Comparing User Interfaces for Customizing Multi-Objective Recommender Systems

Patrik Dokoupil[1], Ludovico Boratto[2] and Ladislav Peska[1]

[1]Faculty of Mathematics and Physics, Charles University, Prague, Czechia
[2]University of Cagliari, Italy

Abstract

The goal of Multi-Objective Recommender Systems (MORSs) is to adapt to the needs and preferences of the users from different beyond-accuracy perspectives. When a MORS operates at the *local* level, it tailors its results to the needs of each individual user. Recent studies have highlighted that, however, the self-declared propensity of the users towards the different objectives does not always match with the characteristics of the accepted recommendations. Therefore, in this study, we delve into different ways for users to express their preference toward multi-objective goals and observe whether they have some impact on declared propensities and overall user satisfaction. In particular, we explore four different user interface (UI) designs and perform a user study focused on the interactions with both the UI and the recommendations. Results show that multiple UIs lead to similar results w.r.t. usage statistics, but users' perceptions of these UIs often differ. These results highlight the importance of examining MORSs from multiple perspectives to accommodate the users' actual needs when producing recommendations. Study data and detailed results are available from https://osf.io/pbd54/.

Keywords

Multi-objective recommender systems, User study, Recommender systems UI

1. Introduction

Multiple-Objective Recommender Systems (MORSs) produce results that account for the effectiveness perspective but also go beyond it so as to tackle perspectives such as novelty, diversity, and fairness (to name a few) [1]. The optimization for these objectives can happen at the *aggregate* level so that the system can guarantee certain properties (e.g., all providers receive a certain exposure in the recommendation lists). Another alternative is to build MORSs that operate at the *local* (*individual*) level so as to shape results towards the prominence of different goals for individual users (e.g., each user would receive recommendations with a different level of diversity) [2]. Having *local* MORS, one may aim to provide users with additional control over the recommendations and allow them to set their propensities towards individual objectives [3]. This is in line with the general trend of a growing need for understanding and control over recommendations as illustrated, e.g., by the recent EU's Digital Services Act[1]. The regulation requires that the main driving forces of the recommendation process are disclosed and also that users should be allowed to select their preferred options (Article 27). However, recent literature revealed a mismatch between the self-declared propensity of users for the different objectives and the characteristics of the recommendations they accept (i.e., the items they choose among the recommendations are less novel or diverse than what they believe they would like) [3].

In this work, we explore the issue of self-declared propensities from the perspective of UI design. Specifically, we focused on a widely used combination of relevance, diversity, and novelty criteria and designed four different UIs that allow users to express their propensity toward the objectives. In a user study (Section 3), we allowed users to interact with both the UI and the recommendations themselves and evaluated the impact of different *customization UIs*. In particular, we observed whether the UI

IntRS'24: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, October 18, 2024, Bari (Italy)
patrik.dokoupil@matfyz.cuni.cz (P. Dokoupil); ludovico.boratto@acm.org (L. Boratto); ladislav.peska@matfyz.cuni.cz (L. Peska)
0000-0002-1423-628X (P. Dokoupil); 0000-0002-6053-3015 (L. Boratto); 0000-0001-8082-4509 (L. Peska)

[1] https://eur-lex.europa.eu/eli/reg/2022/2065/oj

designs affect how users perceive individual objectives, how they interact with recommendations, and whether there is some impact on perceived recommendation quality and overall satisfaction.

Our results (Section 4) show that there is a trade-off between the perceived usability of the different UIs and their effectiveness at indicating user propensity. Moreover, no UI has clearly shown to be the most effective, as users exploited three of our designs with similar effectiveness.

2. Background and Related Work

2.1. Customization UIs in Recommenders

We are not aware of any previous studies focusing on the comparison of UI designs for *local* MORS setting. However, in the context of MORS, the work that most closely aligns with ours is by Harper et al. [4], where the authors propose an algorithm allowing users to control item popularity and recency. In addition to the algorithm and its offline evaluation, the authors conducted a user study in the movie domain, finding that the tuned recommendations were rated more positively by users. They also highlighted the importance of individual-level optimization, as no single global setting worked equally well for all users. The tuning was done using buttons labeled neutrally as "left" and "right." While this choice was intentional and justified, users responded negatively when asked about the ease of use of the tuning interface. Therefore, we focused on different UI designs for RS tuning in our work.

Several additional UI variants were considered for value setting in RS as well as other HCI tasks [5, 3, 6, 7]. In web design praxis, sliders are considered to be a primary design choice for values specification as long as these do not have to be very precise [7]. This is well-reflected in UIs used for RS tuning, as illustrated, e.g., by the work of Liang and Willemsen [5] on tuneable exploration-oriented music RS. Similarly, sliders were also used in [3] for the customization of *local* MORS. Nonetheless, some researchers pointed out the slider's inferior performance (e.g., w.r.t. response times) in situations with limited options and advocated standard HTML radio buttons instead [6].

However, unlike in other scenarios, the particular value of a MORS objective does not carry an inherent meaning for the user (compared, e.g., to a price setting in an e-shop's faceted search). Therefore, users can only perceive the values relative to their previous settings (i.e., incremental increase/decrease; also denoted as *"relative"* in the literature) or through the comparison with the weights of other criteria (also denoted as *"absolute"* [8]). Naturally, UIs can be tailored to better reflect one of these views. Another open question is the optimal level of response granularity [9] so that the task complexity is minimized while the UI expressive power is still sufficient.

From these points of view, we can understand *sliders* as fine-grained UI collecting absolute feedback. To cover other design options, we propose and evaluate two alternatives to the *sliders* UI. In *options* UI, we provide users with several prompts to *relatively* increase/decrease the importance of individual criteria (as such, coarse-grained feedback with relative answers is received). In a way, this layout is most similar to the left/right buttons described in [4], but without the obfuscated labeling. The plus-minus *buttons* UI is inspired by common RPG gaming designs for character stats and, as such, provides coarse-grained absolute feedback. Finally, in [3], authors reported on an extensive over-weighting of beyond-relevance criteria by the users, and so the particular interpretation of user-provided weights can be questioned as well. This was the main driving force for *sliders_shifted* UI variant, which reduces the weights of beyond-relevance criteria.

2.2. Objectives in MORS

MORS typically aim to balance recommendations' relevance with various beyond-accuracy objectives, including diversity, novelty, serendipity, or fairness [10]. Out of the available options, we adopted the approach from [3], focusing on the following variants of relevance, novelty, and diversity.

- Estimated relevance rel of recommendation list L was set to the mean of estimated relevance scores ($\hat{r}_{u,i}$) predicted by the relevance-only baseline: $rel(L) = \frac{1}{|L|} \sum_{i \in L} \hat{r}_{u,i}$.

