Toward Benchmarking Group Explanations: Evaluating the Effect of Aggregation Strategies versus Explanation

Francesco Barile[1], Shabnam Najafian[2], Tim Draws[2], Oana Inel[2], Alisa Rieger[2], Rishav Hada[1] and Nava Tintarev[1]

[1]*Maastricht University, Netherlands*
[2]*TU Delft, Netherlands*

Abstract

In the context of group recommendations, explanations have been claimed to be useful for finding a satisfactory choice for all the group members and helping them agree on a common decision, improving perceived fairness, perceived consensus, and satisfaction. In this work, we present a preregistered evaluation of the impact of using social choice-based explanations for group recommendations (*i.e.*, explanations that intuitively describe the strategy used to generate the recommendation). Our objective is to conceptually replicate a previous study and investigate whether a) the used *aggregation strategy* or b) the *explanation* affected the most users' fairness perception, consensus perception, and satisfaction. Our results show that the participants are able to discriminate between the different strategies, assigning worse evaluations to the *Most Pleasure* strategy (which chooses the item with the highest of the individual evaluations). In addition to a condition with no (natural language) explanation, we introduce a more *detailed* social choice-based explanation, evaluating whether additional information about the strategy has a positive impact on the evaluation of the group recommendation. However, we surprisingly found *no* effect of level of explanations, either as a main effect or as an interaction effect with the aggregation strategy. Overall, our results suggest that users' perceptions of fairness, consensus, and satisfaction are primarily formed based on the *social choice aggregation strategies* for the studied group scenario. Our work also highlights the challenges of replication studies in recommender systems and discusses some of the design choices that may influence results when attempting to benchmark findings for the effectiveness of group explanations.

Keywords

Social Choice-based Explanations, Group Recommender Systems, Explainable Recommender Systems

1. Introduction

In many domains, such as online communities [1, 2], music, movies or TV programs [3, 4, 5, 6], and tourism [6, 7], people consume recommendations in groups rather than individually. Several approaches in the literature [4, 8, 7] propose social choice strategies, which combine the individual preferences of all group members and predict an item that is suitable for everyone. Each such aggregation strategy, however, has its trade-offs. As stated by Arrow's theorem [9], the performance of an aggregation strategy depends on the evaluation context, meaning that it

Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2021), September 25th, 2021, co-located with the 15th ACM Conference on Recommender Systems, Amsterdam, The Netherlands

is unlikely for an aggregation strategy to outperform other strategies in all situations. Nevertheless, understanding why particular items are recommended is not a trivial task, especially for group recommendations [10]. In general, explanations [11, 12] have been proposed as a means of describing why certain items are recommended. The adoption of explanations has proved to increase user acceptance of recommended items [13, 14]. In the context of group recommendations, however, the role of explanations is even more challenging. Multiple functions need to be met, besides explaining why certain items are recommended [15, 16] — to help users agree on a joint decision, as well as improve users' perceived fairness, perceived consensus, and satisfaction [15, 5, 17].

To the best of our knowledge, however, only a few studies [17] have focused on generating and evaluating explanations based on social choice aggregation strategies to increase fairness and consensus perception of users or their satisfaction. We have identified several limitations in the current literature on group recommendation explanations that we address in this paper. First, social choice-based aggregation strategies and their explanations are not evaluated in isolation. Hence, it is unclear to what extent users' fairness perception, consensus perception, and satisfaction evaluations depend on **a)** the *explanations* or **b)** simply the social choice *aggregation strategies*. Furthermore, while we agree that the field of group recommendation explanations is a young one, there is no precedent of replication studies, let alone benchmarks and baselines to compare explanations against. The challenges the replication crises have posed in the social sciences and medicine suggest that similar difficulties would be present in other fields involving user studies [18, 19].

In this paper, thus, we address the aforementioned limitations by taking the first steps toward an explanation benchmark for group explanations. We conduct a preregistered between-subjects user study with 400 participants, where each participant evaluates one aggregation strategy and one explanation type in terms of perceived fairness, perceived consensus, and satisfaction regarding the group recommendations [1]. In addition, we also test for interaction effects between aggregation strategies and explanation types. Thus, we address the following research questions:

RQ1: Are there differences between *social choice-based aggregation strategies* in group recommendation settings regarding users' fairness perception, consensus perception, or satisfaction?

RQ2: Do explanations that are based on the group recommendation aggregation strategy at hand increase users' fairness perception, consensus perception, or satisfaction?

RQ3: Does the effectiveness of explanations (w.r.t. users' fairness perception, consensus perception, or satisfaction) vary depending on the aggregation strategies at hand?

RQ4: Are users' levels of perceived fairness or perceived consensus related to their satisfaction concerning the group recommendations?

[1] To preregister our study, we publicly determined our research questions, hypotheses, experimental setup, and data analysis plan before any data collection. The (time-stamped) preregistration can be found at https://osf.io/ghbsq.

2. Related Work and Hypotheses

In this section, we introduce the social choice-based aggregation strategies used to generate recommendations for groups. Then, we illustrate the relevant literature on the explanations for group recommender systems. Motivated by relevant literature, we also present the hypotheses that we verify in our study.

