UNICA UNIVERSITÀ DEGLI STUDI DI CAGLIARI

© ACM 2023. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in

Patrik Dokoupil, Ladislav Peska, and Ludovico Boratto. 2023. Looks Can Be Deceiving: Linking User-Item Interactions and User's Propensity Towards Multi-Objective Recommendations. In Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23). Association for Computing Machinery, New York, NY, USA, 912–918.

The publisher's version is available at:
https://doi.org/10.1145/3604915.3608848

When citing, please refer to the published version.

This full text was downloaded from UNICA IRIS https://iris.unica.it/

arXiv:2307.00654v2 [cs.IR] 16 Oct 2023

Looks Can Be Deceiving: Linking User-Item Interactions and User's Propensity Towards Multi-Objective Recommendations

PATRIK DOKOUPIL, Faculty of Mathematics and Physics, Charles University, Prague, Czechia
LADISLAV PESKA, Faculty of Mathematics and Physics, Charles University, Prague, Czechia
LUDOVICO BORATTO, University of Cagliari, Italy

Multi-objective recommender systems (MORS) provide suggestions to users according to multiple (and possibly conflicting) goals. When a system optimizes its results at the individual-user level, it tailors them on a user's propensity towards the different objectives. Hence, the capability to understand users' fine-grained needs towards each goal is crucial. In this paper, we present the results of a user study in which we monitored the way users interacted with recommended items, as well as their self-proclaimed propensities towards relevance, novelty, and diversity objectives. The study was divided into several sessions, where users evaluated recommendation lists originating from a relevance-only single-objective baseline as well as MORS. We show that, despite MORS-based recommendations attracting fewer selections, their presence in the early sessions are crucial for users' satisfaction in the later stages. Surprisingly, the self-proclaimed willingness of users to interact with novel and diverse items is not always reflected in the recommendations they accept. Post-study questionnaires provide insights on how to deal with this matter, suggesting that MORS-based results should be accompanied by elements that allow users to understand the recommendations, so as to facilitate the choice of whether a recommendation should be accepted or not. Detailed study results are available at https://bit.ly/looks-can-be-deceiving-repo.

CCS Concepts: • **Information systems** → **Recommender systems**.

Additional Key Words and Phrases: Multi-objective recommender systems, User study, Novelty, Diversity

ACM Reference Format:
Patrik Dokoupil, Ladislav Peska, and Ludovico Boratto. 2023. Looks Can Be Deceiving: Linking User-Item Interactions and User's Propensity Towards Multi-Objective Recommendations. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3604915.3608848

1 INTRODUCTION

Motivation and context. *Beyond-accuracy objectives* are gaining more and more attention in Recommender Systems (RSs). Indeed, it is now paramount to pair recommendation effectiveness with properties that account for user perspectives (such as novelty and diversity [3] or consumer fairness [1]), or that are aligned with the recommended items (such as behavioral biases or provider fairness [1]). *Multi-objective recommender systems* (MORS) support this paradigm by generating results that account for multiple properties [23]. Recent literature has studied how to account for multi-objective goals from different angles. The user perspective was tackled by Li et al. [13], which balance recommendation accuracy for users with different levels of activities. From an item perspective, Ge et al. [7] proposed an approach to balance item relevance and exposure. Considering both the user and item perspectives, Naghiaei et al. [15] propose a re-ranking approach to account for consumer and provider fairness. Other studies blend the multiple objectives into a

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

single function, in order to obtain a Pareto-optimal solution [14, 22]. Recent advances have also proposed MORSs in sequential settings, by optimizing the results for accuracy, diversity, and novelty [19].

MORS can account for multiple objectives at the *aggregate* level, by balancing these objectives over the entire user base (e.g., the system is capable of offering a certain level of diversity), or at the *individual* level, by matching the beyond-accuracy needs of each user in a different way (e.g., the recommendations of one user might intentionally be more diverse than those of another one) [10]. MORS that operate at the individual level have optimized the recommendation process mainly via online interactions, such as conversational approaches [6, 12] or via critiquing [4, 21], but approaches aiming at learning individual propensities from past interactions also exist, e.g., [11].

