Measuring the Impact of Explanation Bias: A Study of Natural Language Justifications for Recommender Systems

Krisztian Balog
Google
Stavanger, Norway
krisztianb@google.com

Filip Radlinski
Google
London, UK
filiprad@google.com

Andrey Petrov
Google
London, UK
apetrov@google.com

ABSTRACT

Despite the potential impact of explanations on decision making, there is a lack of research on quantifying their effect on users' choices. This paper presents an experimental protocol for measuring the degree to which positively or negatively biased explanations can lead to users choosing suboptimal recommendations. Key elements of this protocol include a preference elicitation stage to allow for personalizing recommendations, manual identification and extraction of item aspects from reviews, and a controlled method for introducing bias through the combination of both positive and negative aspects. We study explanations in two different textual formats: as a list of item aspects and as fluent natural language text. Through a user study with 129 participants, we demonstrate that explanations can significantly affect users' selections and that these findings generalize across explanation formats.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Empirical studies in HCI**; • **Information systems** → **Recommender systems**.

KEYWORDS

Explainable recommendation; natural language justifications; evaluating explanations; explanation types; explanation bias

1 INTRODUCTION

Recommender systems have become pervasive in modern society, leading to many studies of the degree of trust people place in provided recommendations [14, 20]. As a result, significant attention has been paid to equip these systems with explanation facilities to help users make informed decisions [25, 31, 34]. In one of the earliest studies, Herlocker et al. [17] showed that explanations can make it more likely that people will adopt the recommendations made. Yet,

studies measuring the eventual satisfaction with recommendations found that certain types of explanations can cause users to over- or underestimate the real value of a recommended item [3, 11]. Even though explanations are known to be able to significantly affect the decision-making process of users [31], significant gaps remain when it comes to measuring and understanding their effects [28]. The bulk of studies of explanations focus on the subjective question of how people perceive the recommendations [1, 5, 11, 22, 25]. Less attention has been paid to how to objectively quantify the degree to which explanations affect the choices that people make. This is closely related to the concept of *persuasiveness*, which is the ability of an explanation "to convince the user to accept or disregard certain items" [11]. Past research has differentiated between over- and underestimate-oriented persuasiveness of *explanation types* [3, 11]. Instead, we study in quantitative detail how bias in the *content* of natural language explanations can impact users' choices.

A key difference to past work is that prior research has focused almost exclusively on explanations that highlight positive aspects of items, i.e., endorse a selection, perhaps by drawing similarities to other items the user has previously indicated that they liked. Explanations can also help users make decisions by highlighting why *not* to select something. We perform the first study that intentionally biases explanations towards positive or negative aspects of items in a controlled manner in the context of items where we have a prior understanding of which items users should find most relevant. We can thus quantify the degree to which positively or negatively biased explanations lead to users choosing suboptimal recommendations. To achieve this, we are inspired by the work of Musto et al. [22] who "*processed and analyzed the reviews in order to obtain a set of characteristics that are often discussed in the reviews (with a positive sentiment, of course) and can induce the user in enjoying the recommended item.*" However, we identify positive as well as *negative* aspects—as argued by Bilgic and Mooney [3], "*the goal of an explanation should not be to 'sell' the user on the item but rather to help the user to make an informed judgment.*"

The primary research question we address is how to quantify the degree to which explanations influence user decisions, either in the positive or negative direction when presented with a set of recommendations from which to choose. Our key contribution is the design of an experiment protocol that allows for a quantitative study of the impact of explanations on users' choices. Our second contribution is a preliminary analysis of the results obtained with 129 subjects. We provide baseline statistics showing how often users presented with biased explanations can end up selecting less relevant suggestions. Beyond this, we study the effect of the *format* of explanations, directly comparing two common formats, namely lists of item aspects and natural language text. We show

that while these formats lead to a similar overall behavior, the effect of explanation bias is more pronounced in case of aspect lists than for natural language text. These may be considered as factors when designing explainable recommender systems. Overall, this work provides a framework for measuring and optimizing explanations to help guide users to make informed decisions.

2 RELATED WORK

Research on explainable recommendation has focused on different dimensions, including (1) explanation goal (e.g., transparency, trust, effectiveness, persuasiveness), (2) style (e.g., content-based, collaborative-based, knowledge/utility-based), (3) scope (e.g., user model, process, recommended item), and (4) format (e.g., textual, visual) [25, 31, 34]. Another way of categorization is by the method used to generate explanations: self-explainable recommender models vs. post-hoc explanations (i.e., *justifications*) [4]. Our approach can be classified as content-based, post-hoc, natural language explanation generation. We do not aim to explain *why* a given item was recommended, but rather provide the user with additional information (relevant characteristics of items) with the purpose of aiding them in their decision making.