2.1. Social Choice-based Aggregation Strategies

There are two main approaches to generate group recommendations: *(i)* aggregated models: aggregate individual preferences (*e.g.*, existing ratings) into a group model, generating then the group recommendations based on such a group model; and *(ii)* aggregated predictions or strategies: aggregate individual item-ratings predictions and recommend items with the highest aggregated scores to the group [15]. Several aggregation strategies inspired by Social Choice Theory have been proposed to aggregate individuals' information [8]. An overview of these strategies, known as social choice-based aggregation strategies, can be found in Masthoff [4]. Following, we describe six of the most utilized social choice-based aggregation strategies: (i) *Additive Utilitarian (ADD)* is a consensus-based strategy [20], so it takes into account the preferences of all group members, recommending the item with the highest sum of all group members ratings; (ii) *Fairness (FAI)* is a consensus-based strategy [8] well suited in the context of repeated decisions, in which the items are ranked as the individuals are choosing them in turn; (iii) *Approval Voting (APP)* is a majority-based strategy [20], so it focuses on the most popular items among group members, recommending the item which has the highest number of ratings greater than a predefined threshold; (iv) *Least Misery (LMS)* is a borderline strategy [20], so it takes into account only a subset of group members preferences and recommends the item which has the highest of all lowest ratings; (v) *Majority (MAJ)* is a borderline strategy [20] which recommends the item with the highest number of all ratings representing the majority of item-specific ratings; (vi) *Most Pleasure (MPL)* is a borderline strategy [20] which recommends the item with the highest of all individual group members ratings. Social choice-based aggregation strategies are widely used in the group recommenders literature [8]. In Masthoff [8], several experiments are presented to identify the best strategy in terms of perceived group satisfaction. The results, however, show that there is no winning strategy — different strategies perform well in different scenarios. This consideration leads us to the following hypotheses related to **RQ1**[2]:

- **H1a**: There is a difference between social choice-based aggregation strategies in group recommendation settings regarding users' fairness perception.
- **H1b**: There is a difference between social choice-based aggregation strategies in group recommendation settings regarding users' consensus perception.
- **H1c**: There is a difference between social choice-based aggregation strategies in group recommendation settings regarding user satisfaction.

[2]We note here that we slightly changed the preregistered hypotheses according to the change made to the research question. The intention is to compare all five aggregation strategies and not only the ones that are categorized as consensus-based.

2.2. Explaining to Groups

In general, explanations can be seen as additional information that is associated with the recommendations to achieve several goals, such as increasing the transparency (explaining how the recommendation system works), effectiveness (helping the user in making good decisions), and usability of the system, as well as user satisfaction [21]. Several studies in different domains showed the benefits of using explanations for recommendations in increasing users acceptance rate and satisfaction [22], or trust in the system [23]. In group recommendations, explanations can achieve further goals: fairness (consider all group members' preference as much as possible); consensus (help group members to agree on the decision) [15]; privacy-preserving (preserve group members' confidential data, to avoid concerns about a possible loss of privacy by, *e.g.*, disclosing the preference information of individual group members in the explanation) [24, 25, 26]. However, most of the research on explanations for recommender systems focus on single-user scenarios, while only a few studies investigate the problem of generating explanations for groups. Typically, such explanations are related to the underlying mechanism of the employed social choice-based aggregation strategy [5, 27, 17]. Natural language explanation styles based on the underlying social choice aggregation strategies were introduced in Najafian and Tintarev [5], while Kapcak et al. [27] extended this work using the wisdom of the crowd to improve the quality of the initially proposed explanations. Quijano-Sanchez et al. [28] introduced explanations including the social factors of personality and tie strength between group members to generate tactful explanations (*e.g.*, explanations that avoid damaging friendships). In a more extensive study, Tran et al. [17] propose three types of explanations for six social choice-based aggregation strategies (ADD, FAI, APP, LMS, MAJ, and MPL), by considering: *(1)* the aggregation strategy itself - *Type 1*, *(2)* the aggregation strategy itself and the decision history - *Type 2*, and *(3)* the aggregation strategy itself and the future decision plan - *Type3*. In a user study, they evaluated these explanations and showed that explanations related to the ADD and MAJ strategies help the most to increase the fairness and consensus perception, and satisfaction regarding the group recommendation. They also found that users' perceived fairness or consensus correlates with their satisfaction.

Although these works present valuable ways to generate explanations for the most used benchmark aggregation strategies in group recommender systems research, it is unclear whether the effects attributed to the explanations might not, in fact, depend on the aggregation strategies themselves. Starting from this consideration, we formulated a second set of hypotheses that we intend to validate, related to **RQ2**:

- **H2a**: Explanations based on the aggregation strategy at hand increase users' fairness perception concerning group recommendations.
- **H2b**: Explanations based on the aggregation strategy at hand increase users' consensus perception concerning group recommendations.
- **H2c**: Explanations based on the aggregation strategy at hand increase users' satisfaction concerning group recommendations.

Furthermore, an aspect that has not been investigated is the level of detail that the explanation can achieve concerning the aggregation strategy used and whether this affects the users' fairness

perception, consensus perception, and satisfaction. To this end, we introduce a third set of hypotheses which we intend to test, related to **RQ3**:

- **H3a**: The effect of aggregation strategy-based explanations on users' fairness perception concerning group recommendations is moderated by the type of aggregation strategy at hand.
- **H3b**: The effect of aggregation strategy-based explanations on users' consensus perception concerning group recommendations is moderated by the type of aggregation strategy at hand.
- **H3c**: The effect of aggregation strategy-based explanations on user satisfaction concerning group recommendations is moderated by the type of aggregation strategy at hand.

Finally, we also validate the correlation between user satisfaction and perceived fairness and consensus, *c.f.*, [17]:

- **H4a**: Users' perceived fairness is positively related to user satisfaction concerning group recommendations.
- **H4b**: Users' perceived consensus is positively related to user satisfaction concerning group recommendations.

3. Method

We conducted an online between-subjects user study to test our hypotheses.[3] We presented users with scenarios that reflected one of five different social choice-based aggregation strategies for group recommender systems and that included either no explanation or one of two different explanation types. This section outlines the materials, variables, procedure, participant sample, and statistical analyses related to our user study.