Open issues. Even though MORS that operate at the individual level have as a goal the optimization of the needs of each user, their functioning and evaluation either requires continuous interaction with the users or is based on offline data without any feedback from the users. Having these two extremes as the only options leads to two main questions that so far remain unanswered. At the RS functioning level, we need to understand how to incorporate the propensity of users towards certain beyond-accuracy properties into the recommendation process. This is not possible in offline approaches, while online ones work until a recommendation is accepted (i.e., the conversation or the critiques stop appearing). At the evaluation level, we do not know to what extent the recommendations accepted by the users are driven by these beyond-accuracy goals. Hence, *understanding directly from the users their propensity towards beyond-accuracy goals and how they should be reflected in the recommendations* is a key open problem for the functioning of MORS that operate at the individual level.

Our contributions. To address the aforementioned issues, we present the results of a user study aimed at linking the self-proclaimed propensity of users towards relevance, novelty, and diversity criteria with their actual acceptance of provided recommendations. In particular, we asked users to iterate through several recommending sessions in the Movie domain. We confronted them with results of a relevance-based single-objective RS and two MORS variants balancing relevance, novelty, and diversity criteria. We further allowed them to tune MORS by defining their propensity towards the aforementioned criteria. Therefore, for the first time in the literature, we can link (i) the propensity of the users to interact with items characterized by certain beyond-accuracy properties, with (ii) their propensity to accept recommendations offering these same properties.

Our results provide interesting insights into how users' propensity towards beyond-accuracy goals can be reflected in the individual-level MORS. Indeed, despite the users' self-declared propensity towards multi-objective goals, single-objective recommendations attracted more selections than those generated by MORS. In the evaluation, we argue that the presence of MORS recommendations (and selections) is crucial for long-term user satisfaction. We also discovered that users' selection behavior exhibits interesting deviations from the distributions induced by displayed items (impressions) and propensities towards individual objectives (weights). Indeed, in the case of single-objective RS, users on average selected items with lower estimated relevance scores than the average of recommended items. Likewise, in the case of MORS, users selected less diverse and novel items with higher estimated relevance than what was the average of those metrics w.r.t. the recommended items. This propensity towards relevant items in MORS-based recommendations happened regardless of the fact that the users could manually fine-tune the level of novelty and diversity. We briefly analyze the possible causes of these phenomena and suggest plausible mitigation strategies.

2 USER STUDY DESIGN

The study was conducted online[1] and consisted of the following steps: informed consent and basic demographics, preference elicitation, recommendation sessions (8x), and a post-study questionnaire.

Dataset and pre-processing. The study was conducted on top of the MovieLens-Latest dataset [8], which was selected for its relative novelty and high familiarity with the movie domain. The dataset was utilized in two ways: to populate collaborative filtering algorithms and as a starting point to gather item metadata. In order to comply with the gathered user selections, the feedback was binarized. Furthermore, to only focus on the relevant portion of the dataset, we filtered out movies released before 1990, ratings older than 2010, movies that have less than 50 ratings per year, users with less than 100 ratings, and movies without 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.

Recommender systems. In the study, three RSs variants were evaluated (one single-objective and two multi-objective), denoted as *Beta*, *Gamma*, and *Delta*. *Beta* (single-objective baseline) follows a generalized matrix factorization [9] example from tf.recommenders[2]. We used the embedding size of 32 and 5 training epochs. *Gamma* and *Delta* utilized the predictions of *Beta* as its relevance component, but additionally incorporated also diversity and novelty viewpoints. In particular, *Delta* utilized RLProp algorithm [16] and *Gamma* utilized incremental weighted average [16].[3] Both algorithms were parameterized by the user's propensity towards individual objectives (described in Sec. 2.2).

Beta algorithm was first trained on the MovieLens dataset and then fine-tuned separately (i.e., each study participant received their own private copy of the algorithm). Fine-tuning was done after the preference elicitation step as well as after each recommending session. Note that since the *Beta* algorithm was utilized as a source in both *Gamma* and *Delta*, the feedback received on all recommended items was utilized for *Beta* fine-tuning. Also note that to enhance engagement and coverage, we prohibited repeated recommendations of items that were previously shown to the user.

2.1 Study flow

In the ***initial*** phase, users received a description of the study and were asked for basic demographics (e.g., gender, age, education) as well as to provide informed consent on the study procedure and publication of anonymized results.

In the ***preference elicitation*** phase, participants were asked to select previously known and liked movies out of a randomized list. Depicted movies were sampled on the basis of three objective criteria: overall relevance, novelty, and diversity. For each criterion, we constructed bins of movies with high and low values, and from each bin, we randomly sampled four movies.[4] This procedure aimed to minimize the historical biases present in the source data. Note that users were allowed to load more movies (based on the same procedure) as well as search for a specific movie manually. There were no strict limits on the volume of selected movies, but participants were instructed to try to select at least 5-10 movies (the median volume of actually selected items at this phase was 10).

