

From a user perspective, aspects such as control and transparency have been found to be at least as important for the success of RS as algorithmic accuracy [13, 20]. The presentation format, however, has not received the same attention, although it can have a strong impact on the user experience. The most influential study on this topic is that of Bollen et al., who investigated list length in relation to item diversity and choice overload [3]. The meta analysis by Scheibehenne et al. showed that the occurrence of choice overload depends on personal characteristics such as domain knowledge and

RecSys '23, September 18–22, 2023, Singapore, Singapore

decision-making strategy [28]. However, while MLRI have become the de-facto standard in real-world systems (cf. Section 1), these works were focused on conventional lists, with items sorted according to a single criterion, even if arranged as a grid. Some studies on critique-based RS can be seen as exceptions, with recommendations grouped by critiquing options [4]. This grouping, however, was intended to improve the critiquing process rather than to offer a multi-list view that facilitates decision making. Numerous other studies likewise have shown that recommendations can be visualized in a more informative and appealing way than in conventional lists [8, 17]. These academic examples, e.g., based on graphs or maps, are yet far too complex to apply them widely. This highlights the lack of empirical research on today's most popular presentation format, i.e., in carousels, especially under consideration of individual differences in decision making. In general, only few studies in the RS area have considered personal characteristics such as domain expertise, maximization behavior, or need for cognition—and only in terms of their influence on, e.g., the preferred level of control [11, 12], the perception of explanations [9, 23], or higher-level user behavior [14, 18].

So far, only a few (exploratory) user studies have focused on MLRI: Jannach et al. [10] conducted a large online experiment to study the effects of design alternatives on user behavior, providing first insights into the grouping of similar-item recommendations. Starke et al. [32] found that even though the carousels in their RS had descriptive labels, there were no positive effects on choice satisfaction or difficulty in comparison to a conventional grid without explanations. Other aspects that may affect the user experience, e.g., user or interface characteristics, were not considered. In contrast, the study presented in [31] examined the impact of explanation styles and corresponding algorithms. The labels did have an effect on the subjective assessment, but the study focused again on similar-item recommendations—in the recipe domain, with an interface as in the critique-based RS mentioned above, different from movie and music streaming platforms, with fewer recommendations and no personalization. The authors also asked about cooking experience, but found no effect for the different interfaces. Hence, neither this nor other studies help obtain a general understanding of the impact of carousel-specific design aspects, and whether individual decision making plays a similar role as in conventional lists (see above). Moreover, by focusing on subjective dimensions, other forms of user interaction than item clicks were largely ignored, especially scrolling and navigation, on which aspects such as the visible length of the carousels or their type could have a great effect. Finally, by presenting reference items, the interfaces were different from most real-world systems, where carousels are usually displayed independently on the landing page.

In this regard, it is worth noting that the positive effects of carousel-based interfaces have also been demonstrated in several publications from industry, where algorithmic solutions have been developed, e.g., to optimize the way carousels are put together, ordered, or labeled [2, 19, 22, 30, 33, 34]. However, most findings stem from comparisons against conventional lists, performed in online A/B tests or offline experiments, with metrics based purely on behavioral data.

Finally, it is important to mention the work of Dacrema et al. [5, 6] and Rahdari et al. [26, 27]: Inspired by studies on search interfaces, Dacrema et al. assumed that user behavior follows a "golden triangle," i.e., attention decreases linearly from the top-left corner [5]. From this, they extended the well-known NDCG metric, and showed that the algorithms behind the different carousel types perform differently when they are combined in a MLRI [6]. Rahdari et al. presented a carousel click model based on the assumption that before users start to examine the items, they explore vertically until they find a label that catches their attention [27]. In a simulation experiment, only with genre-based labels, their main finding was that users were more efficient than with a conventional list. The authors also proposed an approach to interactively control the importance of the topics represented by the carousels [26]. The absence of user studies, however, again points out that more research is needed to strengthen the empirical basis for designing multi-list views.

