Evaluating explainable social choice-based aggregation strategies for group recommendation

Barile, Francesco; Draws, Tim; Inel, Oana; Rieger, Alisa; Najafian, Shabnam; Ebrahimi Fard, Amir; Hada, Rishav; Tintarev, Nava

DOI
10.1007/s11257-023-09363-0

Publication date
2023

Document Version
Final published version

Published in
User Modeling and User-Adapted Interaction

User Modeling and User-Adapted Interaction (2024) 34:1–58
https://doi.org/10.1007/s11257-023-09363-0

Evaluating explainable social choice-based aggregation strategies for group recommendation

Francesco Barile[1] · Tim Draws[2] · Oana Inel[3] · Alisa Rieger[2] · Shabnam Najafian[2] · Amir Ebrahimi Fard[1] · Rishav Hada[1,4] · Nava Tintarev[1]

Received: 1 August 2022 / Accepted in revised form: 17 March 2023 / Published online: 21 June 2023

Abstract

Social choice aggregation strategies have been proposed as an explainable way to generate recommendations to groups of users. However, it is not trivial to determine the best strategy to apply for a specific group. Previous work highlighted that the performance of a group recommender system is affected by the internal diversity of the group members' preferences. However, few of them have empirically evaluated how the specific distribution of preferences in a group determines which strategy is the most effective. Furthermore, only a few studies evaluated the impact of providing explanations for the recommendations generated with social choice aggregation strategies, by evaluating explanations and aggregation strategies in a coupled way. To fill these gaps, we present two user studies (N=399 and N=288) examining the effectiveness of social choice aggregation strategies in terms of users' fairness perception, consensus perception, and satisfaction. We study the impact of the level of (dis-)agreement within the group on the performance of these strategies. Furthermore, we investigate the added value of textual explanations of the underlying social choice aggregation strategy used to generate the recommendation. The results of both user studies show no benefits in using social choice-based explanations for group recommendations. However, we find significant differences in the effectiveness of the social choice-based aggregation strategies in both studies. Furthermore, *the specific group configuration* (i.e., various scenarios of internal diversity) seems to determine the most effective aggregation strategy. These results provide useful insights on how to select the appropriate aggregation strategy for a specific group based on the level of (dis-)agreement within the group members' preferences.

Keywords Group recommender systems · Social choice functions · Explainable recommender systems · Social choice-based explanations

✉ Francesco Barile
f.barile@maastrichtuniversity.nl

Extended author information available on the last page of the article

1 Introduction

Recommender systems have become ubiquitous in people's lives, but there is increasing demand for recommendations that serve several people simultaneously. For instance, in domains such as online communities (Chen et al. 2008; Kim et al. 2010), music, movies or TV programs (O'connor et al. 2001; Masthoff 2004; Najafian and Tintarev 2018; Cao et al. 2018), and tourism (Cao et al. 2018; Najafian et al. 2020a), people often consume recommendations in groups rather than individually. Group recommender systems (GRSs) (Masthoff 2015; Masthoff and Delić 2022) are designed to provide recommendations that meet different group members' preferences to support the group decision-making process. Several approaches have been proposed for performing this task, with most of them aiming to aggregate the individual group members' preferences or recommendations (Senot et al. 2010). This aggregation is typically performed by applying *social choice-based aggregation strategies*, which combine the individual preferences of all group members following different approaches to predict an item that is suitable for everyone (Masthoff 2004, 2015; Najafian et al. 2020a).

Each group recommendation aggregation strategy has its trade-offs: *Arrow's theorem* (Arrow 1950) states that the performance of an aggregation strategy depends on the evaluation context, meaning that it is unlikely for an aggregation strategy to outperform other strategies in all situations. However, previous user studies have demonstrated that some strategies perform better than others in different experimental conditions, in terms of perceived group satisfaction (Masthoff 2015). Other research has accordingly proposed to adaptively select the social choice-based aggregation strategy based on characteristics of the considered group; specifically, they applied strategies focused on avoiding misery, considering the average satisfaction, or maximizing happiness depending on the *relationship strength* between group members (Gartrell et al. 2010; Zhang et al. 2019). However, the selection of the best strategy to apply to each specific group was based on intuitions or assumptions rather than determined by empirical findings with people. Another factor that was analyzed regarding the group decision-making performance and each group member's satisfaction with the outcome is the *intra-group diversity* in terms of individual preferences. Delic et al. (2020) showed that a higher preference diversity generally has negative effects on these factors. However, determining which aggregation strategy performs better for a specific group is still an open problem that inspired the studies described in this paper.

If the aggregation strategy results in a recommendation that is not intuitive or not ideal for some group members, an *explanation* could help the group members to make a decision or reach consensus. Traditionally, explanations in recommender systems have been designed for single users and have achieved goals such as transparency, trust, and scrutability (Chen et al. 2013; Gedikli et al. 2014; Jannach et al. 2010; Tintarev and Masthoff 2022). However, explanations for groups need to meet additional goals besides explaining why certain items are recommended (Felfernig et al. 2018; Ntoutsi et al. 2012)—they need to help users agree on a joint decision and improve users' perceived fairness, perceived consensus, and satisfaction with the group's decision (Felfernig et al. 2018; Najafian and Tintarev 2018; Tran et al. 2019). To the best of our knowledge, few studies have focused on generating and evaluating explanations based on social choice aggregation strategies aiming to increase fairness and consensus

perception of users or their satisfaction (Tran et al. 2019). Crucially, though, they evaluated explanations and aggregation strategies in a coupled way, not distinguishing whether the participants' evaluations referred to the explanation or the underlying aggregation strategy.

In this paper, we present two user studies investigating the performance of different social choice-based aggregation strategies in terms of users' *fairness perception*, *consensus perception*, and *satisfaction*. Furthermore, we study the impact of the level of (dis-)agreement within the group on the performance of several social-choice aggregation strategies. We define this *group configuration* on the basis of the similarity between group members' individual preferences. Finally, given that the social choice strategies are highly explainable, we also explore the added value of explanations. These explanations describe to the group the aggregation strategy used to produce the recommendation.

Our first experiment, EVALUATING THE EFFECTIVENESS OF EXPLAINABLE SOCIAL CHOICE- BASED AGGREGATIONS (see Sect. 3; a version of which was previously published in Barile et al. (2021)) addresses the research question **RQ1**: *"Do explainable social choice-based aggregation strategies increase users' fairness perception, consensus perception, or satisfaction?"*.

To answer this question, we conducted a preregistered, between-subjects user study with 399 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] We experimented with five aggregation strategies (i.e., *Additive Utilitarian*; ADD, *Approval Voting*; APP, *Least Misery*; LMS, *Majority*; MAJ, and *Most Pleasure*; MPL) and three types of explanations and thus 15 conditions in total. In addition, we also tested for interaction effects between aggregation strategies and explanation types. Our results show differences between the social choice aggregation strategies for the studied group scenario in terms of users' perceptions of fairness, consensus, and satisfaction. However, in contrast to earlier work (Tran et al. 2019), we found no added value in accompanying the aggregation strategies with social choice-based explanations.

We conducted a second user study to further investigate which factors influence the effectiveness of social choice-based aggregation strategies and their related explanations. This second experiment, THE IMPACT OF SCENARIO COMPLEXITY (see Sect. 4), investigated the impact of scenario *complexity* in terms of the number of group members, the number of possible items, and the diversity of group members' preferences. Specifically, we defined a set of *group configurations* based on the internal (dis-)agreement between group members to present complex scenarios to the evaluators: (i) *uniform*, which characterizes a group with high internal agreement between group members; (ii) *divergent*, a group with low internal agreement; (iii) *coalitional*, a group characterized by two disjoint subgroups with high internal agreement and low inter-subgroup agreement; and (iv) *minority*, a group with high internal agreement, except for one member who has a low agreement with all the other group's members. This experiment addresses the research question **RQ2**: *"Do explainable social*

[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.

choice-based aggregation strategies increase users' fairness perception, consensus perception, or satisfaction, in complex group recommendation scenarios?".

To answer this research question, we conducted a randomized controlled trial using a mixed design with two between-subject factors (6x2=12 groups) and one within-subject factor (4 conditions).[2] In this experiment, we focus on six aggregation strategies, namely *Additive Utilitarian* (ADD), *Fairness* (FAI), *Approval Voting* (APP), *Least Misery* (LMS), *Majority* (MAJ), and *Most Pleasure* (MPL) and two types of explanations. We found significant differences between social choice-based aggregation strategies in terms of users' fairness perception, consensus perception, and satisfaction. Furthermore, our results show differences in the effectiveness of the social choice-based aggregation strategies *depending on the specific configuration of the group for which the aggregation strategies are applied*. A deeper investigation of the performances of the aggregation strategies in the specific group configuration revealed useful insights on which strategies perform better for each group configuration: the MPL strategy performs worst for minority groups but is one of the best strategies for uniform groups; the FAI strategy has good effectiveness for uniform and coalitional groups, while for divergent groups the ADD strategy obtains the best results. However, as in the first study, we found no added value in adding social choice-based explanations.

In sum, this paper makes the following contributions:

- We conduct two preregistered user studies ($N = 399$ in the first study and $N = 288$ in the second study) to evaluate the effectiveness of social choice-based aggregation strategies and explanation types, and the impact of the group configuration, defined based on the internal (dis-)agreement among the group members.
- We show significant differences among the aggregation strategies in terms of users' fairness perception, consensus perception, and satisfaction, related to the provided group recommendations.
- We found that the effectiveness of aggregation strategies depends on the configuration of the group on which the strategy is applied: (i) Most Pleasure (MPL) should be avoided for a minority group configuration, while it is the preferable strategy for a uniform group; (ii) the Fairness (FAI) strategy is preferred for uniform and coalitional groups; (iii) the Additive (ADD) strategy may be used in the situations in which the group configuration is not clearly identifiable.

2 Related work

In this section, we introduce the social choice-based aggregation strategies used to generate recommendations for groups in the two studies. Then, we describe the relevant literature on explanations for group recommender systems. We conclude the section by introducing the most recent lines of research in the Group Recommender state-of-the-art. Given relevant studies and findings, we also provide an overview of the gaps in the literature that we address in our studies.

[2] We preregistered the user study, publicly determining our research questions, hypotheses, experimental setup, and data analysis plan before any data collection. All the preregistration material can be found at https://osf.io/3dcht/.

2.1 Social choice-based aggregation strategies

There are two main approaches to generating group recommendations: *(i) aggregated models* that aggregate individual preferences (e.g., existing ratings) into a group model and then generate the group recommendations based on such a group model and *(ii) aggregated predictions* or strategies that aggregate individual item-ratings predictions and recommend items with the highest aggregated scores to the group (Felfernig et al. 2018). Several aggregation strategies inspired by *Social Choice Theory* (Kelly 2013) have been proposed to aggregate individuals' information for group recommendations (Masthoff 2015). Masthoff (2004) present an overview of these *social choice-based aggregation strategies*. Six of the most utilized social choice-based aggregation strategies are:

- *Additive Utilitarian (ADD)* is a *consensus-based* strategy that considers the preferences of all group members and recommends the item with the highest sum of all group members' ratings (Senot et al. 2010).
- *Fairness (FAI)* is a *consensus-based* strategy well-suited for repeated decisions, as it ranks items according to how individuals choose them in turn (Masthoff 2015).
- *Approval Voting (APP)* is a *majority-based* strategy, focusing on the most popular items among group members, recommending the item with the highest number of ratings above a predefined threshold (Senot et al. 2010).
- *Least Misery (LMS)* is a *borderline* strategy, considering only a subset of group members' preferences and recommends the item which has the highest of all lowest ratings (Senot et al. 2010).
- *Majority (MAJ)* is a *borderline* strategy that recommends the item with the highest number of all ratings representing the majority of item-specific ratings (Senot et al. 2010).
- *Most Pleasure (MPL)* is a *borderline* strategy that recommends the item with the highest individual group member rating (Senot et al. 2010).