During each of eight ***recommendation sessions***, the results of two RS were shown to the user. Each time we depicted an output of the single-objective RS (*Beta*) accompanied by one of the MORS (either *Gamma* or *Delta*). Recommendation lists were kept separated, and displayed at randomized positions. The procedure for choosing the MORS variant was as follows. Before the first session, either *Delta* or *Gamma* RS was selected at random. This algorithm is then used in the

[1] https://bit.ly/looks-can-be-deceiving-study
[2] https://www.tensorflow.org/recommenders/examples/basic_retrieval
[3] I.e., recommended items were selected one by one, while their marginal gains were iteratively updated.
[4] The overall relevance was considered w.r.t. average user profile. The novelty was defined as the item's mean popularity complement. The diversity was defined as collaborative intra-list diversity (ILD) w.r.t. the already selected movies from popularity and novelty bins.

first four sessions, while in the last four sessions, we switch to the other MORS variant. As such, algorithm-specific sequence-aware patterns can be observed and the usage MORS variant can be considered as *within-subject* variable.[5]

At each recommending session, we asked study participants to provide both implicit feedback (i.e., select items that they would consider watching tonight) and to provide explicit feedback (i.e., rate the overall performance of depicted RSs on a one-to-five stars scale). After completing the feedback phase, participants were also allowed to modify their propensity (i.e., weights) towards individual objective criteria. This was conducted via a slider depicting the current values for each objective and forcing it to maintain a unit sum of all objectives.

Finally, in the ***post-study questionnaire***, we asked participants to fill in responses (on a 5-point Likert scale) to a series of questions regarding both the general performance of RS as well as questions specifically targeting the user interface (UI) for changing objective weights. Questions were inspired by the ResQue framework [17], but extended to also cover the specifics of the UI for criteria propensity setting (see https://bit.ly/looks-can-be-deceiving-repo for details). The questionnaire also contained several attention checks to remove unreliable participants.

2.2 Considered objectives and their importance weights

Both the *Gamma* and *Delta* RSs aim to incrementally construct the list of recommendations w.r.t. several objective criteria. In particular, they utilize the normalized marginal gains (NMG) individual items provide in terms of these objectives. In this paper, we focused on relevance, novelty, and diversity, defined as follows. For relevance, we considered the sum of estimated relevance scores (predicted by *Beta* algorithm) as an objective, so the marginal gain of each item was its own relevance score: $MG_{i,rel} = \hat{r}_{u,i}$. Normalization is then applied as empirical cumulative distribution function (CDF) w.r.t. all items' marginal gains (see [16] for more details). Similarly, marginal gain w.r.t. novelty was defined as item's mean popularity complement [20]: $MG_{i,nov} = 1 - |u \in U : r_{u,i} \text{ exists}|/|U|$, where $r_{u,i}$ is the feedback of user u on item i and U is the set of all users. Marginal gain w.r.t. diversity is defined as the mean collaborative distance of the item to the list of already selected recommendations:[6] $MG_{i,div} = \frac{1}{|L|} \sum_{\forall j \in L} d(i,j)$, where L is a list of already selected recommendations and $d(i,j)$ is a distance metric – cosine distance on items' ratings in our case.

Both the *Gamma* and *Delta* algorithms used the propensity weights assigned to individual objectives. These were iteratively modified by the users after each session, but their initial values had to be trained based on the data from preference elicitation. We used a similar procedure to [11]. In particular, we calculated the normalized marginal gains (NMG) for each objective and each selected movie. Note that because the user's profile was not established yet, relevance gain was calculated as the mean estimated relevance of the selected items w.r.t. all train set users. Diversity gain was calculated as the mean distance of the selected items from all the displayed ones, and novelty gain remains unchanged. Gains of all selected movies were normalized via CDF defined on the population of all displayed movies. Final estimated propensities were obtained as the mean of all items' NMGs and linearly scaled to unit sum.

3 STUDY RESULTS

The study was conducted in April 2023. In total, 120 participants were recruited using the Prolific.co service. Participants were pre-screened for fluent English, no less than 10 previous submissions, and 99% approval rate. Twelve users did not finish the study and, in addition, we rejected 2 participants due to failed attention checks, which resulted in 106 completed participations.