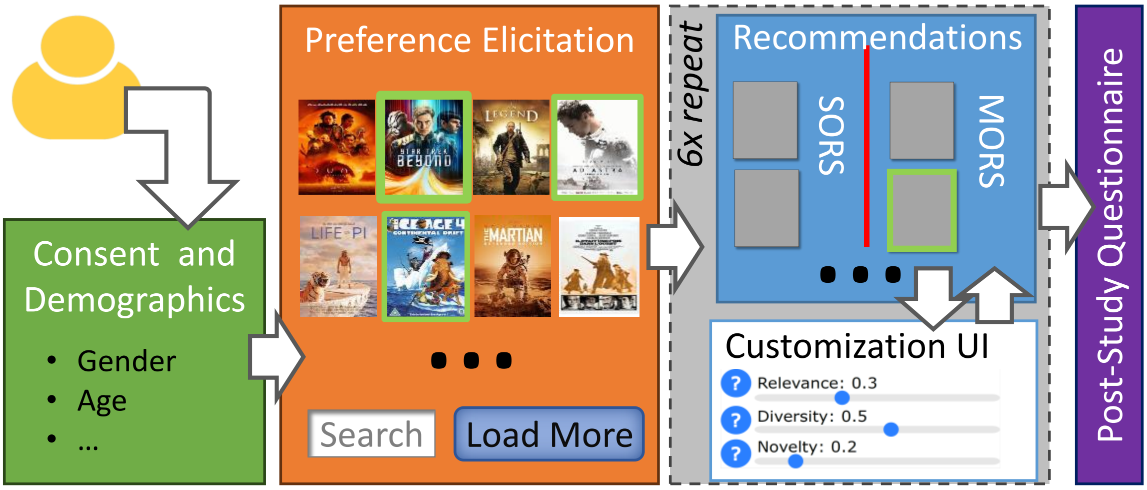


Figure 1: Schema of the study flow. Informed consent and basic demographics are required first, followed by preference elicitation. Then, participants are directed to in total 6 iterations of recommendations, where single-objective and multi-objective results are displayed side-by-side (i.e., within-user variable). After each iteration, users may modify their propensities towards individual objectives via a designated GUI (between-user variable). Finally, users are directed to the post-study questionnaire.

- Novelty was defined as mean popularity complement: $nov(L) = \frac{1}{|L|}\sum_{i \in L}\left(1 - \frac{|u \in U : r_{u,i} \text{ exists}|}{|U|}\right)$, where $r_{u,i}$ is the feedback of user u on item i and U is the set of all users.
- Diversity was set to collaborative intra-list diversity: *CF-ILD*$(L) = \frac{1}{|L|*(|L|-1)}\sum_{\forall i,j \in L, i \neq j} d(i,j)$, where $d(i,j)$ is cosine similarity on items' ratings.

3. User Study

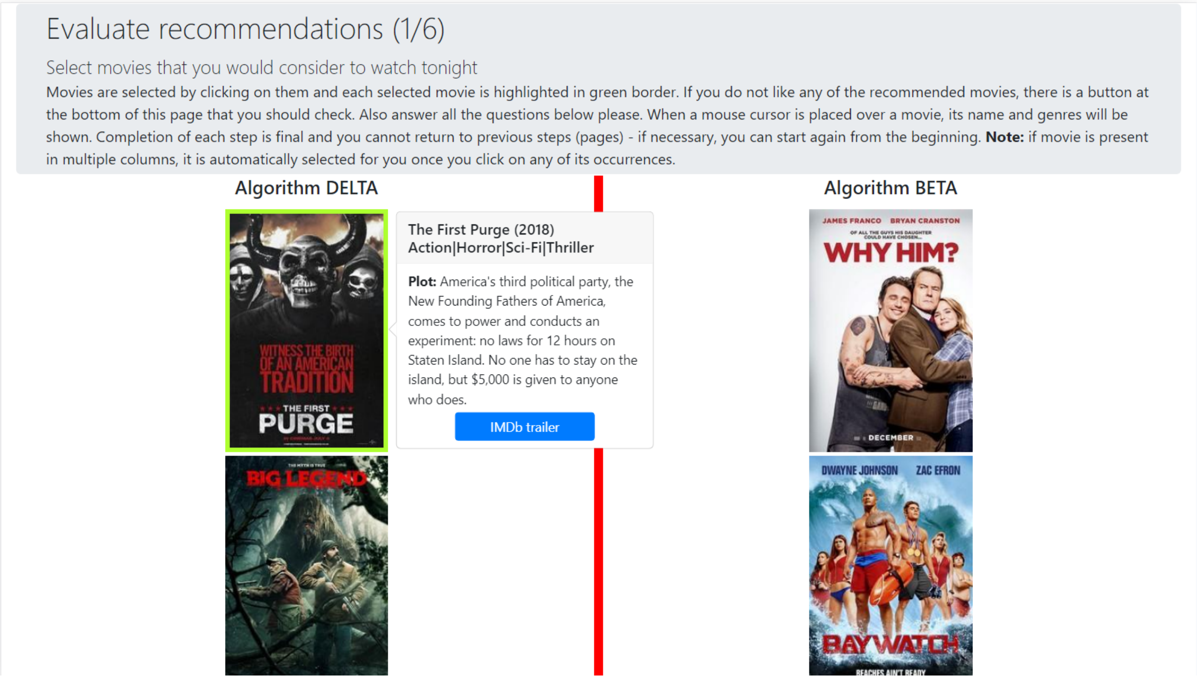
The study was conducted on a movies domain using the EasyStudy framework [11] and the experimental setup was largely based on [3]. In particular, we utilized the same filtered MovieLens Latest [12] dataset, preference elicitation process, objective criteria definitions, items presentation, and task definition. In the rest of this section, we provide details about the data pre-processing, describe the study flow, and specify the *customization UI* variants we evaluated.

3.1. Dataset

For the purpose of the study, we utilized an augmented version of the MovieLens-Latest dataset [12]. The dataset was selected for its relative novelty and the general familiarity and popularity of the movie domain among the general public. Both factors should contribute to the realisticness of the study. The dataset was utilized both to train the collaborative filtering algorithms and as a starting point to gather necessary item metadata. As the feedback collected during the study was binary, we binarized the dataset as well (4* and above counts as positive). Furthermore, we only considered the more recent and less obscure portion of the data. In particular, we removed movies released before 1990, ratings older than 2010, movies that have less than 50 ratings per year, and users with less than 100 ratings. This resulted in 9K users, 2K movies, and 1.5M ratings. In order to properly visualize the items, additional metadata were collected from respective IMDb profiles: movie descriptions, posters, and links to movie trailers.

3.2. Study flow

The user study was organized in four phases: *informed consent*, *preference elicitation*, *recommendation comparison*, and *post-study questionnaire*. The schematic of the study flow is depicted in Figure 1, while



Evaluate recommendations (1/6)

Select movies that you would consider to watch tonight

Movies are selected by clicking on them and each selected movie is highlighted in green border. If you do not like any of the recommended movies, there is a button at the bottom of this page that you should check. Also answer all the questions below please. When a mouse cursor is placed over a movie, its name and genres will be shown. Completion of each step is final and you cannot return to previous steps (pages) - if necessary, you can start again from the beginning. **Note:** if movie is present in multiple columns, it is automatically selected for you once you click on any of its occurrences.

The First Purge (2018)
Action|Horror|Sci-Fi|Thriller

Plot: America's third political party, the New Founding Fathers of America, comes to power and conducts an experiment: no laws for 12 hours on Staten Island. No one has to stay on the island, but $5,000 is given to anyone who does.

Figure 2: Screenshot of a recommendation iteration. Instructions and the head of the two lists are visible, as well as a description of the movie with a mouse hover focus.

the detailed description of individual steps follows.

3.2.1. Pre-study

Prior to the study commencement, users were shown a study mission statement and detailed instructions and were asked for informed consent on the publication of anonymized data. Since all participants were recruited using the Prolific service, we relied on the demographic information participants submitted there and did not ask participants for additional demographics.