3 METHOD

To complement the few user studies on MLRI, we conducted an online experiment focusing on a domain typical for carousels, i.e., movies, and on the most frequently used form, i.e., without a reference item. In this way, we aimed to address our research questions, i.e., how the most fundamental yet underexplored design criteria for carousels, their type and visible length, affect the subjective assessment (RQ1) and the user interaction, including scrolling and navigation (RQ2). As we expected inter-individual differences, we also wanted to examine the role of specific decision-making practices (RQ3). Similar to previous studies on MLRI, we implemented the carousels in a web-based RS, thus requiring a laptop or desktop computer (see screenshot in the supplementary online material). Even though most real-world systems can be used with touch (e.g., Spotify app) or remote control (e.g., Netflix on a TV), we deliberately chose this modality to increase flexibility, to be able to record user behavior, and to ensure comparability. The study was approved by the ethics committee of our department.

3.1 Study design

We designed the study with a 3x2x2 mixed design. First, as a between-subjects factor, we decided to vary the *visible length* of the carousels L_n, with $n \in \{4, 6, 8\}$. We assumed that the total number of simultaneously presented recommendations would affect the perception of and the interaction with a MLRI. The condition with six items displayed per carousel was chosen to resemble popular movie and music streaming platforms. This number was also found to be similar to studies on the length of conventional recommendation lists [3, 35]. Participants were randomly assigned to the three resulting conditions. Second, based on the assumption that also the *type* of a carousel would affect the choice of and the satisfaction with an item, we considered the type T as a two-level within-subjects factor: The interface was either composed of homogeneous carousels (T_{hom}), i.e., each carousel had a content-based label with genre information, or heterogeneous carousels with different explanation styles (T_{het}), i.e., user-based, item-based, content-based (with references to tags, movie duration, or release year), and item popularity. Regardless of the type, personalization took place as described in the next section. There it is also explained how recommendations

were selected for inclusion in the carousels, depending on their type. Third, we added a within-subjects factor to control for the effect of the personalization: The *vertical order* of the carousels O was either randomized (O_{rand}) or based on the average prediction for the contained items (O_{pred}). These four within-subject conditions were randomly intermixed for each participant.

3.2 Prototype and carousel generation

We implemented the aforementioned carousel types in a prototypical movie RS (see screenshot in the supplementary material). For this, we extended our existing web-based system [21], which uses content-boosted matrix factorization [20] based on the *MovieLens 20M* and *Tag Genome* datasets, and displays recommendations based on metadata gathered from *The Movie Database* (TMDB). For each carousel, we created a list of 60 candidate items from the matrix factorization results: First, to obtain a larger set of personalized recommendations, we used the user-factor vector learned via online updating after an initial preference elicitation phase, in which participants had to rate 10 out of the most popular movies with 1 to 5 stars. Second, filtering or re-ranking took place: For carousels with a content-based explanation using genre information (as in T_{hom}, see above), we considered high-scoring items with `genre='scifi'`. For the item-based style (as in T_{het}), we re-ranked the items in terms of their latent factor similarity to one of the movies rated in the beginning. For the sake of space, we omit the details of the filtering/re-ranking process for the other carousel types in T_{het}.

Either way, for the final presentation, we selected 24 items for each carousel based on the Lin-20 method proposed in [3], and displayed them in random order. Which carousels were presented was determined randomly (O_{rand}) or by ordering them based on the average of the scores predicted for the contained items (O_{pred}), which was always possible because of the underlying matrix factorization. Six carousels were included in a single view. Similar to most real-world systems, we decided to display three of them at a time, with visible items depending on L. As suggested in [5], we made sure that each item was shown only once. However, across the four iterations per participant (for each level of T and O), items could appear multiple times. Due to the selection and randomization described above, and the variety of the presented carousel types, we did not expect this to have a strong effect on participants' decision making. Finally, labels were chosen in line with the explanation styles, using information from the underlying metadata dataset.

3.3 Procedure and task

After a brief introduction and the initial preference elicitation (see above), participants were assigned to one of the between-subjects conditions and then exposed to the MLRI four times, based on the levels of T and O (see Section 3.1). In each iteration, the task was to interact with the presented carousels to find a single movie worth watching. The visible length of the carousels was kept constant, based on the level of L. After selecting a movie, participants were briefly redirected to a questionnaire to rate the quality of this item and the selection process (see below). At the very end, they were shown the final part of the questionnaire, with more questions about the system and personal characteristics. If participants asked for study credit (see further below), a supervisor was present for the entire study via video call.