Various graphical and textual explanation formats have been considered in the past [11, 17], with natural language being the most popular both historically [25] and in recent years [1, 5, 22, 23, 27]. Text-based explanations range from tags or keywords [3, 32] to single or multiple sentences [1, 5, 22, 27, 30]. In this work, we consider both short descriptive phrases (aspects) and fluent natural language summaries as explanations. Unique to our approach is that the summaries are generated directly from the aspects, allowing for a direct comparison between the two explanation formats.

Reviews have been exploited for improving recommendations, leading to a line of work on review-aware recommender systems [6, 18], naturally endowing them with a higher degree of explainability and transparency [16]. In contrast, our work continues the thread of research on leveraging reviews for the purpose of generating post-hoc recommendation justifications, treating the recommender engine as a black box [5, 21–23, 27]. Muhammad et al. [21] highlight the most important features (pros and cons) of an item that are likely to matter to the user (based on their own reviews). Similarly, Chen and Wang [7, 8] extract sentiments on specific product attributes and use them to present the user with alternatives to a given recommendation and explaining the trade-offs, i.e., which attributes would be improved and which ones would be compromised. Chang et al. [5] employ a multi-step pipeline for generating natural language explanations using crowdsourcing, by (1) refining algorithmically generated tag clusters, (2) writing explanations for these clusters by synthesizing review text, and (3) selecting the best explanations by voting. We follow a similar approach, but present crowd workers with a significantly simpler task: given a set of reviews, they only need to extract positive and negative aspects from them. We then either present these aspects as a list, or turn them into fluent natural language text using state-of-the-art neural language modeling techniques. Ni et al. [23] identify review segments that can serve as justifications and explore the use of neural language models to generate convincing and diverse justifications. Penha et al. [27] choose a helpful sentence for an item from its review,

predicted by a classifier, and use that to generate a template-based explanation of an item or a pair of items. Most closely related to ours is the work by Musto et al. [22], which presents a fully automated pipeline for generating natural language justifications by (1) extracting a set of aspects that characterize the item, (2) identifying the most relevant ones, and (3) extracting and aggregating review sentences discussing these aspects. Though there are similarities in the explanation generation workflow, their focus is on the algorithmic aspects of automation, while ours is on understanding the *impact* of explanations. Our aspects are short descriptive phrases, not limited to nouns as in [22], which we extract and curate manually. Most importantly, we include aspects with negative sentiment as well, not only positive ones, in the explanations.

Evaluation of explanations may target interface-related aspects [3, 8, 11, 17, 21], quantifiable properties of generation approaches (e.g., readability of natural language [10] or fidelity of post-hoc explanations [26]), or subjective perceptions of quality (in terms of transparency, trust, effectiveness, etc.) through questionnaires [1, 5, 11, 22, 25]. Most relevant to our work is the evaluation protocol proposed in [3] (also followed in [11]) for measuring the *persuasiveness* of explanation types, in terms of differences between an initial rating, given based on the explanation, and a second rating, given after the "consumption" of the item. Consumption is approximated by subjects receiving more details (description or user reviews) about the item. Then, they are asked adjust their rating based on this additional information. Our experiment design is more realistic in that we do not assume item consumption, but rather measure the impact of explanations via the item selection choices users make, with a prior understanding of what they would most likely pick.

3 STUDY DESIGN

This section presents the study we designed in order to answer the following two research questions:

- **RQ1** How much can explanations influence users' decisions when presented with a set of item recommendations?
- **RQ2** How big a role does the presentation format of explanations play in the results?

Our experiment centers around a set of item selection tasks, with participants asked to choose among three items that are presented to them as recommendations. These items are selected, based on an initial preference elicitation phase, such that we know which of the three items the user would probably like the most and which one the least. By applying a controlled amount of bias via explanations to a randomly chosen item, we are able to measure how much users may be influenced in their choices—it is of particular interest to observe how often they would be inclined to pick a different item than that which they would most likely choose (statistically speaking). Using a within-subject design, participants are exposed to different experimental conditions, varying the amount of bias and the format of explanations.