3.2.2. Preference elicitation

After the initial step, participants were routed to the preference elicitation page to collect information to train the collaborative recommenders. In this phase, 24 movies were displayed to the users, asking them to select movies they previously watched and liked. The displayed movies were sampled from the dataset using the following procedure. We calculated estimated relevance (w.r.t. average user profile), novelty (w.r.t. item's mean popularity complement [13]), and diversity (w.r.t. CF-ILD [14]) characteristics for each item in the dataset. For each characteristic, we divided items into "low" and "high" buckets and sampled four items from each bucket.[2]

The displayed items were organized in a grid, and each item was represented by its poster image, title, and genres. Users could express their preferences by simply clicking on the ones they watched and liked before. We recommended selecting at least 5-10 items, but users were allowed to continue even if fewer items were selected. To make the elicitation more thorough, users could dynamically load additional items (repeating the procedure above) or search for specific ones via a text prompt.

After the elicitation phase, we estimated the initial user's propensities of users toward individual objectives based on the normalized marginal gains calculated for each selected vs. each displayed movie - i.e., as compared to all displayed movies, how much were the ones selected by the user relevant/novel/diverse. Please see [3, 15] for more details on the procedure.

[2] Note that we first sampled from the relevance and novelty buckets and only then calculated the diversity w.r.t. already selected items.

3.2.3. Recommendation comparison

The main part of the study comprised six iterations, where users received two lists of top-10 recommendations side-by-side. One list of recommendations was supplied by the relevance-only baseline algorithm, while the other was provided by the multi-objective RS. In particular, we employed generalized matrix factorization as a relevance-only baseline.[3] Note the relevance-only baseline will also be referred to as single-objective RS or simply SORS in the results. For the multi-objective RS, we employed the RLprop algorithm [15] aiming to maintain the proportionality between user-defined propensities and the fraction of individual objectives in the results. The considered relevance, novelty, and diversity objectives were used as defined in Section 2.2 while noting that internally, the RLprop algorithm normalizes the objectives using empirical cumulative distribution function (CDF) to make them comparable against each other.

Note that each participant received their own copy of the recommending algorithms (and objectives' weights) that were updated after each step. That is, for each participant, the algorithms were gradually fine-tuned w.r.t. recommended items the user selected in previous iterations.[4] However, these "sandbox" updates did not affect the recommendations given to the other users. Also note that if the item was recommended in one iteration, it was removed from the set of candidates in the subsequent ones so that the user was not overwhelmed with repeating recommendations.

Regarding the number of iterations, we opted for a rather lower number to maintain a reasonable study duration. Otherwise, too much of users' attention could be lost, which would compromise the results. Following the findings of [16], recommendation lists were organized into columns, and the placement (i.e., left or right) of RS variants was randomized. Each item was represented with its poster image, title, genres, a short plot summary, and a link to its trailer to allow users to thoroughly inspect the previously unknown ones.[5] See Figure 2 for a screenshot of the layout.

At each iteration, users were asked for both low-level and high-level feedback. Similarly to the preference elicitation phase, the low-level feedback was a simple click on relevant items. However, we used a different prompt: *"Select items that you would consider watching tonight."* As for the high-level feedback, users were required to assign 1-5 stars to both recommendation lists so as to compare their overall quality. Once the users finished the feedback provision, they were directed to the *customization UI*, where they could modify their propensities towards individual objectives. These were then supplied to the MORS algorithm to generate the next list of recommendations. We evaluated in total four variants of the customization UIs (details in Section 3.3), while one variant was assigned to the user for the whole duration of the study (i.e., a between-user variable). We opted for this due to supposedly substantial carry-over effects that could otherwise compromise our results.

3.2.4. Post-study questionnaire

During the final phase, users were asked to fill out a post-study questionnaire containing 19 questions together with 6 attention checks (instruction manipulation, nonsensical questions, and memory-based questions). The questionnaire was inspired by the ResQue framework [17], but we altered it to primarily cover the effect of customization UIs (see Figure 3 for the exact wording). Users were allowed to reply in the form of a 5-point Likert scale. The exact prompts were "Strongly Disagree", "Disagree", "Neutral", "Agree", "Strongly Agree", and we also allowed users to answer "I don't understand". In the subsequent analysis, we assigned the numeric values of -2, -1, 0, 1, and 2 to these prompts, while discarding the "I don't understand" answers.

In the evaluation, we grouped the question into individual evaluated aspects of users' attitudes towards the system: through *perceived relevance*, *novelty*, and *diversity*, we aim to observe to what extent these correspond to the users' feedback and measurable characteristics of the resulting recommendations. Several questions aim to determine whether users receive *sufficient information* to participate in the study

[3] Based on the implementation from https://www.tensorflow.org/recommenders/examples/basic_retrieval.

[4] Unlike [3], algorithms were only updated by selections originating from that particular algorithm.

[5] Only the movie's poster was initially visible; other information was displayed on mouse hover.

Q1. **[Perceived Relevance]** The movies recommended to me matched my interests.

Q2. **[Perceived Novelty]** The recommended movies were novel to me.

Q3. **[Perceived Diversity]** The recommended movies were diverse.

Q4. **[Information Sufficiency]** The information provided for the recommended movies was sufficient to judge whether I'm gonna like them.

Q5. **[RS Ease-of-use]** I found it easy to tell the system what movies I like or dislike.

Q6. **[RS Ease-of-use]** I found it easy to tell the system whether it did a good job.

Q7. **[Information Sufficiency]** The description of relevance, diversity, and, novelty was clear and sufficient.

Q8. **[Information Sufficiency]** I understood the purpose of tweaking relevance, diversity, and novelty.

Q9. **[UI Initial State]** The initial values of sliders already provided good recommendations.

Q10. **[UI Effect]** Being able to change relevance, diversity, and novelty ratios was useful for me.

Q11. **[UI Effect]** Overall, the effect of tweaking relevance, diversity, and novelty fulfilled my expectations.

Q12. **[UI Effect]** Overall, after modifying the relevance, diversity, and novelty ratios, recommendations improved.

Q13. **[UI Effect]** Modifying relevance, diversity, and novelty values had a visible effect on the upcoming recommendations.

Q14. **[UI Sufficiency]** The mechanism (slider) provided me with sufficient control over the recommendations.

Q15. **[UI Sufficiency]** The mechanism (slider) was sufficient to tell the system what recommendations I want.

Q16. **[UI Sufficiency]** I was not able to describe my preferences w.r.t. relevance, diversity, and novelty.

Q17. **[UI Understandability]** The mechanism (slider) for tweaking the objectives was understandable and intuitive.

Q18. **[UI Ease-of-use]** Appropriate values of relevance, diversity, and novelty ratios were easy to set.

Q19. **[Perceived Satisfaction]** Overall, I am satisfied with the recommender.

Figure 3: The exact wording of questionnaire questions. The evaluated aspect corresponding to each question is displayed in the brackets. Note that "slider" was replaced with other UI names where appropriate.

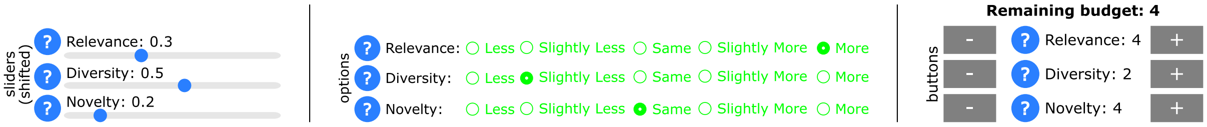

sliders (shifted)
? Relevance: 0.3
? Diversity: 0.5
? Novelty: 0.2
options
? Relevance: ○ Less ○ Slightly Less ○ Same ○ Slightly More ● More
? Diversity: ○ Less ● Slightly Less ○ Same ○ Slightly More ○ More
? Novelty: ○ Less ○ Slightly Less ● Same ○ Slightly More ○ More
buttons
Remaining budget: 4
- ? Relevance: 4 +
- ? Diversity: 2 +
- ? Novelty: 4 +

Figure 4: Different variants of the customization UIs: *sliders*/*sliders_shifted*, *options*, and *buttons*. The following prompt was displayed for all layouts: "How much of the specified quality should be present in the next recommendation iteration?"

and whether the study interface was *easy to use*. Then, a series of questions focused on the customization UIs: whether the *initial state* (i.e., estimated propensities) already provided good recommendations, whether the *effect* of changing propensities was both positive and substantial, and whether the UI was *understandable*, *easy to use* and gave the users *sufficient control* to express their preferences. Finally, we also enquired about the overall *perceived satisfaction* of the users.