3.1. Materials

Aggregation Strategies

Our study considered five different social choice-based aggregation strategies for group recommender systems, that have been evaluated in prior work [17]. Each of these strategies aggregates the preferences of several users to obtain a recommendation for the group as a whole [20]. Differently than in [17], we do not consider the *Fairness* aggregation strategy because the explanation types that we propose can not be generated for this strategy. Each aggregation strategy is applied to the rating scenario presented in Table 1, where each item (*i.e.*, the three restaurants, Rest A, Rest B, and Rest C) is rated on a 5-star rating scale (*i.e.*, 1 - the worst and

[3] All material for analyzing our results and replicating our user study, (*i.e.*, document with preregistration of all the hypotheses tested, user study materials, data gathered in the user study and the analysis scripts) is publicly available – https://osf.io/5xbgf/.

5 - the best). Specifically, we consider the following aggregation strategies, from Section 2.1: *Additive Utilitarian* (ADD); *Approval Voting* (APP) considering a threshold equal to 3, as in [17]; *Least Misery* (LMS); *Majority* (MAJ); *Most Pleasure* (MPL).

Explanations

Each explanation is presented after showing the scenario in Table 1 and a recommendation generated with one of the aggregation strategies considered (see Section 3.2 for more details). We evaluate three types of explanations (see Table 4): (i) *Basic explanations*, which explain the aggregation strategy at hand. These explanations have been adopted from Tran et al. [17], and refer to *Type 1*; (ii) *Detailed explanations*, that explain the aggregation strategy in greater detail by describing the specific reason why a given item has been recommended; additionally, we included a condition *no explanation*, where the aggregation strategy is applied, but no explanation is given. Participants did, however, see the ratings of the other group members in this condition.

3.2. Procedure

Our study consisted of two subsequent steps. During the first step (after participants had agreed to informed consent), we introduced participants to the study and asked them for their gender and age. The second step of our study started with the following scenario (taken from Tran et al. [17]):

"Assume, there is a group of four friends (Alex, Anna, Sam, and Leo). Every month, a group decision is made by these friends to decide on a restaurant to have dinner together. To select a restaurant for the dinner next month, the group again has to take the same decision. In this decision, each group member explicitly rated three restaurants (Rest A, Rest B, and Rest C) using a 5-star rating scale (1: the worst, 5: the best). The ratings given by group members are shown in the table below:"

After that, Table 1 was shown. Participants saw a group recommendation either with or without an explanation depending on which aggregation strategy and explanation type they had been assigned to (see Table 4). We then measured perceived fairness, perceived consensus, and satisfaction (see Section 3.3). We also included an attention check where we specifically instructed participants on what option to select. Finally, participants could explain their answers in a text field. The study had been approved by the ethics committee of our institution.

Table 1: Ratings of group members for the restaurants (1: the worst, 5: the best) from Tran et al. [17].

	Alex	Anna	Sam	Leo
Rest A	2	2	4	4
Rest B	1	4	4	4
Rest C	5	1	1	1

3.3. Variables

Independent Variables

(i) **Aggregation strategy** (categorical, between-subjects). Each participant was exposed to a scenario that reflected one of the five aggregation strategies (i.e., ADD, APP, LMS, MAJ, or MPL; see Section 3.1). (ii) **Explanation type** (categorical, between-subjects). Each participant saw either *no explanation*, a *basic explanation*, or a *detailed explanation* (see Section 3.1).

Dependent Variables

We measured each of our three dependent variables by asking participants to respond to a statement on a seven-point Likert scale ranging from "strongly agree" (scored as -3) to "strongly disagree" (scored as 3). We have: (i) **Perceived fairness** (ordinal): "The group recommendation is fair to all group members"; (ii) **Perceived Consensus** (ordinal): "The group members will agree on the group recommendation"; (iii) **Satisfaction** (ordinal): "The group members will be satisfied with regard to the group recommendation".

Descriptive Variables

We collected data on two demographic variables: (i) **Age** (categorical), participants could select one of the options *18-25*, *26-35*, *36-45*, *46-55*, *>55*; (ii) **Gender** (categorical). Participants could select one of the options *female*, *male*, or *other*. Participants could also select a "prefer not to say" option for these variables.

3.4. Participants

Before data collection, we computed the required sample size for our study in a power analysis for a between-subjects ANOVA (Fixed effects, special, main effects, and interactions; see Section 3.5) using *G*Power* [29]. Here, we specified the default effect size f = 0.25, a significance threshold $\alpha = \frac{0.05}{11} = 0.005$ (due to testing multiple hypotheses; see Section 3.5), a power of $(1 - \beta) = 0.8$, and that we will test $5 \times 3 = 15$ groups (i.e., 5 different aggregation strategies for 3 different explanation scenarios). We performed this computation for each hypothesis using their respective degrees of freedom. This resulted in a total required sample size of at least 378 participants. We thus recruited 400 participants from the online participant pool *Prolific*[4], all of whom were proficient English speakers above 18 years of age. To maintain high-quality answers, we selected only participants that had an approval rate of at least 90% and participated in at least 10 prior studies. Each participant was allowed to participate in our study only once and received £0.63 as a reward for participation. We excluded one participant from data analysis because they did not pass the attention check we included in the experiment. The resulting sample of 399 participants was composed of 61% female, 38% male, and 1% other participants. They represented a diverse range of age groups: 28% were between 18 and 25, 29% between 26 and 35, 17% between 36 and 45, 14% between 46 and 55, and 13% were above 55 years of age. We