3.1 Overview

The study is conducted in the movies domain. As watching films is one of the top media consumption and entertainment activities,

Stage 1: Item consumption and preference elicitation

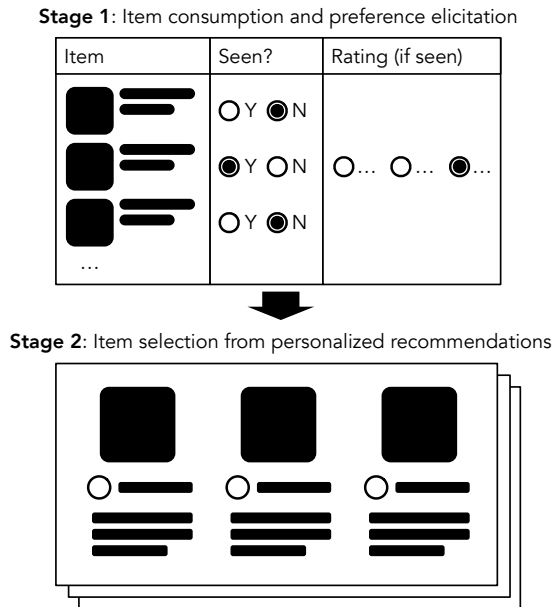**Stage 2**: Item selection from personalized recommendations

Figure 1: Overview of our experiment design.

most people can easily relate to it without any training or prerequisite knowledge. Indeed, movies have been one of the most studied domains for explainable recommendation [1, 2, 5, 11, 12, 17, 22, 30, 32]. Note that even though the instructions are tailored specifically to movies, the proposed design is generic and can be applied in other domains.

To be able to provide participants with personalized recommendations, the study is divided into two stages, as shown in Fig. 1.[1] In Stage 1, participants indicate for each movie, in our pool of 400 candidates items, whether they have watched it, and if yes, how much they liked it (on a three-point scale of disliked, neutral, and liked). In Stage 2, participants complete a total of 12 item selection tasks, corresponding to various experimental conditions in a random order (cf. Table 1). In each task, they are offered three movies to choose from, selected among the items they have not watched yet. For each movie the title, poster, and synopsis are shown, followed (optionally) by an explanation (see Fig. 6 in Appendix A.3). The item selection task is situated in the following scenario: *Imagine that you are traveling alone, and while in transit, you have been offered a free movie rental from a service that only allows you to choose from three possible options. In each of the following tasks, select the movie you would prefer to watch based on the information provided.*

We hypothesize that the explanations accompanying the recommendations can influence users' item selection choices. To test this, we contrast a baseline setting, showing neutral explanations for all three items, with a "biased" setting: showing neutral explanations for two of the items, while making the explanation for the third item overall more positive or negative in a controlled manner. Next, we provide details on each element of the experiment design.

Table 1: Experimental conditions. Each participant experiences each of the listed conditions exactly once (in random order).

ID	Bias	Explanation	Aspects$^+$	Aspects$^-$
Baselines				
#1	No bias	-	-	-
#2				
#3		Itemized	2	2
#4		Fluent-NL		
One random item biased by a more positive/negative explanation				
#5	Bias +	Itemized	3	1
#6		Fluent-NL		
#7	Bias -	Itemized	1	3
#8		Fluent-NL		
#9	Bias ++	Itemized	4	0
#10		Fluent-NL		
#11	Bias --	Itemized	0	4
#12		Fluent-NL		

3.2 Item Collection

We use the MovieLens 25M collection [15] along with the Movies and TV subset of the Amazon Reviews 2018 dataset [23]. Mappings between the two are provided by the Reviews2Movielens dataset (v2 mappings) released in [33]. A total of 400 movies are sampled, using a slightly modified version of the stratified sampling approach employed in [2]: (i) the top 150 movies of all time by the number of ratings received and (ii) a random movie for each year between 1992 and 2016, and for each of the top 10 most popular genres (action, adventure, documentary, comedy, crime, drama, horror, romance, sci-fi, and thriller), that is not already in the top-150 set.[2] We require each movie to have at least 100 associated reviews.

3.3 Generating Personalized Recommendations

Study participants are presented with a set of item selection tasks. In each task, they need to make a choice between three items x_a, x_b, and x_c, shown in random order. These items are selected, based on the initial preference elicitation phase, such that the user's estimated preference ordering is $x_a > x_b > x_c$. For that, we employ an ensemble of three established recommender algorithms, representing different classes of collaborative filtering approaches that performed best in a prior user study on movie recommendations [2]: Item-based k-Nearest Neighbors [29], Weighted Regularized Matrix Factorization [19], and a Sparse Linear Method [24]. The predictions of these algorithms are combined into an ensemble recommendation using a consensus-based voting system (Borda count). This ensemble is expected to yield better performance than any of the individual recommenders. Crucially, only movies not yet seen by the user are eligible for recommendation, as showing already seen items might affect the measurements of the impact of explanations [1].