3.3. Customization UIs

The user study evaluated four different UI variants: *sliders*, *sliders_shifted*, *options*, and *buttons* (see Figure 4). The *sliders* layout comprised three sliders, one for each objective, that automatically normalize values to unit sum, i.e., when one value was being increased, others decreased proportionally. Note that the sliders were initialized with the previous values of each objective. The *sliders_shifted* layout appeared the same from the user's point of view, but in line with the findings of [3], the relative weight of relevance was increased. In particular, upon receiving the user-defined weights, we silently increased the relevance's weight by the factor of $f = 0.5$ and then re-normalized the weights again. As such, both *sliders* variants provide users an interface with a well-perceivable tradeoff between individual objectives and a chance to provide fine-grained preferences.

The *options* layout provided five radio buttons for each objective that allowed users to manipulate the objectives relative to their previous weights. At k-th iteration, objective weights $w^{[k]}$ were calculated as $w^{[k-1]} * f$, where the factor f was derived from selected options (*less*: 1/2, *slightly less*: 2/3, *same*: 1/1, *slightly more*: 3/2, and *more*: 2/1). As such, the *options* UI gives users a chance to relate their feedback to the previous recommendations while allowing them to supply coarse-grained responses only.

Table 1
Overall comparison of single-objective and multi-objective RS. Average results per user and recommending algorithm are displayed. Note that "IMP" stands for metrics evaluated on impressed (displayed) items, and "SEL" stands for metrics evaluated on items selected by the user. Stat. sign. results (Paired t-test p-value ≤ 0.05) are marked with an asterisk (*).

Algorithm		Estimated relevance	CF-ILD	CB-ILD	Novelty	Recency	Topic_Coverage
Single-objective	IMP	***1.537**	0.881	0.324	0.972	2014.0	0.739
Multi-objective		1.136	***0.957**	***0.375**	***0.990**	***2016.8**	***0.791**
Single-objective	SEL	***1.551**	0.857	0.299	0.966	2013.7	***0.569**
Multi-objective		1.282	***0.931**	***0.338**	***0.984**	***2016.2**	0.535

Finally, the plus-minus *buttons* layout utilized "virtual coins" to allow users to increase/decrease the objective's importance. At the very beginning, ten coins were assigned w.r.t. preference elicitation, and at each iteration, the user received four additional coins to assign. Users could also transfer the coins already assigned to other objectives (via a minus button). This is a very similar setting to many RPG games, where players define their avatar's statistics when the game beginnings and then incrementally update them after some level-ups are accumulated. Therefore, we believe users may be quite familiar with such a UI as well. Similarly as sliders, *buttons* UI is tuned to visualize the tradeoff between individual objectives. However, it only allows for a coarse-grained response and nudges users towards smaller, incremental changes.

4. Results

The study was conducted in June 2023. In total, 142 participants were recruited using the Prolific.com service. Participants were pre-screened for fluent English, no less than 10 previous submissions, and a 99% approval rate. Twelve users did not finish the study, and, in addition, we rejected 9 participants due to failed attention checks, which resulted in 121 participants uniformly distributed along individual UIs (i.e., at least 30 participants evaluated each UI). The study sample size was constrained by the funds allocated for participants' compensations. Nonetheless, we also conducted a sensitivity analysis in G*Power software [18] using ANOVA with four groups, $\alpha = 0.05$, and $1 - \beta = 0.8$, concluding that the study should be capable of discovering medium effects (Cohen's $f = 0.305$) with reasonable probability.

As for the participant's demographics, the sample was rather well-balanced regarding gender (50% female, 49% male, 1% unspecified). Participants were rather younger in general (mean age = 27, standard deviation = 7.7, median age = 24), mostly white (64%) or black (21%), and mostly from South Africa (21%) or several European countries (55% in total). The average time to complete the study was 15 minutes.

In the analysis of the results, we focused on three main aspects: (i) whether different UIs affected the received implicit and explicit user feedback, (ii) whether the UIs affected perceived RS qualities as expressed in the questionnaire, and (iii) how individual questionnaire answers correlate with each other.

4.1. Users Feedback Analysis

4.1.1. Comparison of single-objective and multi-objective RS

Let us first focus on the overall difference between the results of single and multi-objective RS. We first analyzed, whether the single- and multi-objective RS actually supplied users with different lists of recommendations. To do so, we compared the corresponding pairs of lists given to the user at each iteration w.r.t. the size of their intersection. Depending on the customization UI, the mean intersection ranged from 12% (buttons UI) to 28% (sliders_shifted). Therefore we can conclude that the lists were sufficiently different to perform the subsequent analyses.

Table 2
Overall results of the multi-objective RS w.r.t. customization UIs. The highest results are in bold, while the lowest results are in italics. Results significantly lower (p-value < 0.05 w.r.t. Fisher's exact test for hit rate and one-sided T-test for ratings and weights) than the highest ones are marked with an asterisk (*). Results significantly higher than the lowest ones are denoted with a circle (○).

	Feedback			Mean propensity scores		
UI variant	Selects fraction	Hit ratio	Mean rating	Relevance	Diversity	Novelty
Sliders	0.744	○ **0.316**	2.737	*○ 0.508	*○ 0.257	*○ 0.235
Sliders_shifted	**0.774**	○ 0.312	○ **2.946**	○ **0.618**	* *0.195*	* *0.187*
Options	0.728	○ 0.307	○ 2.853	*○ 0.527	*○ 0.254	*○ 0.219
Buttons	*0.597*	* *0.247*	* *2.571*	* *0.415*	○ **0.324**	○ **0.262**

Next, Table 1 contains estimated relevance and beyond-accuracy metrics evaluated on the resulting lists. Similarly as in [3], we observed that MORS provided recommendations of higher diversity (CF-ILD) and novelty. MORS also provided more diverse recommendations w.r.t. *content-based ILD* (cosine similarity of associated genres; denoted as *CB-ILD*), more *recent* movies (mean year of release), and had higher *coverage of topics* (w.r.t. associated genres). We evaluated these metrics both w.r.t. individual lists and w.r.t. all items the algorithm recommended to a particular user throughout the six recommendation iterations - yet the conclusions were the same. Also, all considered customization UIs exhibited the same trend. However, the improvements in beyond-accuracy metrics were achieved at the expense of a significant drop in the estimated relevance of recommended items (1.537 vs. 1.136). It seemed that the deficiency w.r.t. relevance was perceived also by the users, who selected items with significantly higher estimated relevance than the average values (1.282 vs. 1.136, T-test p-value: 1.8e-50). While a similar trend was also observed for SORS selections, its magnitude was much smaller.

Overall, single-objective RS obtained more user selections (3096 vs. 2200) and also received a higher average rating from participants (3.36 vs. 2.78). On the other hand, significantly higher diversity, novelty, and recency were also maintained within the selected items recommended by MORS as compared to those recommended by SORS. Furthermore, the ratio of selected items recommended by SORS tends to drop with subsequent iterations, while the volume of selected items recommended by MORS remained roughly the same throughout all iterations. As such, we can conclude that despite not beating the *single-objective* RS w.r.t. short-term utility, MORS brings favorable features that might pay off in the long run.