[4] https://prolific.co

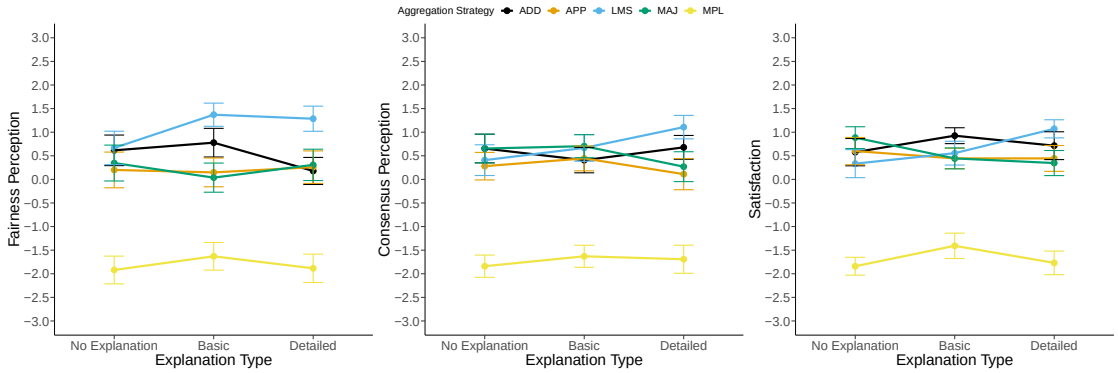

Figure 1: Participants' mean *fairness perception*, *consensus perception*, and *satisfaction* across explanation types on scales from -3 ("strongly disagree") to 3 ("strongly agree"; see Section 3.3). Colors indicate aggregation strategies: Additive Utilitarian (ADD), Approval Voting (APP), Least Misery (LMS), Majority (MAJ), Most Pleasure (MPL). Error bars represent the standard error of the mean.

randomly distributed participants over the 15 conditions (i.e., exposing them to 1/5 aggregation strategies and 1/3 explanation types).

3.5. Statistical Analyses

For each of the three dependent variables in our study (i.e., *fairness perception*, *consensus perception*, and *satisfaction*), we conducted a two-way analysis of variance (ANOVA) using *aggregation strategy* and *explanation type* as between-subjects factors. These three ANOVAs were used to test nine hypotheses (i.e., **H1a** – **H3c**). Specifically, each of them tested main effects of *aggregation strategy* (**H1a** – **H1c**) and *explanation type* (**H2a** – **H2c**) as well as the interaction between these two variables in affecting the dependent variables (**H3a** – **H3c**). We chose this type of analysis despite the anticipation that our data may not be normally distributed (i.e., violating an ANOVA assumption) because ANOVAs are usually robust to Likert-type ordinal data [30]. We additionally performed two Spearman correlation analyses to test hypotheses **H4a** and **H4b**. We thus tested 11 different hypotheses. Applying a Bonferroni correction [31], we lowered the significance threshold to $\alpha = \frac{0.05}{11} = 0.0046$. Since we found significant main effects related to our first six hypotheses (**H1a** – **H2c**; see Section 4), we conducted Tukey posthoc analyses to investigate specific differences between the aggregation strategies and explanation types. The p-values from these posthoc analyses were adjusted to correct for multiple testing (i.e., written as p_{adj}).

4. Results

Descriptive Statistics

Participants' distribution over the 15 different conditions (i.e., all possible combinations between the five aggregation strategies and the three explanation types) was balanced: each condition

Table 2
Results of three two-way ANOVAs for the dependent variables (DVs) *fairness perception* (left), *consensus perception* (center), and *satisfaction* (right). Per effect, we report the F-statistic, p-value, and $\eta^2_{partial}$ effect size. The terms "aggr" and "expl" represent the independent variables *aggregation strategy* and *explanation type*. Colons indicate interaction effects, asterisks statistical significance.

DV: Fairness Perception				DV: Consensus Perception				DV: Satisfaction			
	F	p	η^2_p		F	p	η^2_p		F	p	η^2_p
(H1a) aggr	36.19	<0.001*	0.27	**(H1b)** aggr	38.89	<0.001*	0.29	**(H1c)** aggr	49.57	<0.001*	0.34
(H2a) expl	0.35	0.71	0.00	**(H2b)** expl	0.14	0.87	0.00	**(H2c)** expl	0.15	0.86	0.00
(H3a) aggr:expl	0.68	0.71	0.01	**(H3b)** aggr:expl	0.75	0.65	0.02	**(H3c)** aggr:expl	1.25	0.27	0.03

was shown to 6% to 7% of participants. On average, participants spent 2.9 (sd = 2.2; no notable difference between conditions) minutes on the task. Qualitative feedback from participants suggested that the scenario and task were understandable. Participants had a slight overall tendency to perceive fairness, consensus, and satisfaction in the scenarios, as 51%, 51%, and 56% at least somewhat agreed to these three items, respectively. Figure 1 shows participants' mean *fairness perception*, *consensus perception*, and *satisfaction* across explanation types and split by aggregation strategies.

RQ1: differences between social-choice based aggregation strategies regarding explanation effectiveness. We found significant differences between the five aggregation strategies concerning all three dependent variables *fairness perception*, *consensus perception*, and *satisfaction* (**H1a** – **H1c**; $F = [36.19, 49.57]$, all $p < 0.001$, $\eta^2_p = [0.27, 0.34]$; see Table 2). So, overall, participants expressed different levels regarding these three variables based on which aggregation strategy they were exposed to. Posthoc analyses revealed that MPL led to lower levels on all three variables compared to all other aggregation strategies (all $p_{\text{adj}} < 0.001$). The only other significant differences we found between aggregation strategies was that APP ($p_{\text{adj}} = 0.004$) and MAJ ($p_{\text{adj}} = 0.005$) each led to lower fairness perception compared to LMS. In sum, participants – irrespective of which explanation type they saw – viewed MPL as significantly less fair, consensual, and satisfying compared to other strategies, and judged MAJ as well as APP as less fair compared to LMS.