[5]Merely the ordering of MORS variants is a *between-subject* variable. This was a compromise solution. During study dry-runs, we observed that showing all three recommendation lists at once imposed an excessive cognitive burden on the users. On the other hand, making MORS variant a between-subject variable could introduce an excessive user-specific variance in the results.
[6]I.e., the diversity objective corresponds to the incremental collaborative intra-list diversity, ILD [2].

Table 1. Overall results w.r.t. user feedback and normalized marginal gains of recommended and selected items. Best results are in bold, while significantly inferior results (T-test p-value < 0.05) are denoted with an asterisk (*).

	Feedback		Impressions			Selections		
Algorithm	Selections ratio	Mean rating	NMG_{rel}	NMG_{div}	NMG_{nov}	NMG_{rel}	NMG_{div}	NMG_{nov}
Beta	**0.37**	**3.20**	**0.980**	*0.251	*0.653	**0.974**	*0.246	*0.635
Gamma	*0.19	*2.34	*0.875	**0.787**	**0.849**	*0.910	**0.689**	**0.816**
Delta	*0.26	*2.68	*0.913	*0.659	*0.790	*0.943	*0.515	*0.719

Table 1 depicts the overall study results. It can be seen that *Beta* (relevance-only baseline) significantly outperformed both MORS variants w.r.t. total volume of selections as well as mean algorithm ratings. Out of the two MORS variants, *Delta* obtained significantly more selections (Fisher's exact test p-value: 2.6e-15) as well as significantly higher average ratings than *Gamma* (T-test p-value: 8.6e-6). Note that both implicit and explicit feedback modalities were correlated, but there were some discrepancies (Pearson's correlation: 0.63). We also checked several other statistics with rather expectable results: The volume of selections slightly drops for subsequent sessions (up to 28% drop), top-ranked items were selected more often than lower-ranked (up to 33% drop), etc. Some additional details are available from https://bit.ly/looks-can-be-deceiving-repo. Based on these initial results, we formulated the following questions:

- **RQ1.** Are there some qualities, in which evaluated MORS improve over the single-objective baseline? If so, what is the long-term impact of these qualities on user satisfaction?
- **RQ2.** What are the possible causes of the inferior performance of MORS? Could this be somehow mitigated?

3.1 Beyond-accuracy objectives and their long-term impact

In order to answer RQ1, we focused on the beyond-accuracy criteria of recommended (impressions), but also selected items. A natural choice to start with are the normalized marginal gains of considered objectives. By inspecting Table 1, one can observe that both *Gamma* and *Delta* clearly outperformed *Beta* in terms of NMG_{nov} and NMG_{div} for impressions as well as selections. The increased impression-level novelty and diversity is a direct consequence of recommendations construction, while the selection-level increase indicates that users (to some extent) followed the distribution of recommended items. Obtained results also corresponded to other novelty and diversity metrics: collaborative ILD (0.876 for *Beta* vs. 0.988 for *Gamma* vs. 0.961 for *Delta*), content-based (genre-based) ILD (0.323 vs. 0.385 vs. 0.369), genre coverage (0.484 vs. 0.482 vs. 0.496), mean popularity complement (0.973 vs. 0.996 vs. 0.989), temporal novelty (0.647 vs. 0.932 vs. 0.856). Significant differences were also obtained on selections: collaborative ILD (0.868 vs. 0.964 vs. 0.946), mean popularity complement (0.965 vs. 0.992 vs. 0.990), and temporal novelty (0.625 vs. 0.896 vs. 0.859).

However, although the above-mentioned results are interesting, it is yet to be shown whether they have a practical impact. To do so, we focused on (i) the long-term user satisfaction as a function of users' acceptance of MORS recommendations in early sessions, and (ii) the impact of MORS-based selections on training single-objective RS.

Impact of MORS acceptance on long-term user satisfaction. Let us (optimistically) assume that eight recommendation sessions constitute a sufficient base for a long-term evaluation. Our analysis is based on dividing the sessions into early (i.e., *head*) and late (i.e., *tail*). Then, we measure whether the adoption of MORS recommendations in the *head* had a measurable impact on user satisfaction in the *tail*.

In particular, we considered the size of the *head* to be one to four first sessions and defined three metrics (w.r.t. *head*) to describe a user's MORS adoption: the volume of selections on single-objective RS, the volume of selections on

Table 2. Results of the long-term impact of MORS selections (i.e., *tail* sessions). For the sake of space, we only depict the *head* section sizes of two and three. Significantly better results (one-sided T-test p-value < 0.05) are bold, while an asterisk (*) denotes p-values < 0.01.