3.4 Questionnaire and interaction data

The questionnaire shown between and after the tasks was administered using the online tool *SosciSurvey*. Primarily, we used constructs (see Table 1 and 2, or supplementary material) from established RS evaluation frameworks [15, 25]. To assess the general user experience, we used the short UEQ [29]. Regarding personal characteristics, we collected demographics and asked participants about their domain knowledge using self-generated items (DK). We also considered several constructs that are often used in RS research to assess individual decision-making traits (see Section 2), and that we expected to play a role in the usage and perception of differently designed MLRI. Specifically, we included the short maximization scale (MAX) [24], the decision styles scale (DSS) [7], and the short scale for need for cognition (NFC) [1]. All items had 5-point Likert response scales, except for UEQ (7-point bipolar) and NFC (7-point Likert). We also measured task times and logged interaction data such as clicks on items as well as horizontal and vertical navigation.

3.5 Participants

We recruited 186 participants through personal contacts, via a student Facebook group, and online on LinkedIn. Some participants did not complete the questionnaire or did not use a laptop or desktop computer (which was required, see above), leaving us with a sample of $N = 113$. Age ranged from 18 to 63 ($M = 25.11$, $SD = 8.21$), 61 % were female, 37 % were male, and 2 participants did not indicate their gender. 63 % were students, the rest were employed (27 %), self-employed (3.5 %), or did not answer this question. Random assignment to the between-subjects conditions resulted in group sizes of $N_{L_4} = 36$, $N_{L_6} = 40$, and $N_{L_8} = 37$. Students from a specific degree program were rewarded with study credit.

4 RESULTS

4.1 Impact of carousel type and visible length on user perception (RQ1)

Table 1 and 2 show the questionnaire results for the assessment of system and recommendations. Based on our study design, we used 3x2x2 mixed ANOVAs to determine effects of L, T, and O. The first table shows the main effect of the between-subjects factor L. Given the exploratory nature of our experiment, we did not adjust for multiple comparisons. This may have inflated the type I error rate, but even without a correction, none of the differences exceeded the significance level of $\alpha = .05$. The ANOVAs also did not show significant interaction effects, except for system effectiveness, $F(2, 110) = 3.38$, $p = .038$, $\eta_p^2 = .06$, and, between T and O, for perceived transparency, $F(1, 110) = 13.63$, $p < .001$, $\eta_p^2 = .11$. The second table shows the within-subject comparison, which also did not yield significant results.

Since general user experience was only assessed once at the very end, we calculated one-factorial ANOVAs to compare the levels of L. We found a significant effect on the pragmatic UEQ subscale, $F(2, 110) = 3.48$, $p = .034$, $\eta_p^2 = .06$. The highest scores were obtained in L_4 ($M = 1.71$, $SD = 0.97$), followed by L_6 ($M = 1.25$, $SD = 1.03$)

Table 1: Estimated marginal means and standard errors for the between-subjects comparison. Higher values indicate better results for the questionnaire constructs (difficulty and effort are reversed accordingly), with best values highlighted in bold. The last three columns show the mixed ANOVA results for the main effect of L ($df_1 = 2$, $df_2 = 110$), with η_p^2 representing effect size.

	Overall		L_4		L_6		L_8				
Construct / Measurement	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*F*	*p*	η_p^2
Perc. recommendation quality [15]	3.49	0.07	**3.52**	0.13	3.44	0.12	**3.52**	0.12	0.15	.865	.00
Perc. recommendation diversity [15]	3.73	0.07	**3.81**	0.13	3.76	0.12	3.62	0.12	0.63	.534	.01
Choice satisfaction [15]	4.05	0.06	4.08	0.10	3.95	0.09	**4.11**	0.10	0.80	.452	.01
Choice difficulty [15]	3.13	0.08	3.12	0.15	2.98	0.14	**3.30**	0.14	1.32	.273	.02
Perc. system effectiveness [15]	3.41	0.08	3.31	0.15	3.46	0.14	**3.47**	0.14	0.39	.681	.01
Perc. usage effort [15]	3.74	0.07	**3.80**	0.12	3.69	0.11	3.74	0.11	0.20	.821	.00
Perc. transparency [25]	3.30	0.09	**3.56**	0.16	3.03	0.15	3.34	0.16	3.00	.054	.05
Overall satisfaction [25]	3.61	0.08	3.62	0.14	3.57	0.13	**3.63**	0.14	0.06	.944	.00
Number of navigation actions	8.06	0.62	11.13	1.10	8.23	1.04	4.89	1.08	8.20	<.001	.13
Number of clicked items	1.72	0.12	1.76	0.21	1.68	0.20	1.73	0.20	0.05	.951	.00
Number of viewed items	50.40	2.31	46.36	4.10	52.99	3.89	51.84	4.04	0.77	.465	.01
Task completion time (sec.)	80.50	6.48	81.56	11.47	88.16	10.90	71.80	11.32	0.55	.580	.01