Masthoff and Delić (2022) presents several experiments performed to determine the best strategy in terms of perceived group satisfaction. The results, however, show that there is no "winning" strategy, as different strategies perform well in two different experimental settings. In the first study, the participants were asked to determine the best recommendation list for a group by inspecting group members' preferences and explain the strategy they adopted. In the second, participants were presented with recommendation lists provided by different aggregation strategies and asked to determine the best in terms of the group members' satisfaction. The results of these experiments were contradictory, suggesting that the strategies can have different performances in different group recommendation settings. Based on these considerations, in our experiments, we aim to evaluate the differences between aggregation strategies in terms of fairness perception, consensus perception, and satisfaction.

We note that the experiments in this paper are complimentary to both long-standing and more recent research on group recommender systems, as many approaches propose variations of the strategies described above. The most common approach incorporates personal and social factors influencing the group decision-making process into social choice-based aggregation strategies. More specifically, these approaches assign dif-

ferent weights to the user's preferences, considering demographics (Ardissono et al. 2003), roles in the group (Berkovsky and Freyne 2010), user's experience in the domain (Gartrell et al. 2010), centrality in the group social network (Rossi et al. 2015, 2016; Delic et al. 2018), or individuals' personalities (Nguyen et al. 2019; Quijano-Sanchez et al. 2017; Rossi et al. 2018). Another body of work attempts to balance the satisfaction of group members in relation to a sequence of items rather than the satisfaction with an individual item. More precisely, they aim to learn aggregation strategies directly from group interactions (Cao et al. 2018; Vinh Tran et al. 2019; Sankar et al. 2020) instead of trying to ensure greater fairness (Kaya et al. 2020; Malecek and Peska 2021). These new exciting directions show strong predictive performance but also exhibit limitations in terms of explainability; we believe that the explanation methodologies introduced in this paper can provide a foundation on which these approaches can build to fill this gap.

2.2 Explaining to groups

Explanations can generally be seen as additional information that is associated with the recommendations to achieve several goals, such as increasing *transparency* (explaining how the recommendation system works), *effectiveness* (helping the user to make good decisions), and *usability* of the system, as well as user *satisfaction* (Tintarev and Masthoff 2022). Several studies in different domains have shown the benefits of using explanations for recommendations to increase users' acceptance rate, satisfaction, and trust in the system (Sinha and Swearingen 2002). In group recommendations, explanations can achieve further goals: *fairness* (showing consideration for all group members' preferences as much as possible); *consensus* (helping group members agree on the decision) (Felfernig et al. 2018); *privacy-preservation* (preserving 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) (Najafian et al. 2021a, 2020b, 2021b). However, most of the research on explanations for recommender systems focuses 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 (Najafian and Tintarev 2018; Kapcak et al. 2018; Tran et al. 2019).

Natural language explanation styles based on the underlying social choice aggregation strategies were introduced in Najafian and Tintarev (2018), while Kapcak et al. (2018) extended this work using the wisdom of the crowd to improve the quality of the initially proposed explanations. Quijano-Sanchez et al. (2017) included the social factors of personality and tie strength between group members to generate tactful explanations (e.g., explanations that avoid damaging friendships). In a user study, Tran et al. (2019) evaluated explanations for six social choice-based aggregation strategies and found that explanations related to the ADD and MAJ strategies most increased *fairness* and *consensus* perceptions, as well as user *satisfaction* regarding the group recommendation. They also found that users' perceived fairness or consensus correlated with their satisfaction. Although this paper presents valuable ways to generate explanations for the most used benchmark aggregation strategies in group

recommender systems research, the *joint* evaluation of aggregation strategies and explanations raises questions regarding whether the effects attributed to the explanations might not, in fact, depend solely on the aggregation strategies themselves. On the contrary, in our work, we evaluate the effectiveness of explanations in isolation. Furthermore, a second aspect that has not been investigated in the literature is the level of detail that the explanation can achieve concerning the aggregation strategy used and whether this affects users' fairness perception, consensus perception, and satisfaction. For this reason, we will also evaluate the effectiveness of a *detailed* version of the social choice-based explanations proposed in Tran et al. (2019). Finally, we also validate the correlation between user satisfaction and perceived fairness and consensus, *c.f.*, Tran et al. (2019).

2.3 Complex group recommendation scenarios

Although the group recommender systems literature tends to focus on a specific strategy at a time, some comparative studies showed that different aggregation strategies perform better depending on some characteristics of the specific group the system is supporting (Masthoff and Delić 2022). In particular, Gartrell et al. (2010) and Zhang et al. (2019) propose two approaches in which the strength of the relationships between the group members is used to evaluate a *social factor*, and this is used to determine the aggregation strategy to use: MPL for groups characterized by strong relationships, AVG for groups with intermediate relationships, and LMS for groups with weak relationships. However, the motivation behind these choices is based on anecdotal observations on limited numbers of groups (Gartrell et al. 2010), and we are not aware of any studies validating these assumptions in the literature.

Another factor that was analyzed in relation to the performance of the group decision-making process and the satisfaction of each group member with the outcome is the intra-group diversity (in terms of individual preferences). Delic et al. (2020) defined several measures of internal similarity and evaluated the correlations of such metrics with group member satisfaction regarding the presented recommendations, using real groups data collected in the tourism domain. The results showed that, in general, a higher diversity has negative effects on these factors. The internal dissimilarity is also used by Gartrell et al. (2010), where a dissimilarity descriptor is used to correct the aggregated scores.

Based on these findings, we argue that the complexity of the considered group scenario has an impact on the performance of the aggregation strategies, informing the study on THE IMPACT OF SCENARIO COMPLEXITY. We further hypothesize that the complexity of the recommendation scenario increases the difficulty of evaluating the effectiveness of the group recommendations provided for a specific complex group. In such conditions, the users may benefit from the use of social choice-based explanations. To properly evaluate complex group recommendation scenarios, we consider different settings. In contrast to previous work (c.f., Delic et al. (2020)), we consider the composition of the group rather than averaging the dissimilarity between all the pairs of group members and hypothesize that different complex group configurations lead to different performances for the considered recommendation strategies.

Our literature overview highlights the need for more empirical analysis of social choice-based aggregation strategies. Furthermore, there is a clear need for measuring how factors such as scenario complexity and group configuration influence the performance of the aggregation strategies.

3 The effectiveness of social choice strategies

In this first study, EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES, we evaluated the effectiveness of social choice-based explanations for group recommendations in terms of perceived fairness, consensus, and satisfaction. In particular, we decided to use a slightly different methodology than the one proposed in the literature. In Tran et al. (2019), the recommendation was presented together with the explanations, and the user was asked to evaluate how the provided explanation helped increase their fairness perception, consensus perception, and satisfaction with the recommendation. In our user study, we aimed to conceptually replicate and further investigate the findings of Tran et al. (2019) by decoupling the explanation and the recommendation. More specifically, we added a control condition without explanation, and we asked our participants to evaluate the provided recommendation. This choice was motivated by the following consideration: if the explanation is actually helpful in increasing the fairness perception, consensus perception, and satisfaction related to the recommendation, users' evaluations should be more favorable in the scenarios where recommendations are provided together with an explanation. Conversely, the recommendation provided in the control scenario, i.e., without explanation, should receive lower evaluations. To guide our first study, we decomposed **RQ1** (performance of different explainable aggregation strategies) into four sub-questions:

RQ1.1 Are there differences between *social choice-based aggregation strategies* in group recommendation settings regarding users' fairness perception, consensus perception, or satisfaction?
RQ1.2 Do explanations that are based on the group recommendation aggregation strategy at hand increase users' fairness perception, consensus perception, or satisfaction?
RQ1.3 Does the effectiveness of explanations (w.r.t. users' fairness perception, consensus perception, or satisfaction) vary depending on the aggregation strategies at hand?
RQ1.4 Are users' levels of perceived fairness or perceived consensus related to their satisfaction concerning the group recommendations?

3.1 Hypotheses

In this section, we formalize the hypotheses related to the research questions RQ1.1–4 that we investigate in our experiment. First, based on the findings from Masthoff and Gatt (2006); Masthoff and Delić (2022), we hypothesize that we have different performances in terms of fairness perception, consensus perception, and satisfaction

for the aggregation strategies considered. More specifically, we formalize the following hypotheses related to **RQ1.1**[3]:

- **H1.1a**: There is a difference between social choice-based aggregation strategies in group recommendation settings regarding users' fairness perception.
- **H1.1b**: There is a difference between social choice-based aggregation strategies in group recommendation settings regarding users' consensus perception.
- **H1.1c**: There is a difference between social choice-based aggregation strategies in group recommendation settings regarding user satisfaction.

Furthermore, we hypothesize to have a positive impact from the presence of explanations, in line with the findings presented in Tran et al. (2019). Hence, we formulate a second set of hypotheses related to **RQ1.2**:

- **H1.2a**: Explanations based on the aggregation strategy at hand increase users' fairness perception concerning group recommendations.
- **H1.2b**: Explanations based on the aggregation strategy at hand increase users' consensus perception concerning group recommendations.
- **H1.2c**: Explanations based on the aggregation strategy at hand increase users' satisfaction concerning group recommendations.

We also hypothesize that the effectiveness of the explanations is moderated by the underlying aggregation strategy for all the three variables we are measuring, which translates into the following hypotheses related to **RQ1.3**:

- **H1.3a**: The effect of aggregation strategy-based explanations on users' fairness perception concerning group recommendations is moderated by the type of aggregation strategy at hand.
- **H1.3b**: The effect of aggregation strategy-based explanations on users' consensus perception concerning group recommendations is moderated by the type of aggregation strategy at hand.
- **H1.3c**: 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 hypothesize to confirm the finding from Tran et al. (2019) regarding the correlation between users' perceived fairness, perceived consensus, and satisfaction. Hence, we formulate the following hypotheses related to **RQ1.4**:

- **H1.4a**: Users' perceived fairness is positively related to user satisfaction concerning group recommendations.
- **H1.4b**: Users' perceived consensus is positively related to user satisfaction concerning group recommendations.

[3] 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.

3.2 Method

We conducted an online between-subjects user study to test the aforementioned hypotheses,[4] Users were presented with a scenario that reflected one of five different social choice-based aggregation strategies for group recommender systems and included either no explanation or one of two different explanation types.

3.2.1 Materials

Our study considered five aggregation strategies and two explanation types.

Aggregation strategies

We considered five social choice-based aggregation strategies for group recommender systems in our first study. More specifically, we evaluated the following aggregation strategies (see Sect. 2.1 for more details): ADD, APP,[5] LMS, MAJ, and MPL. These strategies aggregate the preferences of a group of users to obtain a recommendation for the group as a whole Senot et al. (2010). All these strategies were also evaluated in prior work by Tran et al. (2019). However, in contrast to Tran et al. (2019), we do not consider FAI because the explanation types proposed in our study cannot be generated for this strategy in the considered scenario, as it needs more interactions with the system.

Explanations

In our user study, each recommendation is paired with one of the following explanation types:

- *No explanation*: the aggregation strategy is applied without explanation.
- *Basic explanation*: illustrates the aggregation strategy with a short sentence. We adopted them from Tran et al. (2019), where they are referred to as *Type 1* explanations.
- *Detailed explanation*: extends basic explanations by providing details about the specific reason why a given item has been recommended.

Table 1 illustrates the specific explanation types for each aggregation strategy.

3.2.2 Procedure

After participants agreed to an informed consent, they were introduced to the study and asked for their gender and age. Then, they saw the scenario from Tran et al. (2019):

[4] 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 at https://osf.io/5xbgf/.

[5] Following Tran et al. (2019), we consider a threshold of 3 for the APP strategy.

Table 1 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 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}, $u_{j_2}, \ldots$ 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}, $u_{j_2}, \ldots$ 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}, $u_{j_2}, \ldots$ and $u_{j_{\bar{n}}}$ gave it the rating r, which is the highest rating among all items' high ratings)."