RQ2: differences between explanation types (i.e., *no explanation*, *basic explanation*, or *detailed explanation*). We found no significant differences between the three explanation types regarding all three dependent variables (**H2a** – **H2c**; $F = [0.14, 0.35]$, $p = [0.71, 0.86]$, all $\eta^2_p = 0.00$; see Table 2). So, our results contain no evidence for a difference between explanation types concerning our three dependent variables.

RQ3: interactions between aggregation strategies and explanation types regarding explanation effectiveness. There were no significant interaction effects between the five aggregation strategies and the three explanation types (**H3a** – **H3c**; $F = [0.65, 1.25]$, $p = [0.27, 0.71]$, $\eta^2_p = [0.01, 0.03]$; see Table 2). The effect of explanation types on participants' *fairness perception*, *consensus perception*, and *satisfaction* thus did not significantly differ based on which aggregation strategy was applied.

RQ4: associations between explanation effectiveness measures. In line with the findings of Tran et al. [17], Spearman correlation analyses revealed significant positive relationships between fairness perception and satisfaction ($\rho = 0.71, p < 0.001$) as well as between consensus perception and satisfaction ($\rho = 0.76, p < 0.001$). This means that, as participants' fairness and consensus perception increased, satisfaction also increased.

5. Discussion

In the following sections, we look closer at our results and their implications. We discuss the difference between aggregation strategies, the difference between different explanation levels, and the effect of the chosen scenario. We conclude with lessons learned for future benchmarking studies in explanations research and limitations of our study.

5.1. The Differences Between Aggregation Strategies

As shown in Section 4, there are differences between the aggregation strategies in terms of perceived fairness, perceived consensus, and satisfaction. The MLP strategy obtains the lowest scores, regardless of the type of explanation received. Furthermore, MAJ and APP are perceived to be less fair than LMS. We discuss how these results may have interacted with the presented scenario in Section 5.5. However, these results are in contrast with the findings of Tran et al. [17], where the same scenario was used. In such work, the MAJ and ADD strategies scored better than the LMS strategy. An explanation of this difference can be the different design of our experiment: we implemented a between-subject design to guarantee the independence between the conditions; on the contrary, in [17] each user evaluated six strategies and was exposed to different explanation types. Although the strategies were presented in a randomized order to reduce biases, it is possible that the user used an explanation type seen first as a reference point to compare with, in the following evaluations, which introduced noise in the users' evaluations. Furthermore, to evaluate the effect of the aggregation strategy separately from the explanation, we asked participants to evaluate the recommendation. In contrast, Tran et al. [17] asked the participants to evaluate the explanation provided, hence the evaluation of the explanation was influenced by the evaluation of the aggregation strategy.

5.2. The Role of Explanations

The results presented showed no significant difference between the different types of explanations. Furthermore, no interaction effects between the explanations and the aggregations regarding the measured dependent variables (perceived fairness, perceived consensus, and satisfaction) were found. However, these results are not enough to claim that the explanations are not useful for group recommender systems. First, it must be considered that the used scenario was particularly simple to evaluate. More complex scenarios might involve a more balanced situation between subgroups with different preferences, or a greater number of options to choose from: such factors might complicate the assessment; in such cases, an explanation of the approach used might have an impact. Moreover, the strategies presented here represent baselines for group recommenders. Therefore, it is necessary to formalise the explanations for

these strategies, as they serve as a reference against which more articulated strategies can be compared. The most recent lines of research in group recommenders, however, try to integrate into the recommendation generation process personal factors (experience in the domain [32] or personality [33, 28, 34]), as well as social factors (tie strength [35], centrality of group members in the group social network [36], group diversity [37]). In such cases, an explanation may have an impact on the transparency and comprehensibility of the system and result in different evaluations regarding fairness perception, consensus perception, and satisfaction. This, of course, also leads to privacy issues, concerning which personal information of one or more individuals can be mentioned in an explanation.

5.3. The Link Between Fairness, Consensus, and Satisfaction

The correlation between fairness perception (or consensus perception) and satisfaction, already reported in Tran et al. [17], and also showed in our results, confirms the close connection between these concepts. A solution perceived as less fair is also perceived as less satisfactory, and a less satisfactory solution is unlikely to be accepted by the group. This confirms that these aspects, sometimes considered secondary, are crucial and that a group recommendation system must take them into account, both in the generation of recommendations and in their evaluation.

5.4. Lessons Learned for Benchmarking

Report on Participant Recruitment

Numerous platforms can be used to outsource user studies [38], such as Prolific and Amazon Mechanical Turk. Recruitment might also focus on particular users, such as students or staff members. Filtering conditions, such as those for quality control also affect which demographics take part in a study. More generally, any selection of study participants can influence the outcome of the evaluation, which should not be generalized outside the scope of the scenario [39]. Therefore, we recommend a thorough reporting on how participants were recruited.

Report Study Design and Statistical Analysis Rigorously

The choice of the quantitative study, between-subjects, within-subjects or mixed designs is also influencing the conclusions that can be drawn, as well as the statistical analysis that should be applied. In any case, randomizing participants to conditions is of paramount importance, regardless of study design. More personalized study designs, such as the one conducted by Tran et al. [17], should clearly specify how each scenario has been allocated to participants, to be able to replicate them. We, in particular, recommend more rigorous reporting of how randomization is performed, as well as sharing scripts to support replication and comparison.