and L_8 ($M = 1.11$, $SD = 1.05$). A Tukey post-hoc test indicated a difference between L_4 and L_8 ($p = .035$). In contrast, we did not observe significant effects on the hedonic subscale, $F(2, 110) = 0.56$, $p = .572$, or the overall score, $F(2, 110) = 1.23$, $p = .283$.

4.2 Impact of carousel type and visible length on user interaction (RQ2)

Table 1 also shows the between-subjects comparison of the interaction data. Since we found no within-subject differences ($p \gg .05$, small effect sizes), we omit these data in Table 2 for brevity. There were also no interaction effects. However, as shown in the first table, we found a significant main effect on the number of navigation actions, i.e., clicks made to scroll (horizontally or vertically) differed depending on L. A Tukey post-hoc test indicated that participants performed more actions in L_4 than in L_8 ($p < .001$). To analyze the exploration behavior in more depth, we examined the effect of the absolute position of an item on its selection using a multiple linear regression. The result was significant, $F(2, 141) = 84.99$, $p < .001$, with a high amount of explained variance (adjusted R^2 of 0.54). The standardized coefficients show that the selection frequency decreased horizontally ($\beta_{hor} = -.416$) more slowly than vertically ($\beta_{vert} = -.611$), i.e., participants were more likely to find a movie by delving into the carousels rather than by viewing a greater number.

4.3 Impact of individual decision-making traits (RQ3)

To obtain a first understanding of whether common decision-making traits play a similar role in MLRI as in conventional lists, we calculated Pearson correlations for all questionnaire constructs. While we did not observe significant correlations with NFC or DSS, we found a number of (small) effects for MAX: Maximizers perceived recommendation quality, $r(111) = .25$, $p = .007$, system effectiveness, $r(111) = .22$, $p = .018$, and transparency, $r(111) = .22$, $p = .018$, to be higher, but found it more difficult to choose an item, $r(111) = -.22$, $p = .021$, as expected. Regarding the interaction data in the lower part of Table 1, we found no significant correlations. However, when we took a closer look at the navigation actions, we again observed a significant effect for MAX: Maximizers scrolled down more often, exploring carousels that were initially hidden, $r(111) = .22$, $p = .017$. While this correlation was small, it should be noted that most participants did not navigate vertically at all (57.5 %), and that we required them to use buttons to scroll (see screenshot in the online material). To analyze the different carousel types with respect to the frequency of item selection, we used median splits to classify participants into low and high groups, and ran chi-square tests. We found only few significant differences. In the T_{het} conditions, the relationship between NFC and the selection from carousels with a content-based label was significant, $\chi^2(1, N = 113) = 12.91$, $p < .001$, $\phi = .34$. Participants with high NFC were more likely to use this carousel type. In contrast, participants with low NFC were more likely to select items from carousels with a user-based explanation, $\chi^2(1, N = 113) = 6.36$, $p = .012$, $\phi = .24$. We also found significant effects of gender (e.g., males used carousels with romantic movies less often), but omit reporting them here because demographics were not part of our hypotheses.

5 DISCUSSION AND CONCLUSIONS

While this initial study was intended to be exploratory, it is first worth noting that it may have been slightly underpowered: The final sample was smaller than the size of 156 suggested by an a-priori power analysis to detect medium between-subjects effects (power of .80, $\alpha = .05$). This is also the reason why we left structural equation modeling for future studies. Moreover, it may be difficult to generalize our findings, as the study was limited to a single (but typical) domain, based on a prototypical web-based RS. However, this is also true for previous studies, and just emphasizes the need for experiments under more common conditions, i.e., on a smartphone or TV, as well as in other domains and with other datasets. Nevertheless, we are certain that the results, even if some of them seem intuitive, provide valuable evidence about the impact of design decisions for MLRI.