4.1.2. Comparison of different customization UIs

Table 2 depicts the results of the MORS separately for individual customization UIs. Here, the selections fraction depicts the ratio between the volume of selections for SORS and MORS, while the hit ratio depicts the fraction between the number of selections and the number of impressions for MORS. As such, these represent the relative and absolute relevance w.r.t. implicit feedback data. Notably, the *buttons* UI attracted the least selections both absolutely (hit rate) and relatively (selects fraction) and also produced the lowest explicit ratings on average. The results of the three remaining variants were mostly comparable with each other. The *options* variant attracted slightly fewer selections than both the *sliders* variants, but the difference was not significant. Similar results were also obtained w.r.t. overall algorithm's ratings.

In order to better understand the inferior performance of the *buttons* UI, we investigated the weights assigned to each objective (depicted in Table 2 as well). Notable differences were observed for the relevance objective, where the *buttons* UI had in average lowest values, *options* and *sliders* UI represent an approximate midpoint, and *sliders_shifted* ended with the highest values in average. Inverse ordering was observed for both novelty and diversity. The reasoning behind *sliders_shifted* is straightforward as we intentionally manipulated its interpretation towards adding more relevance. However, it is not so clear why such low relevance weights were used in *buttons* UI. We did not find any substantial

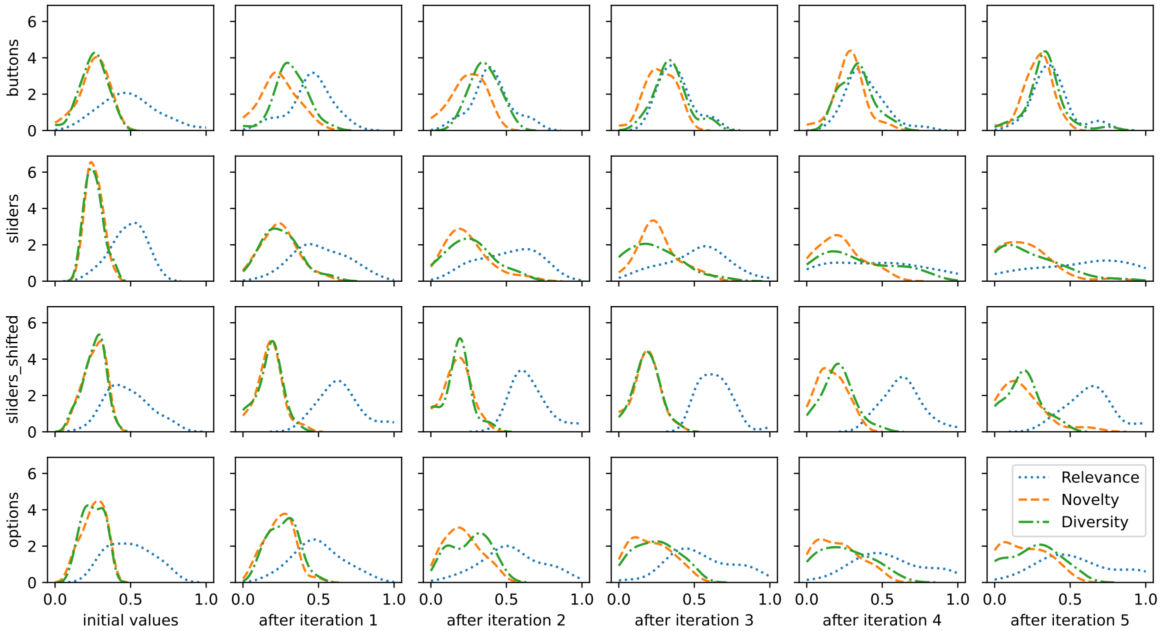
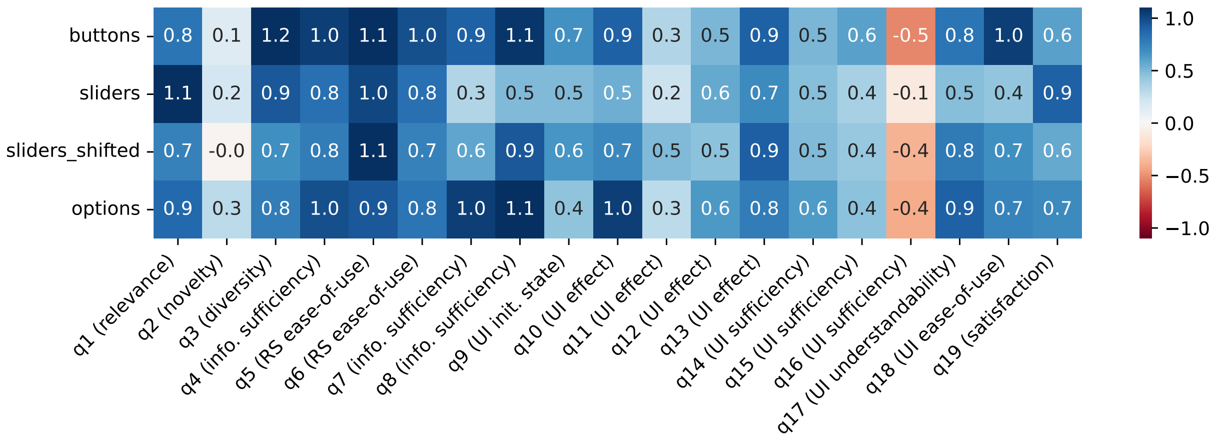


Figure 5: Distribution of the propensity scores provided by the users in each iteration.



	q1 (relevance)	q2 (novelty)	q3 (diversity)	q4 (info. sufficiency)	q5 (RS ease-of-use)	q6 (RS ease-of-use)	q7 (info. sufficiency)	q8 (info. sufficiency)	q9 (UI init. state)	q10 (UI effect)	q11 (UI effect)	q12 (UI effect)	q13 (UI effect)	q14 (UI sufficiency)	q15 (UI sufficiency)	q16 (UI sufficiency)	q17 (UI understandability)	q18 (UI ease-of-use)	q19 (satisfaction)
buttons	0.8	0.1	1.2	1.0	1.1	1.0	0.9	1.1	0.7	0.9	0.3	0.5	0.9	0.5	0.6	-0.5	0.8	1.0	0.6
sliders	1.1	0.2	0.9	0.8	1.0	0.8	0.3	0.5	0.5	0.5	0.2	0.6	0.7	0.5	0.4	-0.1	0.5	0.4	0.9
sliders_shifted	0.7	-0.0	0.7	0.8	1.1	0.7	0.6	0.9	0.6	0.7	0.5	0.5	0.9	0.5	0.4	-0.4	0.8	0.7	0.6
options	0.9	0.3	0.8	1.0	0.9	0.8	1.0	1.1	0.4	1.0	0.3	0.6	0.8	0.6	0.4	-0.4	0.9	0.7	0.7

Figure 6: Results of the questionnaire analysis. The mean of the numeric values corresponding to individual answers is displayed.

differences in the initial weights (i.e., after preference elicitation), so we trust this was a deliberate act of the users. Furthermore, Figure 5 depicts the distribution of propensity scores in each iteration and for each UI. It can be seen that the change is rather gradual for all UIs, but the average vector of the change in *buttons* UI is opposite to those of other UIs (i.e., demoting rather than promoting relevance). Also, while the other UIs tend to disperse the propensities more, these remain fairly compact in the case of *buttons* UI.