Ensure consistency in measurement or motivate changes well

In separating the evaluation of explanations and aggregation strategy, we found it was no longer feasible to ask participants to evaluate the explanations rather than the resulting recommendation. In addition, compared to Tran et al. [17], we ask study participants to rate explanations'

effectiveness on a 7-point Likert scale, instead of a 5-point Likert scale, since this ensures greater robustness in the use of ANOVA analysis, according to [30]. While these changes may not have affected the results, such changes in the design must be described and motivated when attempting to benchmark such user studies.

Report on Completeness

We found that certain aggregation strategies can not be explained in certain instances or scenarios. In this paper, this was the case of the *Fairness* strategy, which is well-suited for repeated decisions, but less applicable for single decisions as in our case. We recommend that future work not only describes the cases where explanations can be generated but also describes the edge cases for which they cannot.

Consider the Effect of the Scenario

The proposed scenario in this work was selected to specifically study groups with heterogeneous preferences. However, this choice is likely to have affected our specific results. For example, the MPL strategy in this specific scenario recommends a solution that displeases at most three out of four group members (Rest C, see Table 1). It is not surprising, therefore, that it is identified as the least fair, least satisfactory strategy, and the one on which it is most difficult to reach an agreement. The result might have been different if it displeased fewer group members. We, therefore, recommend not only to clearly report on the scenario used, but also to discuss its implications.

Consider effects of the role of the participant in a group

The evaluations are given in this paper based on an *external* evaluator who may be more unbiased (than someone within the group). Users within the group may be influenced by their own preferences. Furthermore, the assessment of the fairness of a scenario will likely differ depending on whether it favors the user, *e.g.*, if MLP displeases 2 users and whether the active user is one of them.

5.5. Limitations

Recommendations and explanations are not evaluated by group members

As previously mentioned, in line with the evaluation approach in Tran et al. [17], our study participants were asked to evaluate the recommendations as external evaluators. This means that study participants were not members of the group. We hypothesize, however, that their evaluations could be different when part of the group. Deciding for an evaluator that is part of the group would entail controlling more cases, such as when the evaluator is in the majority preference, minority preference, or a tie preference.

We do not measure nor capture the reasoning process of the study participants regarding recommendations

In the condition with *no explanations*, we provide a mere description of the recommendation. However, we do not capture how study participants reflect on the recommendation or to what extent they understand it. Prior literature [40, 14, 41], however, provides several directions for measuring recommendation understandability, which could be investigated in future work. Nevertheless, our descriptive analysis in Section 4 shows that participants spent a similar amount of time to complete each explanation condition. This suggests that they spent a similar amount of effort analyzing their fairness and consensus perception, as well as satisfaction regarding the recommended restaurant.

Recommendations are provided for unnamed restaurants

We did not want to influence participants' decisions by providing real restaurant names as recommendations. This helped us control for the potential bias that could have been added while showing a real restaurant name. Such normalization, however, could potentially influence the assessments of the study participants, compared to a customized recommendation. Another limitation of our study is that all recommendations are in the restaurants' domain. Different recommendation domains could be differently perceived in terms of fairness, consensus, and satisfaction. In particular, the investment related to the domain considered has shown to have an impact on the evaluation of the recommendations [42]; the restaurant domain is generally perceived as a medium-low investment, compared to other domains suitable for group recommendations, such as tourism.

6. Conclusions

We present a preregistered evaluation of the impact of using social choice-based explanations for group recommendations. Overall, our finding suggests that explanations are not necessarily helpful for improving perceptions of the recommendations. Participants' evaluations were not influenced by the presence of an explanation, and their perceptions of fairness, consensus, and satisfaction were primarily formed based on the social choice-based aggregation strategies. Participants evaluated the Least Misery (LMS) strategy as more fair than the Approval Voting (APP) and the Majority (MAJ), while the Most Pleasure (MPL) was considered the worst in terms of perceived fairness, perceived consensus, and satisfaction. We also discuss some of the challenges and decision points required to benchmark future studies of group explanations. In particular, we highlighted the importance of clarifying and motivating the recruitment process and properly choosing the experimental design, specifying how each condition is assigned to each participant. Furthermore, we discussed how the choice of the scenario to present for the evaluation can influence the results, and that, therefore, the results should always be discussed in relation to it. In future work, we plan to investigate the influence of internal vs. external evaluators. We plan to thoroughly study the reasoning process of evaluators and measure the level of understanding regarding the recommended item. To observe to what extent our results generalize, we also plan to replicate our study with other scenarios and domains.

References

[1] Y.-L. Chen, L.-C. Cheng, C.-N. Chuang, A group recommendation system with consideration of interactions among group members, Expert systems with applications 34 (2008) 2082–2090.

[2] J. K. Kim, H. K. Kim, H. Y. Oh, Y. U. Ryu, A group recommendation system for online communities, International journal of information management 30 (2010) 212–219.

[3] M. O'connor, D. Cosley, J. A. Konstan, J. Riedl, Polylens: A recommender system for groups of users, in: ECSCW 2001, Springer, 2001, pp. 199–218.

[4] J. Masthoff, Group modeling: Selecting a sequence of television items to suit a group of viewers, in: Personalized digital television, Springer, 2004, pp. 93–141.

[5] S. Najafian, N. Tintarev, Generating consensus explanations for group recommendations: an exploratory study, in: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, ACM, 2018, pp. 245–250.

[6] D. Cao, X. He, L. Miao, Y. An, C. Yang, R. Hong, Attentive group recommendation, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 645–654.