With respect to RQ1, it seems that neither the composition nor the order of the carousels played a significant role (cf. Section 4.1). We did not find effects of the within-subject factors T and O, even without correcting for multiple comparisons. On the other hand, the differences in the interface might have been too small, especially for the study situation with lower user engagement due to the artificial task. If this is true, however, this also holds for the labels. In fact, participants seemed to look at the interface as a whole, in line with previous research, where few differences were found

Table 2: Estimated marginal means, standard errors, and ANOVA results for the within-subject comparison ($df_1 = 1$, $df_2 = 110$).

	T_{hom}		T_{het}					O_{rand}		O_{pred}				
Construct	*M*	*SE*	*M*	*SE*	*F*	*p*	η_p^2	*M*	*SE*	*M*	*SE*	*F*	*p*	η_p^2
Perc. recommendation quality [15]	**3.50**	0.08	3.49	0.08	0.03	.854	.00	3.45	0.08	**3.54**	0.08	1.99	.161	.02
Perc. recommendation diversity [15]	**3.77**	0.07	3.69	0.08	1.60	.209	.01	**3.79**	0.08	3.68	0.08	3.16	.078	.03
Choice satisfaction [15]	4.02	0.07	**4.08**	0.07	0.46	.499	.00	**4.05**	0.07	4.04	0.06	0.04	.848	.00
Choice difficulty [15]	**3.16**	0.10	3.11	0.09	0.27	.603	.00	3.13	0.09	**3.14**	0.10	0.01	.940	.00
Perc. system effectiveness [15]	**3.42**	0.09	3.40	0.09	0.38	.541	.00	3.39	0.09	**3.43**	0.09	0.39	.532	.00
Perc. usage effort [15]	**3.75**	0.07	3.74	0.07	0.04	.836	.00	**3.75**	0.07	3.74	0.07	0.10	.752	.00
Perc. transparency [25]	3.29	0.10	**3.34**	0.10	0.49	.484	.00	3.25	0.10	**3.67**	0.10	2.32	.131	.02
Overall satisfaction [25]	3.60	0.08	**3.62**	0.09	0.11	.738	.00	3.59	0.08	**3.62**	0.08	0.17	.683	.00

between grid- and carousel-based interfaces [32]. Consequently, it appears necessary to investigate how labels can provide a greater benefit than the current simple explanation styles, in particular, in situations of actual use.

The results for the between-subjects factor L suggest that the visible length of the carousels was also of limited importance. On the other hand, in L_6, the questionnaire scores were always the lowest. Thus, given the widespread lack of interaction effects with T and O, the tendencies shown in Table 1 could also mean that presenting six items per carousel (as in most real-world systems) is a good compromise between choice overload and the desire for exploration.

The interaction analysis yielded results generally in line with the questionnaire data, thus providing a first answer to RQ2 (cf. Section 4.2). However, it seems that the significant effect of L on the number of navigation actions was not reflected in participants' perception. Of course, it should be noted that horizontal scrolling was naturally limited in L_6 and L_8, since 24 was the maximum number of carousel items in all conditions. Nevertheless, it seems that with the few visible items in L_4, participants had to invest more scrolling effort to reach the same level of satisfaction. Hence, in line with Bollen et al. [3], we assume that a smaller number of visible carousel items reduces cognitive load (as confirmed by the higher pragmatic quality in the UEQ, as well as by qualitative feedback, here not reported for the sake of space), but this is offset by the interaction effort to explore the still hidden items. However, a larger sample would be needed to investigate the causal relations associated with choice overload. So far, only tendencies can be observed in Table 1, e.g., longer carousels seem to be more effective, making it easier to decide.

As the first user study to examine the interaction with MLRI more completely, the analysis also confirmed some of the assumptions made in earlier works (cf. Section 4.2): The position of an item predicts well if it gets selected, with linearly decreasing probability, as suggested by the NDCG2D metric in [5]. In contrast, we did not observe the behavior suggested by the model in [27]. However, this finding should be taken with a grain of salt, as vertical scrolling was not as prominently featured in our system as horizontal exploration. Nevertheless, from a practical perspective, it shows that it may be worth focusing more on carousel pagination rather than on adding carousels or improving their order. In this context, also the role of each carousel needs to be better understood, e.g., to answer the question whether behavior is different when label and recommended items of the first carousel immediately match the user's needs.