Let $G = \{u_1, \ldots, u_n\}$ 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

Table 2 Ratings of group members for the restaurants (1: the worst, 5: the best) from Tran et al. (2019)

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

> "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:"

Participants subsequently saw Table 2 and were presented with a recommendation generated with one of the five considered aggregation strategies. The recommendation was presented either with or without an explanation, depending on the explanation type they had been randomly assigned to (i.e., one of the fifteen possible conditions; determined by the combination of the five considered aggregation strategies and the three explanation types; see Table 1). Finally, we asked them to evaluate the perceived fairness, perceived consensus, and satisfaction (see Sect. 3.2.3) from the point of view of an "external evaluator" (i.e., not a member of the group for which the recommendation was generated). To ensure high quality of the collected results, we included one attention check in which the participant is instructed to select a specific option. Finally, participants could provide a textual explanation for their answers. Before we ran it, the study had been reviewed and approved by the *Human Research Ethics Committee* at TU Delft.[6]

3.2.3 Variables

This section introduces the independent, dependent, and descriptive variables measured in the user study.

Independent variables

The independent variables defined the conditions presented to the participants, in terms of *aggregation strategy* and *explanation type*.

- 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 Sect. 3.2.1).
- Explanation type (categorical, between-subjects). Each participant saw either *no explanation*, a *basic explanation*, or a *detailed explanation* (see Sect. 3.2.1).

[6] https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics.

Dependent variables

Inspired by Tran et al. (2019), we asked the participants to evaluate the provided scenario in terms of perceived fairness, perceived consensus, and satisfaction. For this, we asked the participants to respond to a statement for each variable on a seven-point Likert scale ranging from "strongly agree" (scored as −3) to "strongly disagree" (scored as 3). The statements are adapted from Tran et al. (2019) such that the participant is asked to evaluate the provided recommendation (and not the explanation). Below, we list the statements:

- Perceived fairness (ordinal): "The group recommendation is fair to all group members";
- Perceived Consensus (ordinal): "The group members will agree on the group recommendation";
- Satisfaction (ordinal): "The group members will be satisfied with regard to the group recommendation".

Descriptive variables

In addition to the independent and dependent variables that we used for hypothesis testing, we collected data on two demographic variables:

- Age (categorical), participants could select one of the options *18–25*, *26–35*, *36–45*, *46–55*, *>55*;
- Gender (categorical). Participants could select one of the options *female*, *male*, or *other*.

There was also a "prefer not to say" option for both variables.

3.2.4 Sample size determination

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 Sect. 3.2.6) using *G*Power* (Faul et al. 2007). Here, we specified the default effect size f = 0.25, a significance threshold $\alpha = \frac{0.05}{11} \approx 0.005$ (due to testing multiple hypotheses; see Sect. 3.2.6), a power of $(1 - \beta) = 0.8$, and that we 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.

3.2.5 Participants

We recruited 400 participants from the online participant pool *Prolific*,[7] all of whom were proficient English speakers above 18 years of age. To maintain high-quality

[7] https://prolific.co.

answers, we selected only participants who had an approval rate of at least 90% and participated in at least ten 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 who did not pass the attention check from the data analysis. The resulting sample of 399 participants was composed of 61% (244) female, 38% (153) male, and 1% (2) other participants. They represented a diverse range of age groups: 28% (110) were between 18 and 25, 29% (115) between 26 and 35, 17% (68) between 36 and 45, 14% (55) between 46 and 55, and 13% (51) were above 55 years of age. Additional information on the dataset demographic distributions are available in Appendix A. We randomly distributed participants over the 15 conditions (i.e., exposing them to one out of five aggregation strategies and one out of three explanation types).

3.2.6 Statistical analysis

For each of the three dependent variables in our study (i.e., *fairness perception*, *consensus perception*, and *satisfaction*), we conducted a two-way ANOVA using *aggregation strategy* and *explanation type* as between-subjects factors. These three ANOVAs were used to test nine hypotheses (i.e., **H1.1a** – **H1.3c**). Specifically, each of them tested main effects of *aggregation strategy* (**H1.1a**–**H1.1c**) and *explanation type* (**H1.2a**–**H1.2c**), as well as the interaction between these two variables in affecting the dependent variables (**H1.3a** – **H1.3c**). 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 (Norman 2010). We additionally performed two Spearman correlation analyses to test hypotheses **H1.4a** and **H1.4b**. We thus tested 11 different hypotheses. Applying a Bonferroni correction (Napierala 2012), 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 (**H1.1a**–**H1.2c**; see Sect. 3.3), we conducted Tukey post hoc analyses to investigate specific differences between the aggregation strategies and explanation types. The *p*-values from these post hoc analyses were adjusted to correct for multiple testing (i.e., written as p_{adj}).

3.3 Results

The results of the statistical analyses illustrated in Sect. 3.2.6 are reported in Table 3. First, we report some descriptive statistics about the collected data. Then, we highlight the results related to the research questions RQ1.1–4.

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 was shown to 6–7% 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 the task were

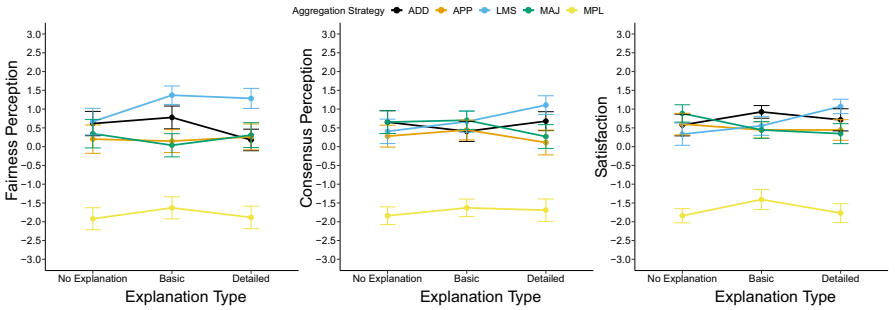


Fig. 1 Participants' mean *fairness perception*, *consensus perception*, and *satisfaction* across explanation types on scales from −3 ("strongly disagree") to 3 ("strongly agree"; see Sect. 4.2.3). Colors indicate aggregation strategies: Additive Utilitarian (ADD), Approval Voting (APP), Least Misery (LMS), Majority (MAJ), Most Pleasure (MPL). Error bars represent the standard error of the mean

understandable. Participants had a slight overall tendency to perceive fairness, consensus, and satisfaction across scenarios, as 51%, 51%, and 56% overall at least somewhat agreed with these three items, respectively. Figure 1 shows participants' mean *fairness perception*, *consensus perception*, and *satisfaction* across explanation types and split by aggregation strategies.

RQ1.1: differences between social choice-based aggregation strategies regarding the recommendation effectiveness. We found significant differences between the five aggregation strategies concerning all three dependent variables *fairness perception*, *consensus perception*, and *satisfaction* (**H1.1a** – **H1.1c**; $F = [36.19, 38.89, 49.57]$, all $p < 0.001$; see Table 3). So, overall, participants expressed different levels regarding these three variables based on which aggregation strategy they were exposed to. Tukey pairwise post hoc analyses revealed that Most Pleasure (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 were that Approval Voting (APP) ($p_{\text{adj}} = 0.004$) and Majority (MAJ) ($p_{\text{adj}} = 0.005$) each led to lower fairness perception compared to Least Misery (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 and APP as less fair compared to LMS.

RQ1.2: 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 (**H1.2a** – **H1.2c**; $F = [0.35, 0.14, 0.15]$, $p = [0.71, 0.87, 0.86]$; see Table 3). So, our results show no difference between explanation types concerning our three dependent variables.

RQ1.3: interactions between aggregation strategies and explanation types regarding explanation effectiveness. There were no significant interaction effects between aggregation strategies and explanation types (**H1.3a** – **H1.3c**; $F = [0.68, 0.75, 1.25]$, $p = [0.71, 0.65, 0.27]$; see Table 3). The effect of explanation types on participants'

Table 3 Results of three two-way ANOVAs for the dependent variables (DVs) *fairness perception* (left), *consensus perception* (center), and *satisfaction* (right)

DV: fairness perception			DV: consensus perception			DV: satisfaction		
	F	*p*		*F*	*p*		*F*	*p*
(H1.1a) aggr	36.19	< 0.001*	**(H1.1b)** aggr	38.89	< 0.001*	**(H1.1c)** aggr	49.57	< 0.001*
(H1.2a) expl	0.35	0.71	**(H1.2b)** expl	0.14	0.87	**(H1.2c)** expl	0.15	0.86
(H1.3a) aggr:expl	0.68	0.71	**(H1.3b)** aggr:expl	0.75	0.65	**(H1.3c)** aggr:expl	1.25	0.27

Per effect, we report the *F*-statistic, and *p*-value. The terms "aggr" and "expl" represent the independent variables *aggregation strategy* and *explanation type*. Colons indicate interaction effects, asterisks statistical significance

fairness perception, *consensus perception*, and *satisfaction* thus did not significantly differ based on which aggregation strategy was applied.

RQ1.4: associations between explanation effectiveness measures. In line with the findings of Tran et al. (2019), 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.

3.4 Discussion

In this section, we look closer at the user study results and their implications. We discuss the difference between aggregation strategies and explanation levels, and the correlation between fairness perception, consensus perception, and satisfaction.

3.4.1 Differences between aggregation strategies

As shown in Sect. 3.3, we found differences between the aggregation strategies in terms of perceived fairness, perceived consensus, and satisfaction. The Most Pleasure (MLP) strategy obtained the lowest scores, regardless of the type of explanation.

Furthermore, participants perceived the Majority (MAJ) and Approval Voting (APP) strategies as less fair than Least Misery (LMS). These results are in contrast to the findings of Tran et al. (2019), where the same scenario was used. There, the Majority (MAJ) and Additive (ADD) strategies scored *better* than the Least Misery (LMS) strategy. An explanation of this difference could be the different design of our experiment: we implemented a between-subject design to guarantee the independence between the conditions; on the contrary, in Tran et al. (2019), 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 their 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. (2019) asked the participants to evaluate the explanation. Hence, the evaluation of the explanation was influenced by the evaluation of the aggregation strategy.

3.4.2 The role of explanations

The results presented showed no significant difference between the different types of explanations. Furthermore, we found no interaction effects between the explanations and the aggregations regarding the measured dependent variables (perceived fairness, perceived consensus, and satisfaction). 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 formalize the explanations for these strategies, as they serve as a reference against which more articulated strategies can be compared.

3.4.3 The link between fairness, consensus, and satisfaction

The correlation between fairness perception (or consensus perception) and satisfaction, already reported in Tran et al. (2019), and also shown 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 consider them both in the generation of recommendations and in their evaluation.

3.4.4 The impact of the considered scenario

One important limitation of this study concerns the used scenario (see Table 2 in Sect. 3.2.2). In particular, we here considered a group with only four people and three items, where three group members mostly agree on the items' evaluations, while only one user has quite different preferences. A more realistic scenario, with more options and different ratios between the group members' agreements and disagreements, could lead to different results and, in general, to different effectiveness for the presented explanations. We addressed this specific limitation in the second user study (see Sect. 4).

4 Complex recommendation scenarios

In our first study on EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES, we found differences between different social choice-based aggregation strategies. However, as we can see from Table 2, the case we studied was reasonably simple (four group members and three candidate restaurants), and it is not clear whether these results generalize to more complex recommendation scenarios. In particular, previous work (Masthoff and Gatt 2006; Delic et al. 2020; Gartrell et al. 2010) suggests that there is a benefit to adapting the aggregation strategy to the group composition, but this has not yet been systematically evaluated. Intuitively, finding a good solution in a group with diverging preferences can be more difficult. This, in turn, can impact the added value of the (aggregated) recommendations and the corresponding explanations. Thus, to investigate the role of group composition, we investigate the effectiveness of aggregation strategies and explanations in a slightly more complex scenario (with five group members and ten items).