Given a total of $m = 12$ item selection tasks to be completed by each participant (cf. Table 1), $m \times 3$ items are sampled as follows.[3] Let n denote the number of movies the person has not seen yet, which are sorted by recommendation score, such that x_1 is the highest and x_n is the lowest ranked suggestion. These items are

---

[1]Additionally, participants are asked to fill out a post-task questionnaire to allow for a qualitative analysis of their preferences in the future.

[2]If it was not possible, then a random movie is sampled from the same year.
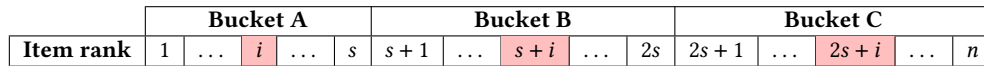[3]Participants with fewer than 3×12 not yet seen movies are excluded from the study.

	Bucket A					Bucket B				Bucket C			
Item rank	1	...	i	...	s	$s+1$	...	$s+i$	...	$2s$	$2s+1$	...	$2s+i$	...	n

Figure 2: Sampling of items for item selection tasks to ensure the largest distance between sampled items across all item samples.

Figure 3: Examples of itemized and fluent-NL explanations, based on the same set of aspects.

Table 2: Controlled generation of explanation sentiment based on the number of positive and negative aspects mentioned.

	Extreme Negative (Bias --)	Negative (Bias -)	Neutral (No bias)	Positive (Bias +)	Extreme Positive (Bias ++)
Aspects$^+$	0	1	2	3	4
Aspects$^-$	4	3	2	1	0

divided into three approximately equal-sized ($s = \lfloor n/3 \rfloor$) buckets A, B, and C. To sample m sets of three items, the items picked in set $i \in [1..m]$ are: $x_a = x_i$, $x_b = x_{s+i}$, and $x_c = x_{2s+i}$. This ensures that the three items shown to the user are as far apart from each other as possible, in terms of recommendation score, across all item selection tasks; see Fig. 2 for a visual explanation. The m sets of items (x_a, x_b, and x_c) are assigned randomly to the m experimental conditions.

3.4 Generating Explanations

To quantify how much explanations can influence user decisions, we need a controlled way of generating explanations which are by default neutral, but can be biased to have a more positive or negative overall sentiment. To operationally define what it takes for an explanation to be overall neutral/positive/negative we assume that for each item $M = 4$ positive and M negative aspects have been identified. An aspect in this context is a short natural language text (typically 1–5 words in length) that expresses why a movie might be liked or disliked, such as "action-packed," "quite corny and unrealistic," or "classic literature brilliantly realized." The sentiment of an explanation can then be controlled by adjusting how many positive and negative aspects it mentions, with Aspects$^+$ = Aspects$^-$ = $M/2$ representing a neutral explanation (cf. Table 2). Aspects are extracted from reviews, as detailed below in Section 3.4.1.

We consider two explanation formats: itemized and fluent natural language text. *Itemized* explanations are comprised of a list of aspects with their corresponding sentiment symbolized by a thumbs up or down icon. Alternatively, the *fluent-NL* format presents the same information as fluent natural language description; see Fig. 3 for an illustration. The generation of fluent-NL explanations from a list of item aspects is detailed below in Section 3.4.2.

3.4.1 Aspect Extraction. Positive and negative aspects are extracted manually using crowdsourcing. Given 10 positive (4-5 stars) and 10 negative reviews (1-2 stars), workers are asked to find aspects that complete the sentence: *This movie may be liked (disliked) because it is/has/contains ___.* The goal is not to be exhaustive, but rather focus on high data quality (i.e., favor precision over recall). To ensure that, extracted aspects are further checked and manually filtered by authors of the paper to remove too harsh or offensive language, aspects that are too generic (e.g., excellent, terrible) or are not about movie itself (concern price, delivery, medium, etc.). Movies with fewer than 4 positive and 4 negative aspects after filtering are removed from the recommendation pool (31 in total). Further details are in Appendix A.1.

3.4.2 Fluent-NL Explanation Generation. We turn the selected aspects into fluent natural language sentences using a large language model with state-of-the-art few-shot performance (PaLM [9], 62b parameter model). A separate prompt is created for each combination of Aspects$^+$ and Aspects$^-$, containing three hand-crafted training examples.