Finally, in response to RQ3, it seems valid to conclude that there are indeed inter-individual differences in the perception and usage of MLRI (cf. Section 4.3). Some carousel types seemed to be more suitable for participants who enjoyed cognitive thinking, while others, e.g., based on collaborative filtering, were easier to approach with an average interest. Admittedly, for most constructs, we did not find significant effects, and the results may have been influenced by the display frequency of the carousels, which depended on the initial preference elicitation (cf. Section 3.3). In addition, the sample may not have been diverse enough to adequately capture all user characteristics. Due to its size, it was also not yet possible to examine potential interaction effects and differences between conditions. Still, while one of the most prominent advantages of MLRI is their ability to serve different contexts at the same time, our findings show that they cannot be seen as a one-fits-all solution. Thus, a better understanding of individual decision making, and, subsequently, better models of user behavior, are needed to further improve carousel-based RS from a user perspective.

ACKNOWLEDGMENTS

To Jens Kohlmann, who conducted this study as part of his Master's thesis. To Yuan Ma, who helped with the presentation.

REFERENCES

[1] Hanna Beißert, Meike Köhler, Marina Rempel, and Constanze Beierlein. 2015. Deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition NFC-K. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)* (2015).

[2] Walid Bendada, Guillaume Salha, and Théo Bontempelli. 2020. Carousel Personalization in Music Streaming Apps with Contextual Bandits. In *RecSys '20: Proceedings of the 14th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 420–425.

[3] Dirk Bollen, Bart P. Knijnenburg, Martijn C. Willemsen, and Mark P. Graus. 2010. Understanding Choice Overload in Recommender Systems. In *RecSys '10: Proceedings of the 4th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 63–70.

[4] Li Chen and Pearl Pu. 2012. Critiquing-Based Recommenders: Survey and Emerging Trends. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 125–150.

[5] Nicolò Felicioni, Maurizio Ferrari Dacrema, and Paolo Cremonesi. 2021. Measuring the User Satisfaction in a Recommendation Interface with Multiple Carousels. In *IMX '21: Proceedings of the 2nd ACM International Conference on Interactive Media Experiences*. ACM, New York, NY, USA, 212–217.

[6] Maurizio Ferrari Dacrema, Nicolò Felicioni, and Paolo Cremonesi. 2022. Offline Evaluation of Recommender Systems in a User Interface With Multiple Carousels. *Frontiers in Big Data* 5 (2022), 910030:1–910030:21.

[7] Katherine Hamilton, Shin-I Shih, and Susan Mohammed. 2016. The Development and Validation of the Rational and Intuitive Decision Styles Scale. *Journal of Personality Assessment* 98, 5 (2016), 523–535.

[8] Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive Recommender Systems: A Survey of the State of the Art and Future Research Challenges and Opportunities. *Expert Systems with Applications* 56 (2016), 9–27.

[9] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2020. Explaining Review-Based Recommendations: Effects of Profile Transparency, Presentation Style and User Characteristics. *i-com – Journal of Interactive Media* 19, 3 (2020), 181–200.

[10] Dietmar Jannach, Mathias Jesse, Michael Jugovac, and Christoph Trattner. 2021. Exploring Multi-List User Interfaces for Similar-Item Recommendations. In *UMAP '21: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 224–228.

[11] Yucheng Jin, Nava Tintarev, Nyi Nyi Htun, and Katrien Verbert. 2020. Effects of Personal Characteristics in Control-Oriented User Interfaces for Music Recommender Systems. *User Modeling and User-Adapted Interaction* 30, 2 (2020), 199–249.

[12] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of Personal Characteristics on Music Recommender Systems with Different Levels of Controllability. In *RecSys '18: Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 13–21.

[13] Michael Jugovac and Dietmar Jannach. 2017. Interacting with Recommenders – Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems* 7, 3 (2017), 10:1–10:46.

[14] Timm Kleemann, Magdalena Wagner, Benedikt Loepp, and Jürgen Ziegler. 2021. Modeling User Interaction at the Convergence of Filtering Mechanisms, Recommender Algorithms and Advisory Components. In *Mensch & Computer 2021 – Tagungsband*. ACM, New York, NY, USA, 531–543.