Furthermore, we defined a scenario in which the system has been used three times in the past. Hence, the provided recommendation is the fourth choice of the consid-

ered aggregation strategy. This also allowed us to consider and properly evaluate the Fairness (FAI) strategy, which was excluded from our previous study (see Sect. 4.2.1).

In this study, we investigate **RQ2** (influence of more complex scenarios), which is divided into four sub-questions:

RQ2.1 Are there differences between the social choice-based aggregation strategies (w.r.t. users' fairness perception, consensus perception, or satisfaction) in complex group recommendation scenarios?

RQ2.2 Do social choice-based explanations increase users' fairness perception, consensus perception, or satisfaction in complex group recommendation scenarios?

RQ2.3 Are there differences between different group configurations in terms of users' fairness perception, consensus perception, or satisfaction?

RQ2.4 Does the effectiveness of social choice-based explanations (w.r.t. users' fairness perception, consensus perception, or satisfaction) vary depending on the underlying aggregation strategies and/or on the group configuration?

4.1 Hypotheses

In this section, we formalize the hypotheses related to RQ2.1–4. First, based on the evidence from the literature (Delic et al. 2020; Masthoff and Gatt 2006), we hypothesize that there are differences between the aggregation strategies in terms of fairness perception, consensus perception, and satisfaction when evaluated in complex recommendation scenarios. Hence, we formulate the following hypotheses related to **RQ2.1**:

- **H2.1a**: There is a difference between social choice-based aggregation strategies regarding users' fairness perception in complex group recommendation scenarios.
- **H2.1b**: There is a difference between social choice-based aggregation strategies regarding users' consensus perception in complex group recommendation scenarios.
- **H2.1c**: There is a difference between social choice-based aggregation strategies regarding users' satisfaction in complex group recommendation scenarios.

Furthermore, we hypothesize that the complexity of the scenario triggers the user's need for explanations, resulting in increased effectiveness of social choice-based explanations. Hence, we formulate the following hypotheses related to **RQ2.2**:

- **H2.2a**: Social choice-based explanations increase users' fairness perception in complex group recommendation scenarios.
- **H2.2b**: Social choice-based explanations increase users' consensus perception in complex group recommendation scenarios
- **H2.2c**: Social choice-based explanations increase users' satisfaction in complex group recommendation scenarios.

We also hypothesize that we observe different effectiveness values according to the specific group configuration of the group for which the recommendation is provided. More specifically, we formulate the following hypotheses related to **RQ2.3**:

- **H2.3a**: There is a difference between group configurations regarding users' fairness perception concerning group recommendations.

- **H2.3b**: There is a difference between group configurations regarding users' consensus perception concerning group recommendations.
- **H2.3c**: There is a difference between group configurations regarding user satisfaction concerning group recommendations.

Finally, regarding **RQ2.4**, we hypothesize that the effectiveness of social choice-based explanations is moderated by the considered aggregation strategy and the specific group configuration:

- **H2.4a**: The effect of social choice-based explanations on users' fairness perception concerning group recommendations is moderated by the underlying social choice aggregation strategy.
- **H2.4b**: The effect of social choice-based explanations on users' consensus perception concerning group recommendations is moderated by the underlying social choice aggregation strategy.
- **H2.4c**: The effect of social choice-based explanations on user satisfaction concerning group recommendations is moderated by the underlying social choice aggregation strategy.
- **H2.4d**: The effect of social choice-based explanations on users' fairness perception concerning group recommendations is moderated by the characteristics of the group to which the recommendation is provided.
- **H2.4e**: The effect of social choice-based explanations on users' consensus perception concerning group recommendations is moderated by the characteristics of the group to which the recommendation is provided.
- **H2.4f**: The effect of social choice-based explanations on user satisfaction concerning group recommendations is moderated by the characteristics of the group to which the recommendation is provided.

4.2 Method

To evaluate the effectiveness of social choice-based aggregation strategies and their explanations for complex group scenarios, we introduced four group configurations, defined based on the internal similarity between group members' evaluations of the possible options. These group configurations are introduced in Sect. 4.2.1, together with the considered *aggregation strategies* and *explanation types*. Then, we present the method of the user study,[8]

4.2.1 Materials

This section introduces the group configurations, aggregation strategies, and explanation types used in this second user study.

[8] All material for analyzing our results and replicating our user study (i.e., document with preregistration of all the hypotheses tested, and the analysis scripts) is publicly available – https://osf.io/3dcht/. The anonymized gathered data is available at https://doi.org/10.34894/8EVX4U.

Group configurations

As mentioned before, inspired by existing literature (Delic et al. 2020), we argue that the complexity of the group configuration, in terms of internal similarity between the group members' evaluations, has an impact on the effectiveness of the recommendations and the corresponding explanations. First, we decided to increase the complexity of the scenario by using a higher number of items and group members: we decided to use five group members and ten items. We introduced four group configurations:

- *Uniform*: characterized by a low internal diversity between group members' preferences.
- *Divergent*: characterized by a high internal diversity between group members' preferences.
- *Coalitional*: characterized by two disjoint subgroups having low inter-group diversity and high intra-group diversity.
- *Minority*: characterized by a subgroup with N-1 users with low internal diversity, where all the N-1 users have a high diversity with the remaining user.

Figure 2 illustrates the group configurations for groups of five people. For more details on the group configurations and the generation of the relative scenarios, the interested reader can refer to Appendix B.

Aggregation strategies

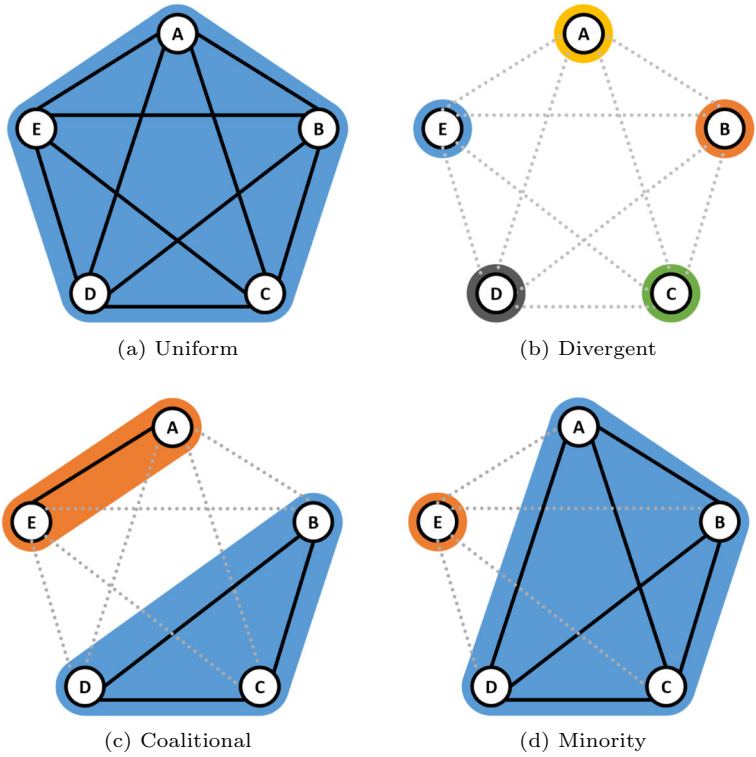
In this second study, we focus on six aggregation strategies for group recommender systems: ADD, APP (considering a threshold equal to 3 as previously), LMS, MAJ, MPL, and FAI. The FAI strategy was included since, in this experiment, the scenario considered multiple past interactions with the system; hence, the strategy can be properly explained and evaluated. However, more details about these strategies can be found in Sect. 2.1.

Explanations

Based on the results of our first experiment, in which we did not find significant differences between the basic and the detailed explanations, we removed the detailed explanation and we only compare the *Basic explanation* with the control condition *No explanation*. For more details, refer to the description provided in Table 4.

4.2.2 Procedure

In this second experiment, we applied a mixed-subject design: each participant was presented with one of twelve possible between-subject conditions (determined by the combination of the six considered aggregation strategies and the two explanation types), and in each such condition, the participant evaluates all the four group configuration scenarios. To reduce the learning effect, the configuration scenarios are shown in random order.



(a) Uniform (b) Divergent

(c) Coalitional (d) Minority

Fig. 2 Graphical representations of the considered group configurations. The nodes represent the group members. Black solid edges represent pairs with high similarity, whereas grey dashed edges represent pairs with low similarity

After participants had agreed to an informed consent, we asked them for their gender and age. Then, we introduced them to the study:

> "In the next steps, you will be presented with four scenarios related to four different groups of people. For each of them, a software system will produce recommendations on the basis of the preferences of the group members. Please read carefully the description of each scenario, and then answer the following questions."

We asked participants to evaluate four different scenarios, one for each group configuration. Each participant was randomly assigned to one aggregation strategy and one explanation type.

For their given aggregation strategy and explanation type, participants were presented with four scenarios, which represent all four possible group configurations (see Sect. 4.2.3). To help participants discriminate among the different scenarios, each scenario was preceded by the text "**Scenario {1, 2, 3, 4} of 4**". Each scenario was then introduced by a text inspired from Tran et al. (2019) and Barile et al. (2021):

Table 4 Generic formulations for each aggregation strategy of the explanations used in this study. Let $G = \{u_1, \ldots, u_n\}$ be a group of users, and $I = \{i_1, \ldots, i_m\}$ be a set of items

Strat.	No explanation	Basic 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 among the available options."
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 θ among the available options."
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 among the available options."
MAJ	"i_k has been recommended to the group."	"i_k has been recommended to the group since most group members like it among the available options."
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 among the available options."
FAI	"i_k has been recommended to the group."	"i_k has been recommended to the group since it is the user u_j turn and this is his/her favourite choice among the available options."

Also, let $\{u_{j_1}, u_{j_2}, \ldots, u_{j_{\bar{n}}}\}$ be a subset of group members who gave a specific rating r to the item i_k selected by the considered strategy

> "Assume that there is a group of friends. 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 ten possible restaurants 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, a table with the friends' preferences for the possible choices of the considered scenario is shown. The specific table depends on the group configuration currently under evaluation.

Participants then saw a group recommendation with or without explanation, depending on which aggregation strategy and explanation type they had been assigned to (see Table 1). The recommendation was introduced by the following statement:

> "The group decided to avoid going in the same restaurant too often; hence, after a restaurant has been selected, it cannot be chosen again for the next 4 dinners. The last 3 restaurants visited are: X, Y and Z."

where X, Y, and Z are replaced with the first three choices obtained with the considered aggregation strategy (again, for the description of each specific scenario, refer to Appendix B.

The recommended restaurant was the fourth choice of the aggregation strategy. According to the type of explanation associated with the specific condition, the recommendation is presented without explanations or with the corresponding social

Table 5 Uniform group configuration

	$Rest_1$	$Rest_2$	$Rest_3$	$Rest_4$	$Rest_5$	$Rest_6$	$Rest_7$	$Rest_8$	$Rest_9$	$Rest_{10}$
Alex	4	5	4	5	4	4	3	2	1	1
Bob	1	5	1	4	3	4	3	5	1	3
Carl	2	4	2	2	5	5	2	5	1	1
David	1	4	1	2	5	5	2	2	2	1
Elle	3	5	2	3	4	5	1	3	1	2

choice-based explanation. The recommendation is introduced by the following statement:

> "Using the provided ratings, the system made a suggestion for the group on the basis of the preferences of the all the group members."

For clarity, we provide an example based on the uniform group configuration. We consider the Additive (ADD) strategy and the condition with the basic explanation. Table 5 provides the group members' preferences considered for this scenario.