	Head size: 2						Head size: 3					
	#MORS		#SORS		MORS ratio		#MORS		#SORS		MORS ratio	
Mean ratings	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
Low-selections	2.56	2.70	2.59	2.66	2.53	2.73	2.54	2.61	2.56	2.59	2.32	***2.83**
High-selections	2.78	**3.07**	2.89	2.95	2.72	***3.12**	2.86	**3.15**	2.99	2.96	2.79	**3.17**
All users	2.61	***2.97**	2.68	2.86	2.61	***2.92**	2.50	***3.02**	2.62	**2.91**	2.57	***2.94**
# selections	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
Low-selections	21.7	**31.6**	19.8	***33.5**	28.3	23.5	18.5	24.4	16.7	***26.1**	22.6	19.9
High-selections	37.1	**47.8**	39.1	45.8	37.0	***47.9**	33.7	**40.5**	33.2	**40.2**	34.5	38.2
All users	28.1	***41.7**	27.3	***40.9**	33.7	34.1	21.2	***35.2**	22.1	***34.8**	27.4	28.7

multi-objective RS, and the ratio of multi-objective selection on all selections (in the results; we denote them as *#SORS*, *#MORS*, *MORS ratio* respectively). For each metric, we divided users into two groups: users with above-median values and with values below-or-equal to the median (denoted as *High* and *Low* clusters). In order to get finer-grained results, we also applied a pre-processing of users to separately consider those who had high or low total volumes of selections in the head segment (denoted as *high-selections*, *low-selections*, and *all users* for no pre-processing). In the *tail* section, we considered the total volume of selections (*#selections*) and the *mean rating* of the provided recommendation lists.

Table 2 shows the results of the long-term impact evaluation. The main outcomes are as follows. For *#MORS*, the *High* cluster exhibited better values of both #selections and mean ratings. However, the inherent flaw is that the same trend appeared in the *head* section as well[7]. This tendency is maintained throughout the study, seemingly without major fluctuations. Therefore, we assume that the high *#MORS* cluster (w.r.t. *head*) merely identifies a cluster of more overall engaged users.

In contrast, *High* and *Low* clusters w.r.t. *MORS ratio* exhibited much more similar performance in the *head* section (at least for shorter *heads* – see further). In the *tail* sections, users of the *High* cluster provided on average significantly higher ratings than the users of the *Low* cluster. To our surprise, the impact on the volume of the selections was much smaller, often insignificant, or even negative. That is, despite selecting similar (or even lower) volumes of items, users of the *High* cluster were in general more satisfied with provided recommendations. Note that the impact is incremental and rather fast. While for the *head* sizes of one and two, there is no substantial difference in the performance of both clusters w.r.t. head, this gradually changes and already for the head size of four, this became noticeable.

Interesting results were also obtained for *#SORS*. While the *high* cluster almost always exhibited a higher volume of all selections in the *tail*, the improvements w.r.t. mean ratings were smaller, mostly insignificant, or even negative. This was despite the fact that quite often, *high* user clusters were associated with higher mean ratings in the head section. We read these results in such a way that, despite being satisfied in the early stages, users who mostly adopted single-objective recommendations struggle to find sufficiently interesting/satisfying recommendations in the later stages. This is despite the fact that there are many "somewhat relevant" items (thus the higher volume of selections). As for the pre-processing variants, the general trend was similar for both sub-groups, but the differences were more

[7]I.e., users who made more MORS selections also made more selections and provided higher ratings in general.

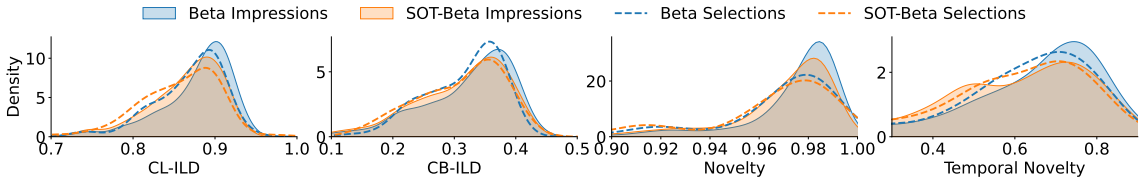

Fig. 1. Comparison of Single-objective trained Beta (*SOT-Beta*) and default Beta, w.r.t. collaborative diversity (*CL-ILD*), content-based diversity (*CB-ILD*), novelty, and temporal novelty on impressions and selections.

pronounced on *high-selection* cluster of users. We assume that this is a natural consequence of the fact that more data is being supplied for fine-tuning.