[7] S. Najafian, D. Herzog, S. Qiu, O. Inel, N. Tintarev, You do not decide for me! evaluating explainable group aggregation strategies for tourism, in: Proceedings of the 31st ACM Conference on Hypertext and Social Media, 2020, pp. 187–196.

[8] J. Masthoff, Group recommender systems: aggregation, satisfaction and group attributes, in: recommender systems handbook, Springer, 2015, pp. 743–776.

[9] K. J. Arrow, A difficulty in the concept of social welfare, Journal of political economy 58 (1950) 328–346.

[10] A. Felfernig, L. Boratto, M. Stettinger, M. Tkalčič, Explanations for groups, in: Group Recommender Systems, Springer, 2018, pp. 105–126.

[11] N. Tintarev, J. Masthoff, Effective explanations of recommendations: user-centered design, in: Proceedings of the 2007 ACM conference on Recommender systems, 2007, pp. 153–156.

[12] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, Recommender systems: an introduction, Cambridge University Press, 2010.

[13] L. Chen, M. De Gemmis, A. Felfernig, P. Lops, F. Ricci, G. Semeraro, Human decision making and recommender systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 3 (2013) 1–7.

[14] F. Gedikli, D. Jannach, M. Ge, How should i explain? a comparison of different explanation types for recommender systems, International Journal of Human-Computer Studies 72 (2014) 367–382.

[15] A. Felfernig, L. Boratto, M. Stettinger, M. Tkalčič, Explanations for groups, in: Group Recommender Systems, Springer, 2018, pp. 105–126.

[16] E. Ntoutsi, K. Stefanidis, K. Nørvåg, H.-P. Kriegel, Fast group recommendations by applying user clustering, in: International conference on conceptual modeling, Springer, 2012, pp. 126–140.

[17] T. N. T. Tran, M. Atas, A. Felfernig, V. M. Le, R. Samer, M. Stettinger, Towards social choice-based explanations in group recommender systems, in: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, 2019, pp. 13–21.

[18] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, et al., Promoting an open research culture, Science 348 (2015) 1422–1425.
[19] B. Nosek, J. Cohoon, M. Kidwell, J. R. Spies, Estimating the Reproducibility of Psychological Science, Science 349 (2015) aac47160. doi:10.1126/science.aac4716.
[20] C. Senot, D. Kostadinov, M. Bouzid, J. Picault, A. Aghasaryan, C. Bernier, Analysis of strategies for building group profiles, in: International Conference on User Modeling, Adaptation, and Personalization, Springer, 2010, pp. 40–51.
[21] N. Tintarev, J. Masthoff, A survey of explanations in recommender systems, in: 2007 IEEE 23rd international conference on data engineering workshop, IEEE, 2007, pp. 801–810.
[22] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM conference on Computer supported cooperative work, ACM, 2000, pp. 241–250.
[23] R. Sinha, K. Swearingen, The role of transparency in recommender systems, in: CHI'02 extended abstracts on Human factors in computing systems, 2002, pp. 830–831.
[24] S. Najafian, A. Delic, M. Tkalcic, N. Tintarev, Factors influencing privacy concern for explanations of group recommendation, in: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, 2021, pp. 14–23.
[25] S. Najafian, O. Inel, N. Tintarev, Someone really wanted that song but it was not me! evaluating which information to disclose in explanations for group recommendations, in: Proceedings of the 25th International Conference on Intelligent User Interfaces Companion, 2020, pp. 85–86.
[26] S. Najafian, T. Draws, F. Barile, M. Tkalcic, J. Yang, N. Tintarev, Exploring user concerns about disclosing location and emotion information in group recommendations, in: Proceedings of the 32st ACM Conference on Hypertext and Social Media, 2021, pp. 155–164.
[27] Ö. Kapcak, S. Spagnoli, V. Robbemond, S. Vadali, S. Najafian, N. Tintarev, Tourexplain: A crowdsourcing pipeline for generating explanations for groups of tourists, in: Workshop on Recommenders in Tourismco-located with the 12th ACM Conference on Recommender Systems (RecSys 2018), CEUR, 2018, pp. 33–36.
[28] L. Quijano-Sanchez, C. Sauer, J. A. Recio-Garcia, B. Diaz-Agudo, Make it personal: a social explanation system applied to group recommendations, Expert Systems with Applications 76 (2017) 36–48.
[29] F. Faul, E. Erdfelder, A. G. Lang, A. Buchner, G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, Behavior Research Methods 39 (2007) 175—-191. doi:10.3758/BF03193146.
[30] G. Norman, Likert scales, levels of measurement and the "laws" of statistics, Advances in Health Sciences Education 15 (2010) 625–632. doi:10.1007/s10459-010-9222-y.
[31] M. A. Napierala, What Is the Bonferroni correction?, 2012. URL: http://www.aaos.org/news/aaosnow/apr12/research7.asp.
[32] M. Gartrell, X. Xing, Q. Lv, A. Beach, R. Han, S. Mishra, K. Seada, Enhancing group recommendation by incorporating social relationship interactions, in: Proceedings of the 16th ACM international conference on Supporting group work, 2010, pp. 97–106.
[33] T. N. Nguyen, F. Ricci, A. Delic, D. Bridge, Conflict resolution in group decision making: insights from a simulation study, User Modeling and User-Adapted Interaction 29 (2019)

895–941.

[34] S. Rossi, F. Cervone, F. Barile, An altruistic-based utility function for group recommendation, in: Transactions on Computational Collective Intelligence XXVIII, Springer, 2018, pp. 25–47.