When we apply the ADD strategy to this group, we obtain the following recommendation list (the complete ranking of the available options according to the considered strategy): $Rest_2$, $Rest_6$, $Rest_5$, $Rest_8$, $Rest_4$, $Rest_1$, $Rest_7$, $Rest_3$, $Rest_{10}$, $Rest_9$. Hence, the table is followed by the following messages:

> The group decided to avoid going in the same restaurant too often; hence, after a restaurant has been selected, it cannot be chosen again for the next 4 dinners. The last 3 restaurants visited are: $Rest_2$, $Rest_6$, *and* $Rest_5$.
>
> Using the provided ratings, the system made a suggestion for the group on the basis of the preferences of the all the group members
>
> *$Rest_8$ has been recommended to the group since it achieves the highest total rating among the available options.*

We then measured perceived fairness, perceived consensus, and satisfaction. We also included an attention check where we specifically instructed participants on what option to select over seven possibilities. The attention check is used to filter out some participants from the analysis. Participants further had the option to explain their answers in an open text field, introduced by the statement "*If you want, you can provide an explanation for your answers in the text below.*".

This procedure was repeated for all four group configurations (shown in a randomized order). Finally, participants had the possibility to provide general feedback on the experiment in an open text field, introduced by the statement "*If you have any further comments or feedback, please provide them in the text below.*".

At the end, a short debriefing message was shown to participants, with a brief explanation of the objectives of the experiment.

Note that, before we ran it, the experiment was approved by the *Ethics Review Committee Inner City Faculties* at Maastricht University.[9]

4.2.3 Variables

This section introduces the independent, dependent, and descriptive variables of this user study.

Independent variables

- **Group configuration** (categorical, within-subjects). Each participant was exposed to all four group configurations, namely *uniform*, *divergent*, *coalitional*, and *minority* (see Sect. 4.2.1), in randomized order (i.e., to reduce learning effects).
- **Aggregation strategy** (categorical, between-subjects). Each participant was exposed to scenarios reflecting one of the six aggregation strategies (i.e., APP, MAJ, ADD, LMS, MPL, or FAI; see Sect. 4.2.1)).
- **Explanation type** (categorical, between-subjects). Each participant saw either *no explanation* or *basic explanations* (see Sect. 4.2.1).

Dependent variables

As in our first study (see Sect. 3.2), we measured each of our three dependent variables by asking participants to rate a statement on a seven-point Likert scale ranging from "strongly agree" to "strongly disagree".

- **Perceived fairness** (ordinal): "The group recommendation is fair to all group members."
- **Consensus** (ordinal): "The group members will agree on the group recommendation."
- **Satisfaction** (ordinal): "The group members will be satisfied with regard to the group recommendation."

Descriptive variables

In addition to the independent and dependent variables that we use for hypothesis testing, we collected data on two different descriptive variables to enable a demographic description of our sample. Participants were able to select a "prefer not to say" option for these variables.

- **Age** (categorical). Participants could select one of the options *18–25*, *26–35*, *36–45*, *46–55*, *>55*.
- **Gender** (categorical). Participants could select one of the options *female*, *male*, or *other*.

[9] https://www.maastrichtuniversity.nl/ethical-review-committee-inner-city-faculties-ercic.

4.2.4 Sample size determination

As we did for the first user study, we computed the required sample size for our experiment before we performed the data collection. We performed a power analysis for a factorial mixed ANOVA (see Sect. 4.2.6) using the software *G*Power* (Faul et al. 2007). Here, we specified the default effect size f = 0.25, a significance threshold $\alpha = \frac{0.05}{15} = 0.003$ (due to testing multiple hypotheses), a power of $(1 - \beta) = 0.8$, and that we test $6 \times 2 = 12$ groups (i.e., 6 different aggregation strategies for 2 different explanation scenarios), and 4 repeated measures for each group (the 4 group configurations). This results in a sample size of 288 participants (24 participants for each of the twelve groups).

4.2.5 Participants

We recruited 388 participants from the online participant pool *Prolific*,[10] all of whom were proficient English speakers above 18 years of age. Each participant was allowed to participate in our study only once and received £0.70 as a reward for participation. We excluded from our analysis participants who did not pass all four attention checks in our experiment (57 participants). Furthermore, we excluded participants who completed the questionnaire in a time which was considered too fast: based on a series of beta tests of the questionnaire performed before running the real study in which the participants spent between three and seven minutes. Thus, we decided to use the threshold of three minutes for filtering out participants contributions. Hence, we removed 43 participants and considered the remaining 288 participants in our analysis.

The resulting sample was composed of 45.4% (131) female, 51.7% (149) male, 2% (6) other participants, while 2 participants did not specify their gender. Regarding the age groups, 40.6% (117) were between 18 and 25, 32.2% (93) between 26 and 35, 12.1% (35) between 36 and 45, 9% (26) between 46 and 55, and 4.5% (13) were above 55 years of age, while 1 participant preferred to not specify the age group. Additional information on the dataset demographic distributions are available in Appendix A. We randomly distributed participants over the 12 between-subject conditions (i.e., exposing them to 1 out of 6 aggregation strategies and 1 out of 2 explanation types).

4.2.6 Statistical analysis

For each of the three dependent variables in our study (i.e., *fairness perception*, *consensus perception*, and *satisfaction*), we conducted a factorial mixed ANOVA using the *aggregation strategy*, and *explanation type* as between-subjects factors, and the *group configuration* as a within-subjects factor. These three factorial mixed ANOVAs were used to test a total of 15 hypotheses (i.e., **H2.1a – H2.4f**). Specifically, each of them tested for main effects of *aggregation strategy* (**H2.1a – H2.1c**), *explanation type* (**H2.2a – H2.2c**), *group configuration* (**H2.3a – H2.3c**), as well as the interaction between these three variables in affecting the dependent variables (**H2.4a – H2.4f**).

[10] https://prolific.co.

Because we tested 15 different hypotheses, we did not handle the typical significance threshold of 0.05. Applying a Bonferroni correction (Napierala 2012), we lowered the significance threshold to $\alpha = \frac{0.05}{15} = 0.003$ (rounded to three digits after the decimal point).

Since we found significant main effects related to six sets of hypotheses (**H2.1a** – **H2.1c** and **H2.3a** – **H2.3c**), we conducted post hoc analyses to investigate specific differences between the analyzed groups. More specifically, we conducted Tukey post hoc analyses to investigate specific differences between aggregation strategies and group configurations.

4.3 Results

Table 6 shows the results of the statistical analyses outlined in Sect. 4.2.6. Below, we report some descriptive statistics about the collected data and describe the results related to the research questions RQ2.1–4.

Descriptive statistics

For each of the 12 between-subjects conditions (i.e., the combinations of the six aggregation strategies and the two explanation types), we collected evaluations from 24 participants. On average, participants spent 450 s on the task (the median value is 359 s). We recall here that we removed all participants who finished the task in less than three minutes. Overall, 60% of the participants somewhat agreed with the fairness statement, 62% agreed with the consensus statement, and 59% agreed with the satisfaction statement. These percentages are higher than in the previous study (see Sect. 3.3).

RQ2.1: differences between social choice-based aggregation strategies regarding recommendation effectiveness in complex group recommendation scenarios. We found significant differences between the six aggregation strategies concerning all three dependent variables, namely *fairness perception*, *consensus perception*, and *satisfaction* (**H2.1a** – **H2.1c**; $F = [6.363, 8.385, 8.746]$, all $p < 0.001$; see Table 6). Hence, in general, participants expressed different agreement levels for the three variables based on which aggregation strategy they were exposed to.

We conducted a Tukey pairwise post hoc analysis to investigate specific differences between the aggregation strategies. We found significant differences for all the dependent variables. Regarding the *fairness perception*, the Approval Voting (APP) was evaluated as less fair than the Majority (MAJ), Additive (ADD), and Fairness (FAI) strategies (all $p_{adj} < 0.005$). Furthermore, the Most Pleasure (MPL) was found to be less fair than ADD and FAI (all $p_{adj} < 0.005$). Finally, the Least Misery (LMS) was found to be less fair than the FAI strategy. The same differences were also found for the *consensus perception*, for which, in addition, we found that LMS was evaluated as worse than ADD ($p_{adj} < 0.005$). Finally, regarding *satisfaction*, all the pairwise differences highlighted for the *fairness perception* and *consensus perception* were confirmed, together with an additional significant difference: MPL was considered to

be less satisfying than MAJ ($p_{adj} < 0.005$). As we can see, these results are not in line with the results of the previous study. We will discuss these differences in Sect. 5.

RQ2.2: impact of the presence of an explanation in complex group recommendation scenarios. We found no significant differences between the two explanation conditions (i.e., the condition with a *basic* explanation and the control condition without explanations) regarding all three dependent variables (**H2.2a** – **H2.2c**; $F = [0.272, 0.000, 0.218]$, $p = [0.603, 1.000, 0.640]$; see Table 3). So, in line with the results of our previous study, our results contain no evidence of an impact of the explanations concerning our three dependent variables, regardless of the increased complexity of the analyzed scenarios.

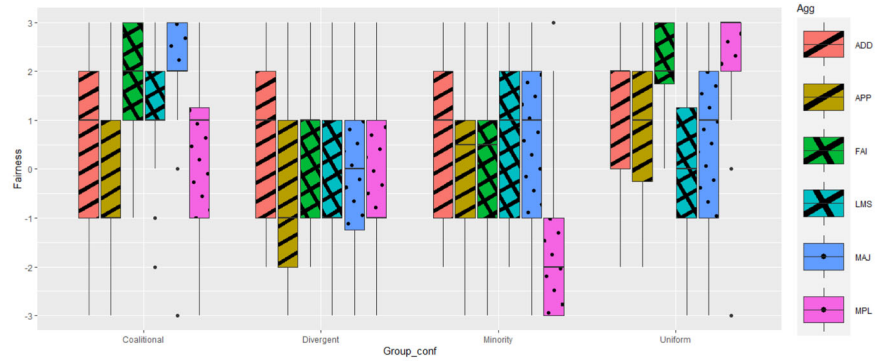
RQ2.3: differences between the group configurations regarding recommendation effectiveness in complex group recommendation scenarios. We found significant differences between the four group configurations for all the three dependent variables *fairness perception*, *consensus perception*, and *satisfaction* (**H2.3a** – **H2.3c**; $F = [67.179, 62.888, 67.418]$, all $p < 0.001$; see Table 6). In general, participants expressed different levels regarding the three variables based on the group configuration they were exposed to. The Tukey pairwise post hoc analysis showed that for all the three dependent variables, the recommendations provided for the *divergent* configuration received significantly lower evaluations than the ones provided for the *coalitional* and the *uniform* configurations (all $p_{adj} < 0.001$). Similarly, the recommendations provided for the *minority* configuration received significantly lower evaluations than the ones provided for the *coalitional* and the *uniform* configurations (all $p_{adj} < 0.001$).

RQ2.4: interactions between aggregation strategies, group configurations, and explanation types regarding recommendations effectiveness. There were no significant 3-way interaction effects between the six aggregation strategies, the four group configurations, and the explanations (**H2.4a** – **H2.4f**; $F = [0.540, 1.307, 1.115]$, $p = [0.918, 0.190, 0.337]$; see Table 6). We also found no interaction effects between the six aggregation strategies and the explanations types (**H2.4a** – **H2.4c**; $F = [0.926, 1.156, 0.809]$, $p = [0.465, 0.331, 0.544]$; see Table 6), or between the four group configurations and the explanations types (**H2.4d** – **H2.4f**; $F = [4.283, 0.408, 0.859]$, $p = [0.005, 0.747, 0.462]$; see Table 6). However, we found significant interaction effects between aggregation strategies and explanation types for all the measured dependent variables ($F = [20.109, 20.478, 24.640]$, all $p < 0.001$; see Table 6). Hence, we found no impact of the explanations on the users' *fairness perception*, *consensus perception*, and *satisfaction*. However, the significant interaction effects between aggregation strategies and group configurations suggest different performances for each aggregation strategy according to each specific group configuration. To investigate specific differences between the aggregation strategies for each group configuration, we conducted a Tukey pairwise post hoc analysis, grouping the observations by group configuration. We found several significant differences, which are reported in Table 7.