3.5 Participant Condition Assignment

The study is performed through crowdsourcing via a web-based platform. Each participant is presented with each of the 12 experimental conditions, shown in Table 1, exactly once in a random order. In fact, there are only 11 unique conditions, as #1 and #2 are the same, but this "no explanations" setting is shown twice in order to establish a robust baseline. The other baseline is to show neutral explanations, in two different formats (#3–#4). In the remaining conditions (#5-#12) one of the three items shown to the rater is randomly selected to be "biased," by changing the number of positive and negative aspects in the explanation that accompanies that item. The explanations for the other two items stay neutral, i.e., showing exactly two positive and two negative aspects. These conditions allow us to make comparisons between different explanation types, i.e., no explanation (#1–#2) vs. itemized (#3, #5, #7, #9, #11) vs. fluent-NL (#4, #6, #8, #10, #12) explanations, and between the amount of bias, i.e., no bias (#1–#4) vs. moderate (#5–#8) vs. large (#9–#12) bias.
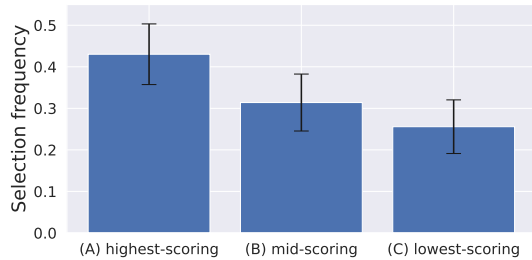
Figure 4: Baseline condition, no explanations.

Figure 5: Baseline condition, neutral explanations.

We acknowledge that listing positive aspects first, followed by negative ones, or the other way around, might have an impact. At the same time, to avoid further complicating the design by introducing yet another dimension, we control for this variable by balancing the two possible settings, i.e., positives-first and negatives-first. Specifically, when both positive and negative aspects are displayed (#3–#8), we make those fully balanced for each participant as well as across all participants by cycling through pre-defined sequences in a Latin Square-like design (see Table 5 in Appendix A.2).

4 RESULTS

We now present quantitative results that show how biasing explanations that accompany recommendations directly affects users' preferences, as expressed by their item selections.

4.1 Participants

A total of 129 participants participated in our study who are paid contractors, and received a standard contracted wage (complying with living wage laws in their country of employment). All of them are US-based and native English speakers. In terms of gender, they are 62.7% women, 35.7% men, 1.6% prefer not to say. Their distribution by age: 5.6% 18–24, 29.4% 25–34, 32.5% 35–44, 19.8% 45–54, 9.5% 55–64, 3.2% over 65. Their self-reported amount of average time spent per week watching movies: 5.7% <2 hours, 37.9% 2–5 hours, 29% 6–10 hours, 16.9% 11-16 hours, 10.5% > 16 hours. Participants on average took 41 seconds to rate a batch of five movies in Stage 1 (seen/liked), and 43 seconds to choose the movie they would want to watch from the three recommendations in Stage 2.

4.2 Effect of the Presence of Explanations

Recall that in each instance, participants are provided with three recommendations where we know which the raters should prefer the most. Our first test validates that the preferences obey the expected order, and whether the presence of neutral explanations affects this order. Specifically, let the three recommendations be $x_a \in A$, $x_b \in B$ and $x_c \in C$, where A, B, and C are buckets with the highest-scoring, mid-scoring, and lowest-scoring items, respectively (cf. Fig. 2). Let the relative frequency with which each bucket is chosen (corresponding to users' item selections), referred to as their *selection frequency*, be denoted as p_A, p_B, and p_C respectively. We then have an expected ordering of $p_A > p_B > p_C$.
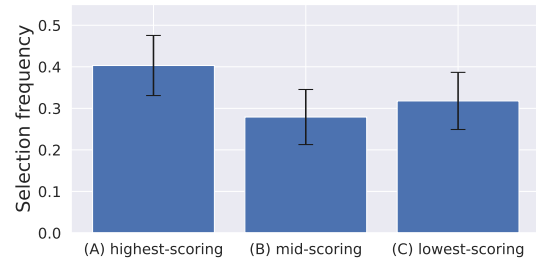
When explanations are not shown, we find that across all participants, $p_A = 0.43 \pm 0.07$, $p_B = 0.31 \pm 0.07$ and $p_C = 0.26 \pm 0.06$ with a 95% confidence interval, computed using the Goodman method [13]; see Fig. 4. Thus, the expected ordering holds, and the difference between the highest- and lowest-scoring buckets is statistically significant.