As an additional observation, we can see that while the feedback manipulation introduced by the *sliders_shifted* UI had a visible effect on the propensity scores, it did not fully translate to the users' feedback. While the fraction of selections and the mean rating were slightly higher for *sliders_shifted* than for *sliders*, the difference was not significant, and also *sliders* achieved a slightly higher hit ratio. We hypothesize that the measured difference in the resulting recommendations was simply not substantial enough to trigger a significantly different response from the users. This is in line with the observations of [19] regarding perceived diversity, where users often perceived the diversity of the presented lists inversely or indifferently, despite relatively large differences in the measured diversity levels.

4.2. Questionnaire Analysis

The feedback analysis revealed that the *buttons* UI leads to inferior results w.r.t. short-term relevance, while the other three UIs perform comparably with a slight preference towards *sliders* and *sliders_shifted*. However, it is not yet clear whether the users perceived the results alike. So, in the questionnaire analysis, we focused on evaluating additional axes of RS's and customization UI's quality.

Figure 6 depicts the results of the questionnaire analysis. Let us start with overall remarks. Generally, users were able to understand and answer required questions; We received less than 1% of "I don't understand" answers in total. The only question with more (9) of such answers was Q16 targeting *UI satisfaction*. This might be partially caused by the fact that it was the only question with negative phrasing. Therefore, we approach Q16 cautiously here and plan to rephrase it in future studies. Overall, users answered that recommendations were sufficiently relevant (Q1) and diverse (Q3)[6], but not quite as novel (Q2). This might be an effect of using a bit older dataset or not taking movie recency directly into account. The experiment environment's validity is supported by overly positive answers to *RS ease-of-use* and *information sufficiency* (Q4-Q8), but the *sufficiency* and *effect* of the customization *UIs* may be questioned to some extent, given slightly less positive answers for Q11, Q12, and Q14-Q16. Nevertheless, answers on all questions except Q16 were above the neutral point (p-vals < 0.0002). We plan to explore this issue in the future by providing users with more options to tune the recommendations. Finally, users of all UI variants agreed that tweaking the values had a visible effect on resulting recommendations (Q13) and that they were generally satisfied with recommendations (Q19).

Moving to compare different UIs, one of the main results was the superiority of the *options* and *buttons* UI w.r.t. information sufficiency. In particular, participants perceived the description of relevance, novelty, and diversity as clearer (Q7) and better understood the purpose of tweaking their values (Q8). This seemingly affected the perceived usefulness of the UI usage (Q10) and the understandability of the UIs' mechanisms (Q17).[7] These findings can be, to some extent, backed by the work of Funke [6] - if we accept that users internally perceive the task as one with a limited option space.

Let us now briefly mention some more speculative observations. Despite its inferior effectivity, the *buttons* UI surpassed both *sliders* and *sliders_shifted* in perceived ease of setting proper weights for objectives (Q18). This might be an artifact of the finer-grained slider's scale [20, 21], but the same was not sufficiently corroborated for the *options* UI. The *perceived diversity* (Q3) of *buttons*-based recommendations was significantly higher than for *sliders_shifted* – in accordance with the differences of the user-defined diversity weights. In contrast, although the average weights of relevance criterion were much higher for *sliders_shifted* than for *sliders*, users perceived *sliders*-based recommendations as significantly more matching to their interests (Q1). This supports our previous hypothesis on the somewhat inconsistent perception of individual objectives. However, a dedicated future study is needed to quantify the magnitude of such inconsistencies.

4.3. Questionnaire correlations

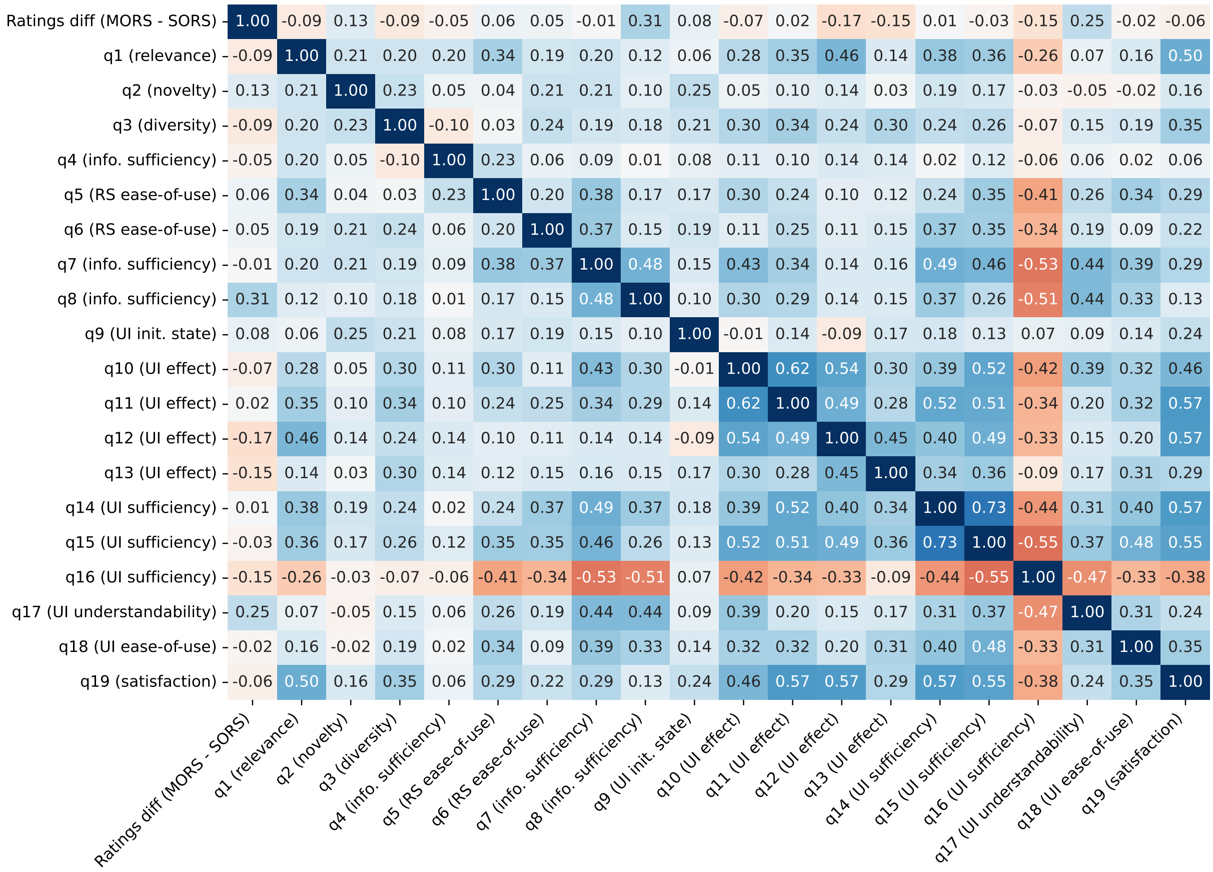
Finally, let us focus on the interdependence of the questionnaire answers. Figure 7 depicts the correlation matrix of the responses to individual questions in the post-study questionnaire. We derive several interesting observations from the results.

First, considering overall satisfaction (Q19) as a target variable, we can see that no other evaluated quality criteria exhibited a substantial negative correlation with satisfaction.[8] Also, while the perceived relevance (Q1) was strongly correlated with the overall satisfaction ($\rho = 0.5$), several questions related to UI's effect and sufficiency had an even larger impact (Q11, Q12, Q14, Q16). This also translates to

[6] Means significantly above the neutral point; one-sample t-test p-vals < 2.6e-19.