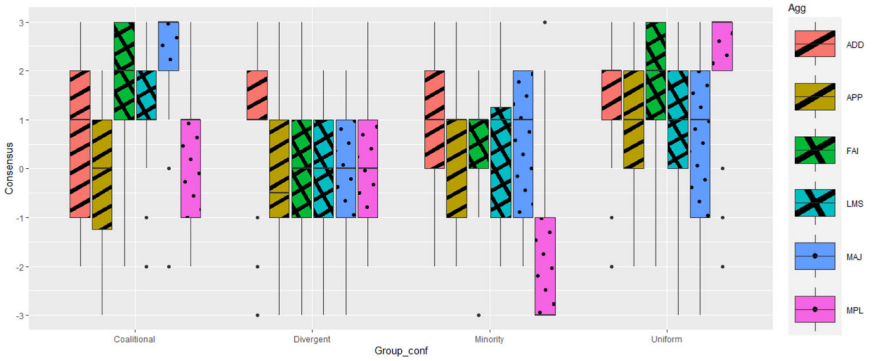
Table 6 Results of three mixed ANOVAs for the dependent variables (DVs) *fairness perception* (left), *consensus perception* (center), and *satisfaction* (right). Per effect, we report the F-statistic and p-value

DV: fairness perception			DV: consensus perception			DV: satisfaction		
	F	p		F	p		F	p
(H2.1a) aggr	6.363	< 0.001*	**(H2.1b)** aggr	8.385	< 0.001*	**(H2.1c)** aggr	8.746	< 0.001*
(H2.2a) expl	0.272	0.603	**(H2.2b)** expl	0.000	1.000	**(H2.2c)** expl	0.218	0.640
(H2.3a) gr_conf	67.179	< 0.001*	**(H2.3b)** gr_conf	62.888	< 0.001*	**(H2.3c)** gr_conf	67.418	< 0.001*
(H2.4a) aggr:expl	0.926	0.4647	**(H2.4b)** aggr:expl	1.156	0.3314	**(H2.4c)** aggr:expl	0.809	0.544
aggr:gr_conf	20.109	< 0.001*	aggr:gr_conf	20.478	< 0.001*	aggr:gr_conf	24.640	< 0.001*
(H2.4d) expl:gr_conf	4.283	0.005	**(H2.4e)** expl:gr_conf	0.408	0.747	**(H2.4f)** expl:gr_conf	0.859	0.462
(H2.4ad) aggr:expl:gr_conf	0.540	0.918	**(H2.4be)** aggr:expl:gr_conf	1.307	0.190	**(H2.4cf)** aggr:expl:gr_conf	1.115	0.337

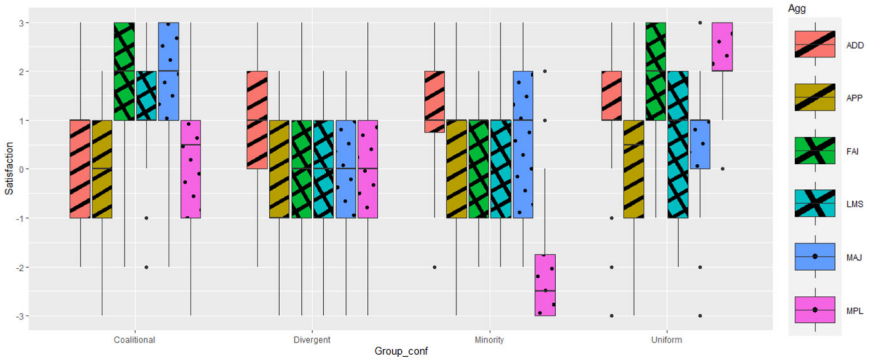
The terms “aggr”, “expl” and “gr_conf” represent the independent variables *aggregation strategy*, *explanation type*, and *group configuration*. Colons indicate interaction effects, asterisks statistical significance



(a) Fairness perception



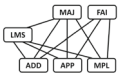
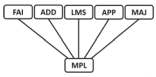
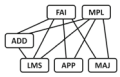
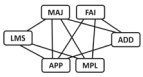
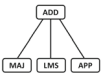
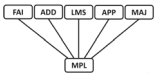
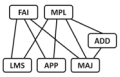
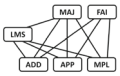
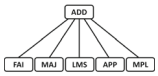
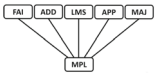
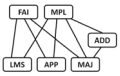
(b) Consensus perception



(c) Satisfaction

Fig. 3 Distribution of user's evaluations for *fairness perception*, *consensus perception*, and *satisfaction*, on scales from −3 ("strongly disagree") to 3 ("strongly agree"; see Sect. 4.2.3), for each group configuration and aggregation strategy

Table 7 Significant differences between social choice-based aggregation strategies for each group configuration

Dependent Variable	Coalitional	Divergent	Minority	Uniform
Fairness Perception				
Consensus Perception				
Satisfaction				

We summarize the results as a partial ordering for fairness perception, consensus perception, and satisfaction for the different aggregation strategies. For example, we see that Most Pleasure (MPL) consistently performs worse for the Minority configuration but better than many strategies for the Uniform configuration

4.4 Discussion

In the second study on COMPLEX RECOMMENDATION SCENARIOS, we studied the differences in effects between aggregation strategies when applied to larger groups and more candidate items. In the following section, we discuss the implications of the results of this second study, focusing on the differences between the defined group configurations and the interaction with different aggregation strategies.

4.4.1 The differences between aggregation strategies

As in EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES (c.f., Sect. 3), the results presented in Sect. 4.3 showed significant differences between the aggregation strategies in terms of perceived fairness, perceived consensus, and satisfaction.

Regarding fairness perception, the post hoc Tukey analysis showed that MPL obtained lower scores than the FAI and ADD strategies. Furthermore, APP obtained lower scores than the MAJ, FAI, and ADD strategies, and LMS is evaluated worse than FAI. The same occurred for consensus perception, for which, additionally, LMS obtained lower scores than ADD, and for satisfaction, for which MPL was also evaluated lower than MAJ. As we can see, some of these differences are partially coherent with the results of our first study EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES, where MPL was perceived as the worst strategy regarding all the considered dependent variables. Other differences are not in line with the previous findings, as we cannot confirm that MPL performs significantly worse than LMS, APP, and MAJ. Furthermore, additional differences were found regarding the APP and LMS strategies. One plausible reason for this difference is the interaction with the considered group configuration on the effectiveness of the aggregation strategies. As we show in Sect. 4.4.3, some aggregation strategies perform better when applied

to some group configurations and worse for others; these differences are not captured when comparing the strategies in general.

4.4.2 The differences between group configuration

In Sect. 4.3, we found significant differences between the group configurations regarding perceived fairness, perceived consensus, and satisfaction. For all the considered variables, the effectiveness of the strategies for the *Minority* and *Divergent* configurations is lower than for the *Coalitional* and *Uniform* configurations. No significant differences were found between the *Divergent* and *Minority* configurations and between the *Uniform* and *Coalitional* configurations. This result is intuitive in the sense that both the *Minority* and *Divergent* configurations present relatively difficult scenarios, in which it is harder to determine a recommendation that can satisfy all of the group members. On the contrary, in the *Coalitional* and *Uniform* configurations, it is easier for the recommender to select an item that satisfies most of the group members, without disregarding the preferences of some group members.

4.4.3 The impact of the group configuration on the effectiveness of the aggregation strategies

The significant interaction effects between aggregation strategies and group configurations shown in Sect. 4.3 suggest that the performances of the aggregation strategies may be affected by the specific configuration of the group to which the aggregation strategy is applied. To explore these differences, we performed grouped Tukey pairwise comparisons, which are reported in Table 7. We also visually compare the differences between the aggregation strategies in Fig. 3 and observe that the results vary according to the specific group configuration.

For the *Minority* configuration, we can see significantly lower performances for the Most Pleasure (MPL) strategy compared to the other aggregation strategies. This is in line with the results we found in our first study, EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES, where MPL performed the worst. To further investigate the relationship between the findings in the two studies, we take a closer look at the scenario used for the first experiment (see Table 2 in Sect. 3.2.2). Indeed, the scenario can be considered a minority group configuration: we have three group members with a general agreement, while the group member Alex may be considered in a minority position. These results suggest that for a group characterized by a *Minority* configuration, the MPL strategy should be avoided. It also seems like the Additive (ADD) strategy is the overall best-performing strategy for this configuration, as it ensures good performances in terms of satisfaction and consensus perception. However, this is not confirmed by significant differences between ADD and the other strategies.

If we focus on the *Divergent* group, the only remarkable observation is that the ADD strategy again performs the best. In particular, ADD is significantly better than LMS in terms of fairness perception, better than MAJ, APP, and LMS in terms of consensus perception, and better than all the other strategies when considering satisfaction. For this configuration, ADD performs well on all three dependent variables. Jointly, the results for the Minority and Divergent groups suggest that the ADD strategy could

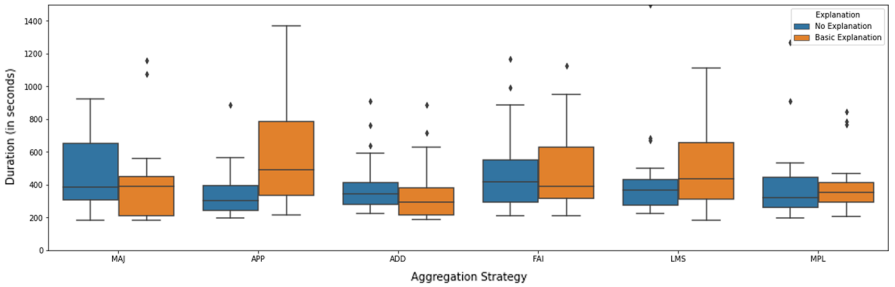


Fig. 4 The differences in time (seconds) to complete each session. This was computed across 4 group configurations. Comparison of the different aggregation strategies (MAJ, APP, ADD, FAI, LMS, MPL) and compared with and without explanations

be the better aggregation strategy to implement when the group was presented with a "hard" configuration, with very different individual preferences, or a user in a minority position which tends to be unhappy with most of the items selected for the other group members. This is an exciting area for future work.

The best strategy to use seems to be different for the remaining group configurations. For both the *Coalitional* and *Uniform* configurations, we can notice good performances for all strategies (in Fig. 3, we can notice that the average values are mostly above zero, which was the "neutral" option in the Likert-scale used for the questionnaire). However, for the *Coalitional* configuration, FAI and MAJ result in higher performances for all the considered dependent variables, and after these, LMS also obtains better evaluations than APP, MPL, and ADD. This result, however, is not surprising since applying the strategy multiple times (as in our scenario) could allow satisfying one or the other coalition roughly in a balanced way. On the contrary, for the *Uniform* group configuration, we have the best evaluations when applying the FAI and MPL strategies. This can be motivated by the fact that when the group has similar preferences, it is natural to assume that the most satisfying item for one of the group members is also good for all the others. To summarize, the fairness-based strategy appears to be a good default when users have more similar preferences. In addition, the MPL strategy seems a good choice for the *Uniform* configuration, in contrast to the more "difficult" group compositions.

4.4.4 The ineffectiveness of social choice-based explanations

Similar to our first study, we find no significant effects of explanations on fairness perception, consensus perception, and satisfaction. Furthermore, we found no interaction effects between explanations, aggregation strategies, and group configurations. This seems to suggest that even when the scenario to evaluate is more complex, the participants experienced little to no benefit from the presence of the explanations, in terms of perceived fairness, perceived consensus, and satisfaction, regarding the provided recommendations.

We also analyzed the time the participants spent on the task to check whether there were other indications of an impact of the presence of the explanation. Figure 4

reports the duration in seconds of the experiment, grouped by aggregation strategy and explanation type.[11]

In sum, explanations appear to decrease the duration notably for MAJ and slightly for ADD, but increase the duration for other aggregation strategies such as APP and LMS. We note that requiring more time is not necessarily negative. Suppose an aggregation strategy is counter-intuitive to a user's initial representation of the group recommendations. In that case, explanations can be useful in improving the understandability of the system, which can improve user confidence and trust in the system. Therefore, it becomes crucial to evaluate these aspects in future studies and make specific measurements for different group configurations.