When neutral explanations are included, we observe $p_A^n = 0.4 \pm 0.07$, $p_B^n = 0.28 \pm 0.07$, and $p_C^n = 0.32 \pm 0.07$, as shown in Fig. 5. While the highest-scoring bucket continues to receive the most selections, the relative ordering between buckets B and C, surprisingly, is now swapped. Also, the selection frequencies for all three buckets come closer together, resulting in overlapping confidence intervals. This means that despite the careful experiment design, explanations seem to have some uncontrolled effects. It could be, for example, that mid- and low-scoring items are not that well distinguished by the recommender system and it is a random effect due to noise in the data. However, the "well-behaving" baseline setting without explanations and the relatively large sample size (n=258) suggest otherwise. It could also be that neutral explanations that highlight both positive and negative aspects invite more "risky" selections by users, giving lower-ranked suggestion a try. In the remainder of our analysis, we take the neutral explanations setting (Fig. 5) as our baseline. However, the relative ordering of buckets B and C warrants further investigation.

4.3 Effect of Explanations Biased towards Positive or Negative

By adjusting the number of positive and negative aspects, we can bias explanations in a positive or negative direction either weakly (e.g., three positive aspects and one negative aspect) or strongly (e.g., four negative aspects and no positives). Table 3 shows the effect of bias on item selections (rows) depending on the position and direction of bias (columns). For simplicity, we do not distinguish between the amount of bias (weakly or strongly positive/negative) nor the type of explanation (fluent-NL or itemized), but report on aggregated counts.

Bias in a given bucket has an effect on selections both in the same bucket (highlighted as grey in Table 3) and in other buckets. For example, biasing positively items in C increases selections in C, but also decreases selections in A and B. Similarly, negatively biasing items in A drives selections down in A, while moving selections up in B. This intuitively makes sense, but there are a few exceptions when this expected behavior cannot be observed, e.g., negative

Table 3: Selection frequency of different buckets (rows) depending on the presence and direction of bias in explanations (columns). Grey cell background indicates when the bias happens in the same bucket as the selection. Green/red arrows show the change in selection frequency with respect to the no bias setting.

Selection	No bias	Bias in A		Bias in B		Bias in C	
		+/++	-/--	+/++	-/--	+/++	-/--
A	0.40 ± 0.07	0.49 ± 0.09▲	0.37 ± 0.09▼	0.37 ± 0.09▼	0.41 ± 0.09▲	0.36 ± 0.08▼	0.50 ± 0.09▲
B	0.28 ± 0.07	0.25 ± 0.08▼	0.34 ± 0.08▲	0.42 ± 0.09▲	0.30 ± 0.08▲	0.27 ± 0.08▼	0.29 ± 0.08▲
C	0.32 ± 0.07	0.26 ± 0.08▼	0.29 ± 0.08▼	0.21 ± 0.08▼	0.28 ± 0.08▼	0.38 ± 0.09▲	0.21 ± 0.07▼

Table 4: Selection frequency of different buckets (rows) depending on the presence and direction of bias in explanations (columns), for itemized explanations (top block) vs. fluent-NL explanations (bottom block); cell annotations are the same as in Table 3.

Selection	No bias	Bias in A		Bias in B		Bias in C	
		+/++	-/--	+/++	-/--	+/++	-/--
Itemized explanations							
A	0.40 ± 0.10	0.49 ± 0.13▲	0.35 ± 0.11▼	0.39 ± 0.12▼	0.47 ± 0.13▲	0.34 ± 0.12▼	0.55 ± 0.12▲
B	0.28 ± 0.09	0.26 ± 0.11▼	0.34 ± 0.11▲	0.45 ± 0.13▲	0.19 ± 0.11▼	0.26 ± 0.11▼	0.27 ± 0.11▼
C	0.32 ± 0.10	0.25 ± 0.11▼	0.31 ± 0.11▼	0.16 ± 0.10▼	0.33 ± 0.12▲	0.40 ± 0.12▲	0.19 ± 0.10▼
Fluent-NL explanations							
A	0.40 ± 0.10	0.49 ± 0.12▲	0.40 ± 0.13	0.34 ± 0.12▼	0.36 ± 0.12▼	0.38 ± 0.12▼	0.45 ± 0.12▲
B	0.28 ± 0.09	0.24 ± 0.10▼	0.33 ± 0.12▲	0.39 ± 0.13▲	0.40 ± 0.12▲	0.27 ± 0.11▼	0.31 ± 0.12▲
C	0.32 ± 0.10	0.26 ± 0.10▼	0.27 ± 0.12▼	0.27 ± 0.12▼	0.24 ± 0.11▼	0.35 ± 0.12▲	0.23 ± 0.11▼