[7] In Q7, *options* improved over *sliders* (one-sided T-test p-value: 0.002) and *sliders_shifted* (p-val: 0.03). Also, *buttons* improved over *sliders* (p-val: 0.02). In Q8, *options* and *buttons* improved over *sliders* (p-vals: 0.002 and 0.02 resp.). In Q10, *options* improved over *sliders* (p-val: 0.027). In Q17, *options* improved over *sliders* (p-val: 0.041). Also, if all *information sufficiency* answers are merged together, *options* UI significantly outperforms *sliders* and *sliders_shifted*, while *buttons* UI outperforms *sliders*.

[8] Note that Q16 was negatively formulated itself, so negative values actually indicate a positive effect.



	Ratings diff (MORS - SORS)	q1 (relevance)	q2 (novelty)	q3 (diversity)	q4 (info. sufficiency)	q5 (RS ease-of-use)	q6 (RS ease-of-use)	q7 (info. sufficiency)	q8 (info. sufficiency)	q9 (UI init. state)	q10 (UI effect)	q11 (UI effect)	q12 (UI effect)	q13 (UI effect)	q14 (UI sufficiency)	q15 (UI sufficiency)	q16 (UI sufficiency)	q17 (UI understandability)	q18 (UI ease-of-use)	q19 (satisfaction)
Ratings diff (MORS - SORS)	1.00	-0.09	0.13	-0.09	-0.05	0.06	0.05	-0.01	0.31	0.08	-0.07	0.02	-0.17	-0.15	0.01	-0.03	-0.15	0.25	-0.02	-0.06
q1 (relevance)	-0.09	1.00	0.21	0.20	0.20	0.34	0.19	0.20	0.12	0.06	0.28	0.35	0.46	0.14	0.38	0.36	-0.26	0.07	0.16	0.50
q2 (novelty)	0.13	0.21	1.00	0.23	0.05	0.04	0.21	0.21	0.10	0.25	0.05	0.10	0.14	0.03	0.19	0.17	-0.03	-0.05	-0.02	0.16
q3 (diversity)	-0.09	0.20	0.23	1.00	-0.10	0.03	0.24	0.19	0.18	0.21	0.30	0.34	0.24	0.30	0.24	0.26	-0.07	0.15	0.19	0.35
q4 (info. sufficiency)	-0.05	0.20	0.05	-0.10	1.00	0.23	0.06	0.09	0.01	0.08	0.11	0.10	0.14	0.14	0.02	0.12	-0.06	0.06	0.02	0.06
q5 (RS ease-of-use)	0.06	0.34	0.04	0.03	0.23	1.00	0.20	0.38	0.17	0.17	0.30	0.24	0.10	0.12	0.24	0.35	-0.41	0.26	0.34	0.29
q6 (RS ease-of-use)	0.05	0.19	0.21	0.24	0.06	0.20	1.00	0.37	0.15	0.19	0.11	0.25	0.11	0.15	0.37	0.35	-0.34	0.19	0.09	0.22
q7 (info. sufficiency)	-0.01	0.20	0.21	0.19	0.09	0.38	0.37	1.00	0.48	0.15	0.43	0.34	0.14	0.16	0.49	0.46	-0.53	0.44	0.39	0.29
q8 (info. sufficiency)	0.31	0.12	0.10	0.18	0.01	0.17	0.15	0.48	1.00	0.10	0.30	0.29	0.14	0.15	0.37	0.26	-0.51	0.44	0.33	0.13
q9 (UI init. state)	0.08	0.06	0.25	0.21	0.08	0.17	0.19	0.15	0.10	1.00	-0.01	0.14	-0.09	0.17	0.18	0.13	0.07	0.09	0.14	0.24
q10 (UI effect)	-0.07	0.28	0.05	0.30	0.11	0.30	0.11	0.43	0.30	-0.01	1.00	0.62	0.54	0.30	0.39	0.52	-0.42	0.39	0.32	0.46
q11 (UI effect)	0.02	0.35	0.10	0.34	0.10	0.24	0.25	0.34	0.29	0.14	0.62	1.00	0.49	0.28	0.52	0.51	-0.34	0.20	0.32	0.57
q12 (UI effect)	-0.17	0.46	0.14	0.24	0.14	0.10	0.11	0.14	0.14	-0.09	0.54	0.49	1.00	0.45	0.40	0.49	-0.33	0.15	0.20	0.57
q13 (UI effect)	-0.15	0.14	0.03	0.30	0.14	0.12	0.15	0.16	0.15	0.17	0.30	0.28	0.45	1.00	0.34	0.36	-0.09	0.17	0.31	0.29
q14 (UI sufficiency)	0.01	0.38	0.19	0.24	0.02	0.24	0.37	0.49	0.37	0.18	0.39	0.52	0.40	0.34	1.00	0.73	-0.44	0.31	0.40	0.57
q15 (UI sufficiency)	-0.03	0.36	0.17	0.26	0.12	0.35	0.35	0.46	0.26	0.13	0.52	0.51	0.49	0.36	0.73	1.00	-0.55	0.37	0.48	0.55
q16 (UI sufficiency)	-0.15	-0.26	-0.03	-0.07	-0.06	-0.41	-0.34	-0.53	-0.51	0.07	-0.42	-0.34	-0.33	-0.09	-0.44	-0.55	1.00	-0.47	-0.33	-0.38
q17 (UI understandability)	0.25	0.07	-0.05	0.15	0.06	0.26	0.19	0.44	0.44	0.09	0.39	0.20	0.15	0.17	0.31	0.37	-0.47	1.00	0.31	0.24
q18 (UI ease-of-use)	-0.02	0.16	-0.02	0.19	0.02	0.34	0.09	0.39	0.33	0.14	0.32	0.32	0.20	0.31	0.40	0.48	-0.33	0.31	1.00	0.35
q19 (satisfaction)	-0.06	0.50	0.16	0.35	0.06	0.29	0.22	0.29	0.13	0.24	0.46	0.57	0.57	0.29	0.57	0.55	-0.38	0.24	0.35	1.00

Figure 7: Pearson's correlation between answers to individual questions. In addition, correlations w.r.t. Ratings diff, i.e., the difference between per-user mean ratings to MORS and SORS recommendations, are displayed in the first row/column.

compound statistics,[9] where the mean *UI effect* and mean *UI sufficiency* answers are strongly correlated with the overall satisfaction ($\rho = 0.57$ and $\rho = 0.62$, respectively). Furthermore, *UI's understandability* (Q17) and *ease of use* (Q18) also exhibited a non-negligible correlation with overall satisfaction. To sum up, these findings indicate a possible strong influence of UI controls and their function on the overall user experience with the recommender systems. In this study, we only evaluated limited graphical user interfaces. However, in light of emerging conversational RS powered by large-language models, it may be crucial to focus on this aspect of user experience also in connection with additional UI and interactional designs.

Second, some of the considered quality axes (*information sufficiency*, *RS ease-of-use*, *UI effect*, and *UI sufficiency*) were targetted by multiple questions. However, while the questions targetting *UI's effect* and *UI's sufficiency* were highly correlated in most cases, this was not true for the *information sufficiency* and *RS ease-of-use*. As such, a finer-grained division of these objectives may be considered in future work.