Overall, we may conclude that early adoption of MORS-based recommendations led to higher satisfaction later on. This is further corroborated by the questionnaire analysis, revealing that the users of *high* cluster w.r.t. MORS ratio provided more positive answers on the questions *"Recommended items were novel to me"* (significant for all *head* sizes), *"Recommended items were diverse"* (significant for *head* sizes of three and four), and *"Recommended items matched my interests"* (significant for *low-selection* users and *head* sizes of three and four).

Impact of MORS selections on the fine-tuning of single-objective RS. In this analysis, we aimed on discovering to what extent was it beneficial to fine-tune single-objective RS with the help of MORS-based selections. To do so, we simulated the behavior of *Beta*, should it be trained only w.r.t. selections made on single-objective recommendations.[8] First, note that while the recommendations of single-objective-trained *Beta* (*SOT-Beta*) gradually departed from the original *Beta*, the intersection remained substantial (decreasing from 80% in the second session to 63% in the last session). This makes the whole procedure feasible, although we can expect that due to the lower volume of impressions, obtained results could somewhat underestimate the true performance of *SOT-Beta*.

Now, let us observe the beyond-accuracy statistics of both *Beta* variants (see Figure 1). *SOT-Beta* exhibited lower collaborative ILD (0.864 vs 0.876; T-test p-value:3.7e-8) w.r.t. impressions. More importantly, this also translated into the inferior ILD w.r.t. selections – that is, when comparing the ILD of all selections of original *Beta* recommendations with those recommended *SOT-Beta* (0.857 vs 0.869, p-value: 0.028). Similar observations can be made also for impression-based content-based ILD, genre coverage, mean popularity complement, and temporal novelty. In these cases, selection-level statistics were also slightly better for *Beta*, but the difference was not significant. To conclude, the existence of MORS-based selections considerably improved the beyond-accuracy properties of the single-objective RS (as long as impressed items are considered), and this partially translated into the improved adoption of items with higher beyond-accuracy statistics by study participants.

3.2 Comparing user's selections with user-defined propensities towards beyond-accuracy objectives

As it can be observed from Table 1, users' selections did not exactly follow the distribution of the impressions. Selections made on *Beta* recommender exhibited significantly lower NMG_{rel} than corresponding impressions (T-test p-value: 4.4e-5). Also, selections made on both MORS exhibited significantly lower NMG_{div} and NMG_{nov} and simultaneously higher NMG_{rel} (p-values < 2.6e-14). Note that while the decrease of selection's NMG_{rel} w.r.t. *Beta* may seem modest, it is due to a very narrow distribution of NMG_{rel} on impressions.

[8]The procedure was incremental, i.e., in each session, we only considered those selections, for which the re-trained *Beta* provided an impression.

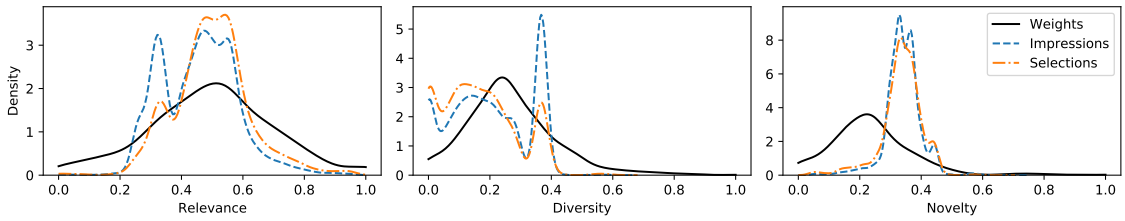

Fig. 2. Comparison of distributions for user-defined propensity weights, and relative marginal gains on impressions and selections.

Differences between selection and impression distributions naturally lead to the question of to what extent this is connected to the users' self-proclaimed propensity weights. Figure 2 depicts distributions of propensity weights and $NMGs$[9] of impressions and selections. Notably, for impressions, the distribution lacks the segment of very high/low values despite the demand expressed in propensity weights. This is mostly due to the existing covariance among selected objectives, which, e.g., prevents from finding diverse items without certain levels of novelty. More importantly, note that both relevance and diversity metrics were under-represented in impressions, if compared to the propensity weights. However, while for relevance, users tend to balance this bias back by only rarely selecting items with very low relevance (see the spike towards the left side of Figure 2), for diversity, items with low NMG_{div} are over-sampled and thus the difference between selection behavior and self-proclaimed propensity amplifies.