bias in A increases selections in B, but not in C. Despite these anomalies that remain to be investigated in the future, it is clear that biasing explanations has a large effect on the selections people make. Two extremes are worth noting: (1) negatively biasing the most relevant recommendation reduces the selection of what is believed to be the best recommendation by 9%, and (2) positively biasing the least relevant recommendation increases the selection of that item by twice as much, almost 19%, compared to the no bias baseline. Notice that the selection frequency of the least relevant suggestion with a positive bias (0.38 ± 0.09) reaches that of the most relevant suggestion with a negative bias (0.37 ± 0.09).

4.4 Effect of Explanation Format: Fluent-NL vs. Itemized Explanations

In our design, explanations were presented in two ways: as a list of attributes, and as fluent text that mentions the same attributes. The last question we ask is: To what extent do our findings depend on the particular explanation format? Table 4 breaks down the previous results by explanation format. Our main findings are as follows. First, there is no difference in results in the no bias setting. Second, when explanations are biased, itemized explanations behave more "as expected," i.e., when bias happens in the same bucket as the selection (grey cells), then positive bias always means an increase and negative bias always causes a drop in selection frequency. This is not the case for fluent-NL explanations. Third, we observe that the differences between the positive and negative bias settings within a given bucket tend to be much larger in case of itemized explanations. This intuitively makes sense, as the positive and negative aspects are made explicit with visual thumbs up/down icons, while fluent text can be more prone to hiding differences. Despite these differences, results indicate a consistent pattern of change across the two explanation formats: the arrows indicating change point in the same direction in 14 out of the 18 cells.

5 CONCLUSION

We have designed a protocol that allows for a quantitative study of the impact of explanations on users' choices in item recommendation. Key elements of this design include preference elicitation that allows for the generation of personalized recommendations, manual identification and extraction of item aspects to include in explanations, a controlled way of introducing bias via the combination of both positive and negative aspects, and the presentation of explanations in two different textual formats. We have conducted a user study and showed that explanations can indeed have a large effect on the item selections that people make, and that these findings generalize across the two explanation formats. The results have also yielded some unexpected findings that warrant further investigation in future work. We also plan to conduct a more detailed statistical analysis of the results and perform qualitative evaluation based on post-survey responses. Further, the differences in terms of absolute impact between itemized and fluent natural language explanations suggest that the specific wording of the latter might play a role. Measuring whether slight differences in phrasing have an impact is an interesting topic for future research. Finally, we focused on movie recommendations, yet our approach is generalizable to other domains where users rely on automatic suggestions due to the size of the item collection (e.g., books, music, recipes). It would be interesting to repeat the experiment in other domains.

REFERENCES

[1] Krisztian Balog and Filip Radlinski. 2020. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 329–338.

[2] Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. 2019. Transparent, Scrutable and Explainable User Models for Personalized Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. 265–274.

[3] Mustafa Bilgic and Raymond J. Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces.* 13–18.

[4] Or Biran and Courtenay V. Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence.*

[5] Shuo Chang, F. Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-Based Personalized Natural Language Explanations for Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16).* 175–182.

[6] Li Chen, Guanliang Chen, and Feng Wang. 2015. Recommender Systems Based on User Reviews: The State of the Art. *User Model. User-Adapt. Interact.* 25, 2 (jun 2015), 99–154.

[7] Li Chen and Feng Wang. 2014. Sentiment-Enhanced Explanation of Product Recommendations. In *Proceedings of the 23rd International Conference on World Wide Web ('14).* 239–240.

[8] Li Chen and Feng Wang. 2017. Explaining Recommendations Based on Feature Sentiments in Product Reviews. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17).* 17–28.

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 [cs.CL]

[10] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic Generation of Natural Language Explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion (IUI '18).*

[11] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. *Int. J. Hum. Comput. Stud.* 72, 4 (2014), 367–382.

[12] Azin Ghazimatin, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. ELIXIR: Learning from User Feedback on Explanations to Improve Recommender Models. In *Proceedings of the Web Conference 2021 (WWW '21).* 3850–3860.

[13] Leo A. Goodman. 1965. On Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics* 7, 2 (may 1965), 247–254.