In addition, we also focused on whether the difference in the feedback users provided on MORS and SORS recommendations can be explained by some of the questionnaire answers. To do so, we also calculated the correlations for the per-user differences in mean ratings to MORS and SORS. In most cases, we obtained close-to-zero results, with the exception of two moderate correlations: Q8 and Q17, both targeting possible understandability issues (Q8: *"I understood the purpose of tweaking relevance, diversity, and novelty."*; Q17: *"The mechanism (slider) for tweaking the objectives was understandable and intuitive."*). Therefore, we can preliminarily conclude that the main driving force behind the adoption of customizable individual MORS is actually whether we did a good job of explaining why & how should

[9] I.e., using the mean of all answers targeting the same evaluated aspect.

users tune their propensities.

5. Conclusions and Limitations

We tackled the problem of allowing users to indicate their propensity towards different recommendation objectives, so as to shape more effective and better-tailored MORS. To this end, we conducted a user study that allowed users to customize MORS via *sliders*, *sliders_shifted*, *buttons*, and *options* UIs. Results show that while multiple UIs can lead to similarly effective recommendations (w.r.t. user feedback), they can significantly vary in some of the user-perceived quality criteria. In particular, *buttons* UI resulted in the lowest consumption-related statistics as well as lowest user ratings, while the other three UIs performed without significant differences from each other. The main driving force behind this inferiority was a different distribution of propensities the users set through this UI. Further research is needed to focus on the causes of this behavior and possible ways to support users in setting the best possible values for their current needs.

A subsequent questionnaire analysis revealed certain advantages of the *options* UI variant as compared to more standard *slider*-based UIs. In particular, *options* UI dominated over *sliders* and *sliders_shifted* in *information sufficiency*, *perceived usefulness*, and *UI's ease-of-use* aspects. As such, we can tentatively recommend *options* UI with prompts relative to the previous criteria values as a good variant for customizing *local* MORS.

As an initial work on a rather complex topic, the study has numerous limitations, which we plan to address in the future. First, when designing the evaluated UIs, we primarily aimed at the most commonly used UI components. Even though, there was a plethora of parameters and design options that we could not test due to the limits imposed on the number of participants. In particular, the current study could not disentangle whether the differences between *sliders* and *options* UIs were mainly caused by the different "grounding" of the choices (i.e., relative to other criteria vs. relative to previous choices), different response granularity, or different appearance. So, although we can conclude that there are viable alternatives to the most common choice (i.e., *sliders* UI), the selection of the best such alternative is a matter for future work.

Similarly as in some related works [3], the study revealed several features that should favor MORS over single-objective RS in the long term. However, a truly long-term study should be conducted to verify these assumptions. As indicated by not-so-positive scores for Q14-Q16, there is some space for revisiting the objectives by incorporating additional criteria or re-defining the current ones. Finally, while the pool of participants was sufficient to reveal the differences in user feedback and corroborate medium-sized effects in the questionnaire analysis, subtle effects might have been overlooked, which could be remedied by contracting more users.

Overall, we plan a series of larger follow-up studies that will focus on a more detailed long-term analysis of user interaction and perception of MORS. This should also include studying the impact of different domains, dataset properties, recommending algorithms, and study designs.

Acknowledgments

This paper has been supported by Czech Science Foundation (GAČR) project 22-21696S, Charles University grant SVV-260698/2023, and Charles University Grant Agency (GA UK) project number 188322.

References

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

[2] D. Jannach, Multi-objective recommendation: Overview and challenges, in: H. Abdollahpouri, S. Sahebi, M. Elahi, M. Mansoury, B. Loni, Z. Nazari, M. Dimakopoulou (Eds.), Proceedings of the 2nd Workshop on Multi-Objective Recommender Systems co-located with 16th ACM Conference on Recommender Systems (RecSys 2022), Seattle, WA, USA, 18th-23rd September 2022, volume 3268 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3268/paper1.pdf.

[3] P. Dokoupil, L. Peska, L. Boratto, Looks can be deceiving: Linking user-item interactions and user's propensity towards multi-objective recommendations, in: Proceedings of the Seventeenth ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3604915.3608848. doi:10.1145/3604915.3608848.

[4] F. M. Harper, F. Xu, H. Kaur, K. Condiff, S. Chang, L. Terveen, Putting users in control of their recommendations, in: Proceedings of the 9th ACM Conference on Recommender Systems, RecSys '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 3–10. URL: https://doi.org/10.1145/2792838.2800179. doi:10.1145/2792838.2800179.

[5] Y. Liang, M. C. Willemsen, Personalized recommendations for music genre exploration, in: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 276–284. URL: https://doi.org/10.1145/3320435.3320455. doi:10.1145/3320435.3320455.

[6] F. Funke, A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales, Social Science Computer Review 34 (2016) 244–254. doi:10.1177/0894439315575477.

[7] B. Shneiderman, Designing the User Interface: Strategies for Effective Human-Computer Interaction, 3rd ed., Addison-Wesley Longman Publishing Co., Inc., USA, 1997.

[8] Q. Zhao, The superior psychological impact of absolute (vs. relative) standing feedback does not depend on the reward criterion, Social Psychology of Education 26 (2023) 473–484. URL: https://doi.org/10.1007/s11218-023-09758-2. doi:10.1007/s11218-023-09758-2.

[9] L. Peska, S. Balcar, The effect of feedback granularity on recommender systems performance, in: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 586–591. URL: https://doi.org/10.1145/3523227.3551479. doi:10.1145/3523227.3551479.

[10] D. Jannach, H. Abdollahpouri, A survey on multi-objective recommender systems, Frontiers in Big Data 6 (2023). URL: https://www.frontiersin.org/articles/10.3389/fdata.2023.1157899. doi:10.3389/fdata.2023.1157899.

[11] P. Dokoupil, L. Peska, Easystudy: Framework for easy deployment of user studies on recommender systems, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1196–1199. URL: https://doi.org/10.1145/3604915.3610640. doi:10.1145/3604915.3610640.

[12] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2015). URL: https://doi.org/10.1145/2827872. doi:10.1145/2827872.

[13] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 109–116. URL: https://doi.org/10.1145/2043932.2043955. doi:10.1145/2043932.2043955.

[14] K. Bradley, B. Smyth, Improving recommendation diversity, in: Proceedings of the twelfth Irish conference on artificial intelligence and cognitive science, Maynooth, Ireland, volume 85, Citeseer, 2001, pp. 141–152.

[15] L. Peska, P. Dokoupil, Towards results-level proportionality for multi-objective recommender systems, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1963–1968. URL: https://doi.org/10.1145/3477495.3531787. doi:10.1145/3477495.3531787.

[16] P. Dokoupil, L. Peska, L. Boratto, Rows or columns? minimizing presentation bias when comparing

multiple recommender systems, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 2354–2358. URL: https://doi.org/10.1145/3539618.3592056. doi:10.1145/3539618.3592056.

[17] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 157–164. URL: https://doi.org/10.1145/2043932.2043962. doi:10.1145/2043932.2043962.

[18] F. Faul, E. Erdfelder, A. Buchner, A.-G. Lang, Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses, Behavior Research Methods 41 (2009) 1149–1160. URL: https://doi.org/10.3758/BRM.41.4.1149. doi:10.3758/BRM.41.4.1149.

[19] P. Dokoupil, L. Boratto, L. Peska, User perceptions of diversity in recommender systems, in: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 212–222. URL: https://doi.org/10.1145/3627043.3659555. doi:10.1145/3627043.3659555.

[20] C. C. Preston, A. M. Colman, Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences, Acta Psychologica 104 (2000) 1–15. URL: https://www.sciencedirect.com/science/article/pii/S0001691899000505. doi:https://doi.org/10.1016/S0001-6918(99)00050-5.

[21] E. I. Sparling, S. Sen, Rating: How difficult is it?, in: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 149–156. URL: https://doi.org/10.1145/2043932.2043961. doi:10.1145/2043932.2043961.