Despite the fact that users had the freedom of choosing objective weights at their discretion, their selection behavior significantly differs from the self-proclaimed propensity towards diversity, amplifying the existing bias of impression data. Seemingly, users overstated their propensity towards diversity. To some extent, we corroborated this hypothesis by analyzing the satisfaction of users as a factor of their propensity towards diversity. Users with a below-median propensity towards diversity on average selected more items from MORS recommendations (3.28 vs. 2.51, p-value 2.2e-7) and also provided higher overall ratings for MORS recommendations (2.84 vs. 2.64, p-value: 0.001). Originally, we expected that based on these inferior results, users would tend to converge towards lower weights for the diversity objective. However, no such evidence was found in the dataset.

This observation could have several causes. The limited number of sessions might be simply insufficient to learn the dependencies between objective weights and self-perceived satisfaction, let alone that the dependence might vary through time as illustrated in Section 3.1. Nonetheless, the misconception or misunderstanding on the level of objective semantics and/or item's marginal gains w.r.t. these objectives may play an important role too. The post-study questionnaire provided some leads on this factor. User's overall user satisfaction (i.e., *"Overall, I am satisfied with the recommender."*) was correlated with the information sufficiency (*"The information provided for the recommended movies was sufficient to judge whether I gonna like them."*, Pearson's correlation: 0.42) and the ability to state one's preferences (*"I was not able to describe my preferences w.r.t. relevance, diversity, and novelty."*, Pearson's correlation: -0.43). Also, while evaluating the user-perceived fulfillment of individual objectives, we found that positive answers on *"The movies recommended to me matched my interests."* implied no significant relations to the estimated relevances. Furthermore, while the positive answers on *"The recommended movies were novel to me."* implied some increase of the novelty metrics (0.985 vs. 0.981 for mean popularity complement and 0.798 vs. 0.751 for temporal novelty), the

[9]In order to make NMGs comparable with propensity weights, we re-scaled them to maintain unit sum object-wise.

magnitude of improvement was much higher for users who answered positively on *"The recommended movies were diverse."*: 0.986 vs. 0.976 for mean popularity complement and 0.795 vs. 0.716 for temporal novelty.

We can conclude the level of misconception between objective metrics and users' perception of these qualities is substantial. This is in line with the observations in related studies, e.g., [5]. Some parts of the post-study questionnaire suggest that this issue may be mitigated by better explanations (i.e., more informative descriptions) of recommended items. One option would be to visualize the degree, to which items fulfill individual objectives. This would allow users to better link their perception with underlying metrics and, e.g., help to adapt the self-proclaimed propensities to this knowledge.

4 CONCLUSIONS AND LIMITATIONS

In this paper, we conducted a user study focused on discovering the dependencies between users' interactions on items with certain beyond-accuracy properties and users' self-proclaimed propensities towards these beyond-accuracy criteria. We observed a considerable drift between both statistics and investigated the possible causes. We also provided some evidence of the benefits of MORS, despite not being the favored option from the user's (short-term) perspective.

The study had several limitations, which we plan to address in the future. First, only a modest volume of recommending sessions was conducted with no time in between. This prevents us from measuring the preference drifts [18] and/or contextual dependencies in long-term impact analysis. Also, users might not have enough time to stabilize their propensities towards individual objectives. Therefore, our future work should include studies with longer trial periods and sufficient time in between. Second, the choice of beyond-accuracy objectives as well as their particular definitions might affect the results. We plan to address this by a future study with a wider set of beyond-accuracy objectives. Similarly, we plan to investigate the impact of particular UIs for setting propensity weights. Third, the fact that *Beta* RS was trained w.r.t. all selections correspond to the situation, where an ensemble model is used. While this is plausible, we would also like to observe the effect of independent evolution for all RS.

Last but not least, the study was rather limited in terms of its size. Only one dataset, only one relevance-based RS, and only around 100 participants were employed. This might impact the stability of the results. Also, a rather strict filter on train set users and items was employed, resulting in a modest volume of candidate items. As such, it may prove difficult to find suitable recommendations in sequential settings without repetition. To mitigate these limitations, we plan to conduct an extensive set of follow-up studies comprising larger pools of participants, more evaluated RS variants, and more domains, including large-scale ones.

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. We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR).