[15] Bart P. Knijnenburg, Martijn C. Willemsen, and Alfred Kobsa. 2011. A Pragmatic Procedure to Support the User-Centric Evaluation of Recommender Systems. In *RecSys '11: Proceedings of the 5th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 321–324.

[16] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 379–390.

[17] Johannes Kunkel and Jürgen Ziegler. 2023. A Comparative Study of Item Space Visualizations for Recommender Systems. *International Journal of Human-Computer Studies* 172 (2023).

[18] Yu Liang and Martijn C. Willemsen. 2021. The Role of Preference Consistency, Defaults and Musical Expertise in Users' Exploration Behavior in a Genre Exploration Recommender. In *RecSys '21: Proceedings of the 15th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 230–240.

[19] Chieh Lo, Hongliang Yu, Xin Yin, Krutika Shetty, Changchen He, Kathy Hu, Justin M. Platz, Adam Ilardi, and Sriganesh Madhvanath. 2021. Page-Level Optimization of E-Commerce Item Recommendations. In *RecSys '21: Proceedings of the 15th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 495–504.

[20] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2019. Interactive Recommending with Tag-Enhanced Matrix Factorization (TagMF). *International Journal of Human-Computer Studies* 121 (2019), 21–41.

[21] Benedikt Loepp and Jürgen Ziegler. 2019. Towards Interactive Recommending in Model-Based Collaborative Filtering Systems. In *RecSys '19: Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 546–547.

[22] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. 2018. Explore, Exploit, and Explain: Personalizing Explainable Recommendations with Bandits. In *RecSys '18: Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 31–39.

[23] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To Explain or Not to Explain: The Effects of Personal Characteristics When Explaining Music Recommendations. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 397–407.

[24] Gergana Y. Nenkov, Maureen Morrin, Andrew Ward, Barry Schwartz, and John Hulland. 2008. A Short Form of the Maximization Scale: Factor Structure, Reliability and Validity Studies. *Judgment and Decision Making* 3 (2008), 371–388.

[25] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender Systems. In *RecSys '11: Proceedings of the 5th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 157–164.

[26] Behnam Rahdari, Peter Brusilovsky, and Alireza Javadian Sabet. 2021. Controlling Personalized Recommendations in Two Dimensions with a Carousel-Based Interface. In *IntRS '21: Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*. 112–122.

[27] Behnam Rahdari, Branislav Kveton, and Peter Brusilovsky. 2022. The Magic of Carousels: Single vs. Multi-List Recommender Systems. In *HT '22: Proceedings of the 33rd ACM Conference on Hypertext and Social Media*. ACM, New York, NY, USA, 166–174.

[28] Benjamin Scheibehenne, Rainer Greifeneder, and Peter M. Todd. 2010. Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload. *Journal of Consumer Research* 37, 3 (2010), 409–425.

[29] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4, 6 (2017), 103–108.

[30] Sanidhya Singal, Piyush Singh, and Manjeet Dahiya. 2021. Automatic Collection Creation and Recommendation. In *RecSys '21: Proceedings of the 15th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 633–638.

[31] Alain D. Starke, Edis Asotic, and Christoph Trattner. 2021. "Serving Each User": Supporting Different Eating Goals Through a Multi-List Recommender Interface. In *RecSys '21: Proceedings of the 15th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 124–132.

[32] Alain D. Starke, Justyna Sedkowska, Mihir Chouhan, and Bruce Ferwerda. 2022. Examining Choice Overload across Single-List and Multi-List User Interfaces. In *IntRS '22: Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*.

[33] Chao-Yuan Wu, Christopher V. Alvino, Alexander J. Smola, and Justin Basilico. 2016. Using Navigation to Improve Recommendations in Real-Time. In *RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 341–348.

[34] Liang Wu, Mihajlo Grbovic, and Jundong Li. 2021. Toward User Engagement Optimization in 2D Presentation. In *WSDM '21: Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, USA, 1047–1055.

[35] Qian Zhao, Shuo Chang, F. Maxwell Harper, and Joseph A. Konstan. 2016. Gaze Prediction for Recommender Systems. In *RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 131–138.