[35] F. Barile, J. Masthoff, S. Rossi, The adaptation of an individual's satisfaction to group context: the role of ties strength and conflicts, in: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, 2017, pp. 357–358.

[36] A. Delic, J. Masthoff, J. Neidhardt, H. Werthner, How to use social relationships in group recommenders: empirical evidence, in: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, 2018, pp. 121–129.

[37] A. Delic, J. Masthoff, H. Werthner, The effects of group diversity in group decision-making process in the travel and tourism domain, in: Information and Communication Technologies in Tourism 2020, Springer, 2020, pp. 117–129.

[38] E. Peer, L. Brandimarte, S. Samat, A. Acquisti, Beyond the turk: Alternative platforms for crowdsourcing behavioral research, Journal of Experimental Social Psychology 70 (2017) 153–163.

[39] J. Beel, C. Breitinger, S. Langer, A. Lommatzsch, B. Gipp, Towards reproducibility in recommender-systems research, User modeling and user-adapted interaction 26 (2016) 69–101.

[40] B. P. Knijnenburg, N. J. Reijmer, M. C. Willemsen, Each to his own: how different users call for different interaction methods in recommender systems, in: Proceedings of the fifth ACM conference on Recommender systems, 2011, pp. 141–148.

[41] X. Wang, M. Yin, Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making, in: 26th International Conference on Intelligent User Interfaces, 2021, pp. 318–328.

[42] N. Tintarev, J. Masthoff, Over-and underestimation in different product domains, in: Workshop on Recommender Systems associated with ECAI, Springer Boston, MA, 2008, pp. 14–19.

A. Appendix - Basic and Detailed Explanations

In this appendix, we specify how to generate the Basic and Detailed explanations used in this work, for each of the aggregation strategies considered (see section 3.1). Let $G = \{u_1, ..., u_n\}$ be a group of users, and $I = \{i_1, ..., i_m\}$ be a set of items. Also, let $\{u_{j_1}, u_{j_2}, ..., u_{j_{\bar{n}}}\}$ be a subset of group members who gave a specific rating r to the item i_k selected by the considered strategy. Hence, we can define the explanations, for each aggregation strategy, as in the Table 3, while the Table 4 shows the explanations obtained on the scenario used in the experiment (see the Table 1).

Table 3
Generic formulations for each aggregation strategy of the explanations used in this study.

Strat.	No explanation	Basic explanation	Detailed explanation
ADD	*"i_k has been recommended to the group."*	*"i_k has been recommended to the group since it achieves the highest total rating."*	*"i_k has been recommended to the group since it achieves the highest total rating (as the sum of the ratings of all members for i_k is r which is higher than other items)."*
APP	*"i_k has been recommended to the group."*	*"i_k has been recommended to the group since it achieves the highest number of ratings which are above a threshold (θ)."*	*"i_k has been recommended to the group since it achieves the highest number of ratings which are above a threshold (as the $\bar{n}$ group members u_{j_1}, u_{j_2}, ... and $u_{j_{\bar{n}}}$ gave it ratings higher than θ)."*
LMS	*"i_k has been recommended to the group."*	*"i_k has been recommended to the group since no group members has a real problem with it."*	*"i_k has been recommended to the group since no group members has a real problem with it (as u_{j_1}, u_{j_2}, ... and $u_{j_{\bar{n}}}$ gave it a rating of r which is the highest rating among the lowest ratings per item)."*
MAJ	*"i_k has been recommended to the group."*	*"i_k has been recommended to the group since most group members like it."*	*"i_k has been recommended to the group since most group members like it (as $\bar{n}$ out of n group members gave it a high rating)."*
MPL	*"i_k has been recommended to the group."*	*"i_k has been recommended to the group since it achieves the highest of all individual group members."*	*"i_k has been recommended to the group since it achieves the highest of all individual group members' ratings (as u_{j_1}, u_{j_2}, ... and $u_{j_{\bar{n}}}$ gave it the rating r, which is the highest rating among all items' high ratings)."*

Table 4
All possible explanation scenarios that participants saw in our study. The explanations describe a restaurant recommendation scenario that participants were exposed to (based on the scenario defined in Table 1).

Strat.	No explanation	Basic explanation	Detailed explanation
ADD	*"Restaurant B has been recommended to the group."*	*"Restaurant B has been recommended to the group since it achieves the highest total rating."*	*"Restaurant B has been recommended to the group since it achieves the highest total rating (as the sum of the ratings of all members for Restaurant B is 13 which is higher than other items).*
APP	*"Restaurant B has been recommended to the group."*	*"Restaurant B has been recommended to the group since it achieves the highest number of ratings which are above 3."*	*"Restaurant B has been recommended to the group since it achieves the highest number of ratings which are above a threshold (as the three group members Anna, Sam, and Leo gave it ratings higher than 3)."*
LMS	*"Restaurant A has been recommended to the group."*	*"Restaurant A has been recommended to the group since no group members has a real problem with it."*	*"Restaurant A has been recommended to the group since no group members has a real problem with it (as Alex and Anna gave it a rating of 2 which is the highest rating among the lowest ratings per restaurant)."*
MAJ	*"Restaurant B has been recommended to the group."*	*"Restaurant B has been recommended to the group since most group members like it."*	*"Restaurant B has been recommended to the group since most group members like it (as 3 out of 4 group members gave it a high rating)."*
MPL	*"Restaurant C has been recommended to the group."*	*"Restaurant C has been recommended to the group since it achieves the highest of all individual group members' ratings."*	*"Restaurant C has been recommended to the group since it achieves the highest of all individual group members' ratings (as Alex gave it the rating 5, which is the highest ratings among all items' high ratings)."*