REFERENCES

[1] Ludovico Boratto and Mirko Marras. 2021. Advances in Bias-aware Recommendation on the Web. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 1147–1149. https://doi.org/10.1145/3437963.3441665

[2] Keith Bradley and Barry Smyth. 2001. Improving recommendation diversity. In *Proceedings of the twelfth Irish conference on artificial intelligence and cognitive science, Maynooth, Ireland*, Vol. 85. Citeseer, 141–152.

[3] Pablo Castells, Neil Hurley, and Saúl Vargas. 2022. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, 603–646. https://doi.org/10.1007/978-1-0716-2197-4_16

[4] Mehdi Elahi, Mouzhi Ge, Francesco Ricci, David Massimo, and Shlomo Berkovsky. 2014. Interactive Food Recommendation for Groups. In *8th ACM Conference on Recommender Systems, RecSys 2014*. CEUR-WS.

[5] Soude Fazeli, Hendrik Drachsler, Marlies Bitter-Rijpkema, Francis Brouns, Wim van der Vegt, and Peter B. Sloep. 2018. User-Centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg. *IEEE Transactions on Learning Technologies* 11, 3 (2018), 294–306. https://doi.org/10.1109/TLT.2017.2732349

[6] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2 (2021), 100–126. https://doi.org/10.1016/j.aiopen.2021.06.002

[7] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. 2022. Toward Pareto Efficient Fairness-Utility Trade-off in Recommendation through Reinforcement Learning. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). ACM, 316–324. https://doi.org/10.1145/3488560.3498487

[8] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (dec 2015), 19 pages. https://doi.org/10.1145/2827872

[9] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*. 263–272. https://doi.org/10.1109/ICDM.2008.22

[10] Dietmar Jannach. 2022. Multi-Objective Recommendation: Overview and Challenges. In *Proceedings of the 2nd Workshop on Multi-Objective Recommender Systems co-located with 16th ACM Conference on Recommender Systems (RecSys 2022), Seattle, WA, USA, 18th-23rd September 2022 (CEUR Workshop Proceedings, Vol. 3268)*, Himan Abdollahpouri, Shaghayegh Sahebi, Mehdi Elahi, Masoud Mansoury, Babak Loni, Zahra Nazari, and Maria Dimakopoulou (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-3268/paper1.pdf

[11] Michael Jugovac, Dietmar Jannach, and Lukas Lerche. 2017. Efficient optimization of multiple recommendation quality factors according to individual user tendencies. *Expert Systems with Applications* 81 (2017), 321–331. https://doi.org/10.1016/j.eswa.2017.03.055

[12] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 9748–9758.

[13] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 624–632. https://doi.org/10.1145/3442381.3449866

[14] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. 2019. A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 20–28. https://doi.org/10.1145/3298689.3346998

[15] Mohammadmehdi Naghiaei, Hossein A. Rahmani, and Yashar Deldjoo. 2022. CPFair: Personalized Consumer and Producer Fairness Re-ranking for Recommender Systems. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 770–779. https://doi.org/10.1145/3477495.3531959

[16] Ladislav Peska and Patrik Dokoupil. 2022. 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* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1963–1968. https://doi.org/10.1145/3477495.3531787

[17] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11)*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/2043932.2043962

[18] Nakarin Sritrakool and Saranya Maneeroj. 2021. Personalized Preference Drift Aware Sequential Recommender System. *IEEE Access* 9 (2021), 155491–155506. https://doi.org/10.1109/ACCESS.2021.3128769

[19] Dusan Stamenkovic, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas. 2022. Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). ACM, 957–965. https://doi.org/10.1145/3488560.3498471

[20] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11)*. Association for Computing Machinery, New York, NY, USA, 109–116. https://doi.org/10.1145/2043932.2043955

[21] Zhaoyuan Wang, Chuishi Meng, Shenggong Ji, Tianrui Li, and Yu Zheng. 2020. Food package suggestion system based on multi-objective optimization: A case study on a real-world restaurant. *Applied Soft Computing* 93 (2020), 106369. https://doi.org/10.1016/j.asoc.2020.106369

[22] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. 2022. Multi-FR: A Multi-objective Optimization Framework for Multi-stakeholder Fairness-aware Recommendation. In *Transactions on Information Systems (TOIS)*. ACM.

[23] Yong Zheng and David (Xuejun) Wang. 2022. A survey of recommender systems with multi-objective optimization. *Neurocomputing* 474 (2022), 141–153. https://doi.org/10.1016/j.neucom.2021.11.041