[14] Jason L. Harman, John O'Donovan, Tarek Abdelzaher, and Cleotilde Gonzalez. 2014. Dynamics of Human Trust in Recommender Systems. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14).* 305–308.

[15] 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.

[16] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. TriRank: Review-Aware Explainable Recommendation by Modeling Aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15).* 1661–1670.

[17] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00).* 241–250.

[18] María Hernández-Rubio, Iván Cantador, and Alejandro Bellogín. 2019. A Comparative Analysis of Recommender Systems Based on Item Aspect Opinions Extracted from User Reviews. *User Model. User-Adapt. Interact.* 29 (apr 2019), 381–441.

[19] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM '08).* 263–272.

[20] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* 1–12.

[21] Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16).* 256–260.

[22] Cataldo Musto, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2019. Justifying Recommendations through Aspect-Based Sentiment Analysis of Users Reviews. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '19).* 4–12.

[23] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.* 188–197.

[24] Xia Ning and George Karypis. 2011. SLIM: Sparse Linear Methods for Top-N Recommender Systems. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining (ICDM '11).* 497–506.

[25] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Model. User-adapt. Interact.* 27, 3-5 (2017), 393–444.

[26] Georgina Peake and Jun Wang. 2018. Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18).* 2060–2069.

[27] Gustavo Penha, Eyal Krikon, and Vanessa Murdock. 2022. Pairwise Review-Based Explanations for Voice Product Search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22).* 300–304.

[28] Alessandro Piscopo, Oana Inel, Sanne Vrijenhoek, Martijn Millecamp, and Krisztian Balog. 2022. Report on the 1st Workshop on Measuring the Quality of Explanations in Recommender Systems (QUARE 2022) at SIGIR 2022. *SIGIR Forum* 56, 2 (dec 2022).

[29] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web (WWW '01).* 285–295.

[30] Nava Tintarev and Judith Masthoff. 2012. Evaluating the Effectiveness of Explanations for Recommender Systems. *User Model. User-Adapt. Interact.* 22 (oct 2012), 399–439.

[31] Nava Tintarev and Judith Masthoff. 2015. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook* (2nd ed.), Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer US, Chapter 10, 353–382.

[32] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: Explaining Recommendations Using Tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09).* 47–56.

[33] Yury Zemlyanskiy, Sudeep Gandhe, Ruining He, Bhargav Kanagal, Anirudh Ravula, Juraj Gottweis, Fei Sha, and Ilya Eckstein. 2021. DOCENT: Learning Self-Supervised Entity Representations from Large Document Collections. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL '21).* 2540–2549.

[34] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14, 1 (2020), 1–101.

A APPENDIX: STUDY DESIGN

We provide further details on the study design and data preparation.

A.1 Aspect Extraction

The data collected using crowdsourcing has been automatically cleaned and pre-filtered; this includes fixing capitalization, removing trailing whitespace, filtering too long aspects, and removing aspects that are substrings of other aspects. However, the data needed further manual cleaning and filtering by the paper authors, as not all extracted aspects fit the template (i.e., human workers did not follow the instructions closely enough), aspects may be too harsh or offensive, or sound too personal to be used as explanations. As part of the manual cleaning process, some aspects were slightly rewritten and near-duplicates were removed. Of the 9,635 movie-aspect pairs collected originally, 7,414 remained after automatic pre-filtering, and 5,948 after the end of the manual filtering.

It is worth emphasizing that the recommendations are personalized, while the explanations accompanying them are not, i.e., all participants receiving the same recommendation under the same experimental condition will see the same explanation for that item, to ensure that there is no uncontrolled bias.

A.2 Explanations

When both positive and negative aspects are displayed (conditions #3–#8), we make those fully balanced for each participant as well as across all participants by cycling through the sequences shown in Table 5. These sequences follow a Latin Square design where the binary value P/N is determined by the least significant bit.

Figure 6: User interface for item selection (Stage 2).

Figure 7: User interface for item consumption and preference elicitation (Stage 1).

Table 5: Sequences determining whether to list positive or negative aspects first in explanations.

Sequence	Condition					
	#3	#4	#5	#6	#7	#8
1	N	P	P	N	N	P
2	P	N	N	P	P	N
3	N	P	P	N	N	P
4	P	N	N	P	P	N
5	N	P	P	N	N	P
6	P	N	N	P	P	N

A.3   User Interfaces

Figures 7 and 6 show screenshots of the user interfaces used in Stages 1 and 2, respectively.