5 General discussion

In this section, we compare the results in EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES (c.f., Sect. 3) with COMPLEX RECOMMENDATION SCENARIOS (c.f., Sect. 4), as well as with the results of the user study from Tran et al. (2019). We look at the differences between social choice-based aggregation strategies regarding users' perceived fairness, perceived consensus, and satisfaction. Table 8 summarizes the results of these studies graphically.

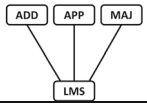
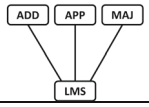
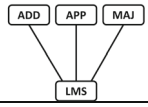
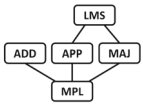
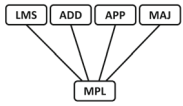
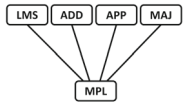
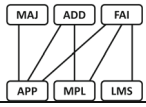
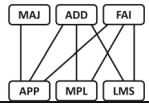
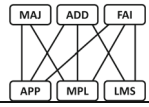
In the following section, we first critically discuss the effectiveness of different social-based aggregation strategies for a single scenario, in light of the findings of all three studies. We then discuss the importance of the group configuration in more detail. We conclude the discussion with an analysis of why no significant results were found for the benefit of explaining these strategies.

5.1 The effectiveness of social choice-based aggregation strategies

In Table 8, we observe that there are significant differences between aggregation strategies in all three experiments. However, the findings are not consistent among the three studies. In the study conducted by Tran et al. (2019), Least Misery (LMS) was consistently performing *worse* than Additive (ADD), Approval Voting (APP), and Majority (MAJ). In contrast, in our first study, EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES, LMS was found to perform the *best* in terms of the same variables. We also found Most Pleasure (MPL) to be the worst strategy in terms of all the measured dependent variables. As mentioned in Sect. 3.4.1, the differences with the results in Tran et al. (2019) might be explained by differences in the methodology: in their study, the aggregation strategies were used as a within-subject factor. While we can only speculate on the nature of the effect, we could expect that learning effects or indirect comparison in those studies could have somehow resulted in LMS being the less preferred aggregation strategy (e.g., learning that there are strategies that result in happier group members compared to this strategy). This also indicates that

[11] Note that each participant evaluated all four group configurations for a specific combination of one aggregation strategy and one explanation at a time, and we only recorded the total duration of the session, so we cannot analyze the differences in duration between group configurations.

Table 8 Summary of the results of the three studies: the original study by Tran et al. (2019), the first replication (c.f., Sect. 3), and the second study in a more complex scenario (c.f., Sect. 4)

Experiment	Fairness	Consensus	Satisfaction
Tran et al (2019)			
EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES [12] (Barile et al, 2021)			
COMPLEX RECOMMENDATION SCENARIOS			

We summarize the results as a partial ordering for fairness perception, consensus perception, and, respectively, satisfaction, for the different aggregation strategies. For example, we see that Least Misery (LMS) consistently performs worse in the study by Tran et al. (2019), but better than many strategies in EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES. In contrast, Most Pleasure (MPL) performs consistently worse than any other strategy in both experiments presented in this paper

further research is required to study what would happen in a system where different aggregation strategies are consecutively applied.

We then compare the results of Tran et al. (2019) with our second study, COMPLEX RECOMMENDATION SCENARIOS. Here, we found that both MPL and LMS were outperformed by other strategies such as Fairness (FAI),[12] Additive (ADD), and Majority (MAJ).

In other words, when comparing our two studies, we see that while LMS performed best in the first study, it performed much worse when considering different group configurations in the second study. Specifically, in Sect. 4.4.3, we saw that the effectiveness of each aggregation strategy depends on the group configuration on which the strategy is applied. We also observed that the scenario considered in EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES was effectively a minority group configuration. Considering this, the results we obtained for the minority configuration in COMPLEX RECOMMENDATION SCENARIOS seem coherent with the first study's results, as ADD, APP, FAI, LMS, and MAJ all obtain better performances. In sum, our results suggest carefully considering the effects of the recommendation scenario when designing a group recommendation user study, as this may influence the results.

[12] Note that in our first study, the Fairness (FAI) strategy was not included, so no comparison is possible for this strategy among our two experiments. However, Tran et al. (2019) included it, but they did not find any significant difference between FAI and the other strategies.

5.2 The importance of the group configuration

As we illustrated in Sect. 2, previous work has proposed adapting the aggregation strategy to the characteristics of the specific group (Gartrell et al. 2010). However, those approaches define the group based on the strength of the social relationship between the group members. Furthermore, the association between the aggregation strategy and the specific group is typically based on anecdotal observations from a limited number of groups. Our work not only suggests that an important factor to characterize a group is the internal (dis-)agreement, but also provides clear definitions of four group configurations that can be used to categorize real groups, and indications on which aggregation strategies are better for each specific group configuration.

In Sect. 4.4.3, we described the impact of the group configuration on the effectiveness of the aggregation strategies in COMPLEX RECOMMENDATION SCENARIOS, and we summarized the results in Table 7. From these findings, we draw the following guidelines. The Most Pleasure (MPL) strategy performs poorly and should be avoided for a minority group configuration. In contrast, MPL is the preferable option for a uniform group. Furthermore, the Fairness (FAI) strategy may be used for uniform and coalitional groups.[13] Finally, the Additive (ADD) strategy may be used in situations where the group configuration is not clearly identifiable, as it is among the best for the more critical configurations (divergent and minority). ADD also obtains good results for the uniform configuration. However, since the knowledge of the individual preferences (or individual predictions/recommendations) is necessary to apply any aggregation strategy, it is, in principle, possible to determine the group configuration for any group recommendation scenario.

These findings also have implications for generating explanations. If the aggregation strategy used is adapted to the group configuration, it is also reasonable to use this information in the explanation. For instance, for the Most Pleasure (MPL) strategy and a uniform group, an explanation could be:

> "Considering that the group members have similar preferences, the system recommended the item i_k as it achieves the highest of all individual group members' ratings".

Graphical representations of the group, as illustrated in Fig. 2, could also be used to accompany the explanations.

5.3 Limitations and their impact on the effectiveness of social choice-based explanations

Both our studies showed no benefits in using social choice-based explanations for group recommendations. However, further investigations are necessary before concluding that such explanations have no use. Here, we discuss some of the main limitations of our approach, which may have determined this result.

[13] We also note here that while FAI outperforms many of the strategies in COMPLEX RECOMMENDATION SCENARIOS for all configurations, we cannot conclude that this result is replicable since this aggregation strategy was not evaluated in EVALUATING THE EFFECTIVENESS OF SOCIAL CHOICE STRATEGIES.

One limitation of our work regards how we present the items to the participants. To avoid influencing participants' decisions, we did not provide 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. This may affect the effectiveness of explanations, as the restaurants' anonymization directly impacts the provided explanation, making the interaction with the system less realistic.

Another limitation of our study is that all recommendations are in the restaurant domain. Different recommendation domains could be perceived differently 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 (Tintarev and Masthoff 2008); the restaurant domain is generally perceived as a medium-low investment compared to other domains suitable for group recommendations, such as tourism. In such a domain, the user's perception of the risk of not making the best decision is lower, negatively impacting the user's need for explanations. It is possible that in a high-investment domain, such as the tourism domain, the explanations may be more effective.

Another important factor to consider is that recommendations and explanations are not evaluated by group members. As previously mentioned, in line with the evaluation approach in Tran et al. (2019), and also to other studies in the literature (Masthoff and Gatt 2006; Masthoff and Delić 2022), 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 that their evaluations in relation to the explanations could be different when part of the group, especially when the system is providing a recommendation that is not the best for the user. 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.

Furthermore, we should consider that 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, however, provides several directions for measuring recommendation understandability, which could be investigated in future work (Knijnenburg et al. 2011; Gedikli et al. 2014; Wang and Yin 2021). In Sect. 4.4.4, we did see that explanations decreased the duration for some strategies (Majority) and increased the duration for others (Approval Voting and Least Misery). Still, the reason for the differences in duration is not known, i.e., it is unclear whether the increased processing time was due to correcting participant expectations or unnecessary complexity. There were no user comments to indicate either, however.

6 Conclusion and next steps

Social choice aggregation strategies have been proposed as an explainable way to make recommendations to groups. However, few studies have empirically and systematically

evaluated how the distribution of preferences in a group influences which strategy is most effective.

To this end, we present two user studies investigating the effectiveness of these strategies in terms of users' fairness perception, consensus perception, and satisfaction. We investigate the impact of the level of (dis-)agreement within the group on the performance of the social-choice aggregation strategies. We call this the *"group configuration"* and define it based on the similarity between group members' individual preferences. Furthermore, given that the social choice strategies are highly explainable, we also explore the added value of explanations. These are presented as sentences explaining the aggregation strategy used to produce the recommendation to the group as a whole.

We find significant differences in the effectiveness of the social choice-based aggregation strategies in both studies. Furthermore, the most effective strategy appears to *depend on the specific group configuration*. In particular, the Most Pleasure (MPL) strategy should be avoided for a minority group configuration, while it is the preferable option for a uniform group. Furthermore, the Fairness (FAI) strategy may be used for uniform and coalitional groups. Finally, the Additive (ADD) strategy may be used when the group configuration is not clearly identifiable. To our surprise, we did not find much added value in accompanying the aggregation strategies with social choice-based explanations (in neither of the two studies). We did, however, see that explanations decreased the duration for some strategies (Majority) and increased the duration for others (Approval Voting and Least Misery).

Our findings emphasize the importance of considering the group configuration when selecting and analyzing the benefit of different aggregation strategies. This is a substantial step in understanding when aggregation strategies benefit group decision-making. In our next steps, we plan to study the dynamics of group decision-making, including supporting discussions among group members (c.f. our work using a chatbot in Najafian et al. (2021b)). It improves the ecological validity of people chatting together about potential recommendations while allowing us to control the flow of information by suggesting gradual revealing of information to users. Finally, we plan to validate our findings in more complex user studies involving real groups. Several works in the literature (Delic et al. 2018; Herzog and Wörndl 2019; Rossi et al. 2015) presented user studies involving real groups, which are observed during the decision-making process, and asked to evaluate recommendations provided by group recommender systems. A similar approach will be used to test if an adaptive recommender system, which decides the best strategy to use according to the detected group configuration, leads to better performance than a fixed aggregation strategy.

Table 9 Distribution of the participants in the two user studies over the two collected demographic dimensions: Gender and Age group

	Dataset 1		Dataset 2	
	Perc	Num	Perc	Num
Gender				
Male	38.3%	153	51.7%	149
Female	61.0%	244	45.4%	131
Others	0.5%	2	2.0%	6
Prefer not to say	0.0%	0	0.6%	2
Age group				
18–25	27.6%	110	40.6%	117
26–35	28.8%	115	32.2%	93
36–45	17.0%	68	12.1%	35
46–55	13.8%	55	9.0%	26
>55	12.8%	51	4.5%	13
Prefer not to say	0.0%	0	0.3%	1

Appendix A Datasets

The datasets collected from the two user studies have been published and are available online. The first dataset, collected for the study EVALUATING THE EFFECTIVENESS OF EXPLAINABLE SOCIAL CHOICE- BASED AGGREGATIONS (see Sect. 3), is available at https://osf.io/5xbgf/. The second dataset, collected for the study THE IMPACT OF SCENARIO COMPLEXITY (see Sect. 4), is available at https://doi.org/10.34894/8EVX4U.

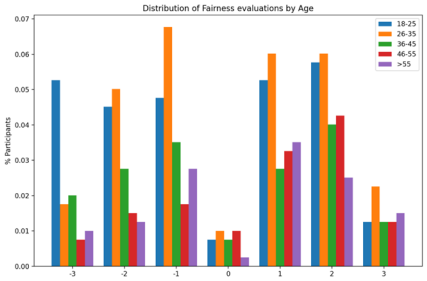
Table 9 illustrates the distribution of the participants over the two collected demographic information: Gender and Age group. Figures ?? and ?? show the distributions of the collected dependent variables (Fairness perception, Consensus perception, and Satisfaction) across the different categories of Age group and Gender in both datasets.

Figures 5, 6, 7, and 8 are provided to further investigate the distribution of the evaluations provided by the participants in the different age groups and self-reported gender groups. For each dependent variable (Fairness perception, Consensus perception, or Satisfaction), we produce a distribution chart and a normalized distribution chart. Note that the normalization is done at group level, to allow comparing groups of different sizes. In these charts, the groups having less than 10 participants are omitted for the sake of clarity. The charts suggest similar distributions across the different groups for both the variables and in both the datasets.

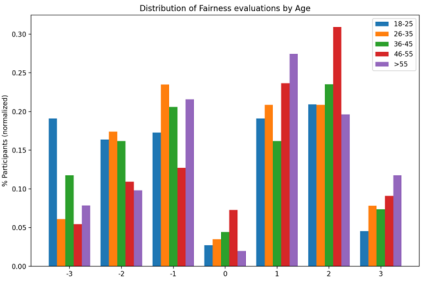
Appendix B Group configurations generation

Appendix B.1 Social graph and groups of interest

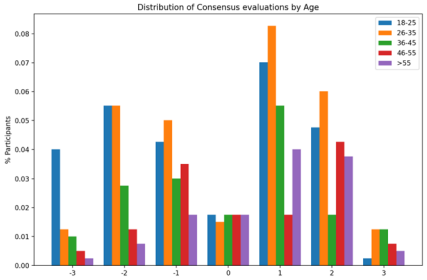
Let $U = \{u_1, \ldots, u_N\}$ be the set of users belonging to a given group. We define the "Group Social Graph" as the graph $G = (U, E)$, where $E = U \times U$. Furthermore,



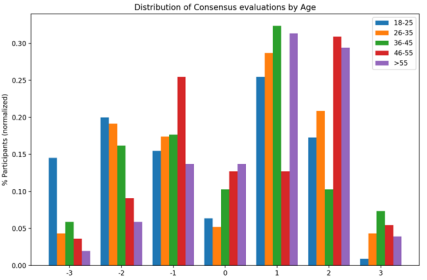
(a) Distribution of the evaluations for the Fairness perception across the different age groups.



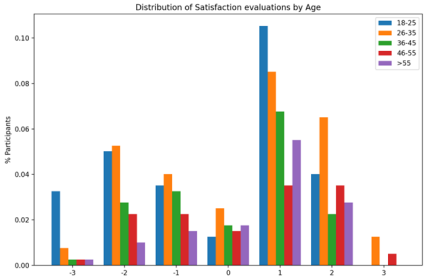
(b) Distribution of the evaluations for the Fairness perception across the different age groups, normalized at group level. The groups with less than 10 participants are omitted.



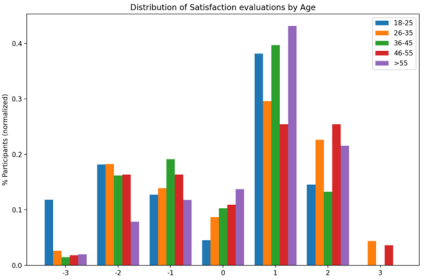
(c) Distribution of the evaluations for the Consensus perception across the different age groups.



(d) Distribution of the evaluations for the Consensus perception across the different age groups, normalized at group level. The groups with less than 10 participants are omitted.

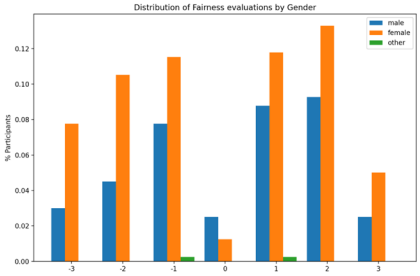


(e) Distribution of the evaluations for the Satisfaction across the different age groups.

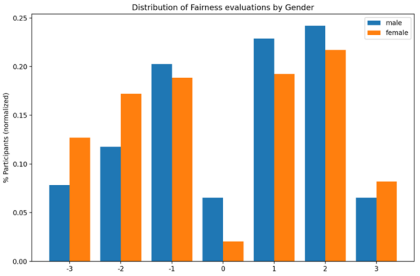


(f) Distribution of the evaluations for the Satisfaction across the different age groups, normalized at group level. The groups with less than 10 participants are omitted.

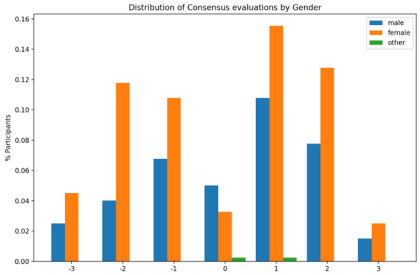
Fig. 5 Distributions of the evaluations for the dependent variables (fairness perception, consensus perception, and satisfaction) across the different age groups, in the dataset 1. In the normalized charts (b, d, and f) the group with less than 10 participants are omitted



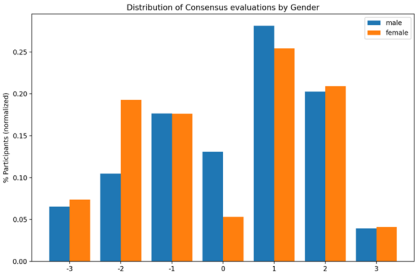
(a) Distribution of the evaluations for the Fairness perception across the different self-reported gender groups.



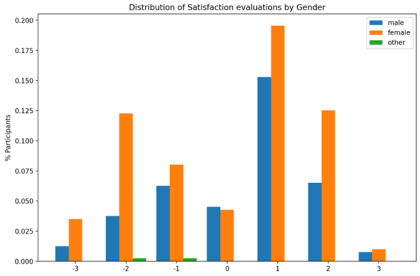
(b) Distribution of the evaluations for the Fairness perception across the different self-reported gender groups, normalized at group level. The groups with less than 10 participants are omitted.



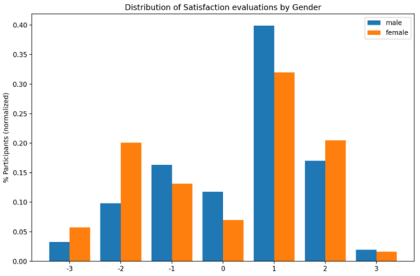
(c) Distribution of the evaluations for the Consensus perception across the different self-reported gender groups.



(d) Distribution of the evaluations for the Consensus perception across the different self-reported gender groups, normalized at group level. The groups with less than 10 participants are omitted.

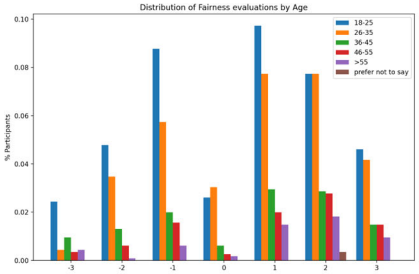


(e) Distribution of the evaluations for the Satisfaction across the different self-reported gender groups.

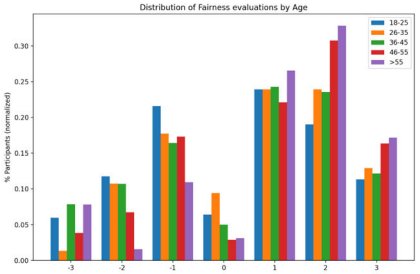


(f) Distribution of the evaluations for the Satisfaction across the different self-reported gender groups, normalized at group level. The groups with less than 10 participants are omitted.

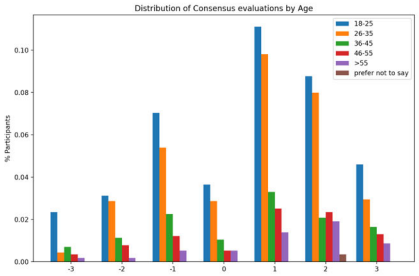
Fig. 6 Distributions of the evaluations for the dependent variables (fairness perception, consensus perception, and satisfaction) across the different self-reported gender groups, in the dataset 1. In the normalized charts (b, d, and f) the group with less than 10 participants are omitted



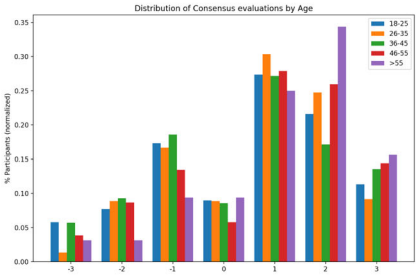
(a) Distribution of the evaluations for the Fairness perception across the different age groups.



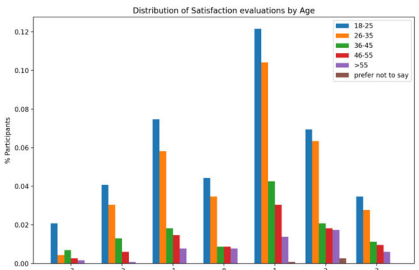
(b) Distribution of the evaluations for the Fairness perception across the different age groups, normalized at group level. The groups with less than 10 participants are omitted.



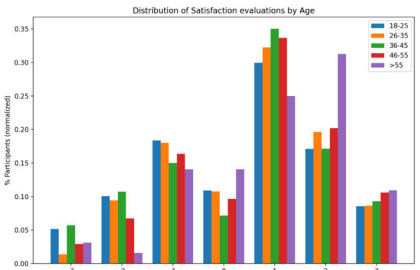
(c) Distribution of the evaluations for the Consensus perception across the different age groups.



(d) Distribution of the evaluations for the Consensus perception across the different age groups, normalized at group level. The groups with less than 10 participants are omitted.

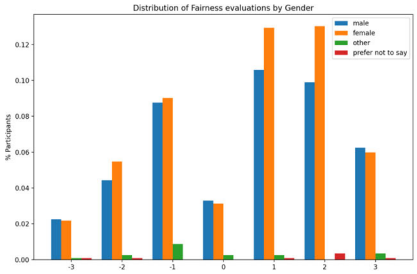


(e) Distribution of the evaluations for the Satisfaction across the different age groups.

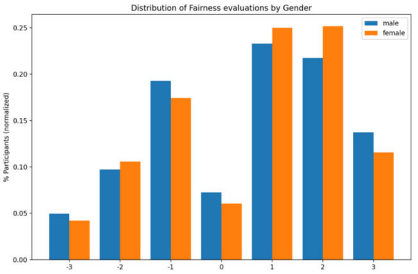


(f) Distribution of the evaluations for the Satisfaction across the different age groups, normalized at group level. The groups with less than 10 participants are omitted.

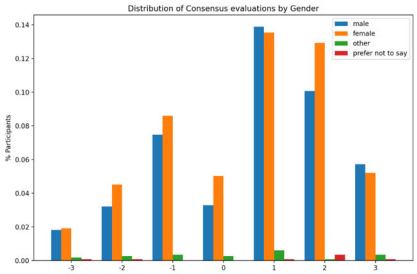
Fig. 7 Distributions of the evaluations for the dependent variables (fairness perception, consensus perception, and satisfaction) across the different age groups, in the dataset 2. In the normalized charts (b, d, and f) the group with less than 10 participants are omitted



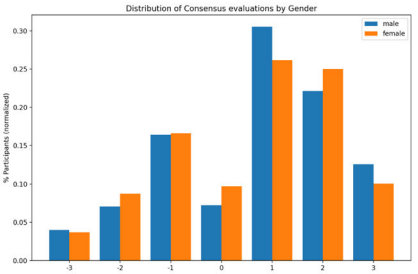
(a) Distribution of the evaluations for the Fairness perception across the different self-reported gender groups.



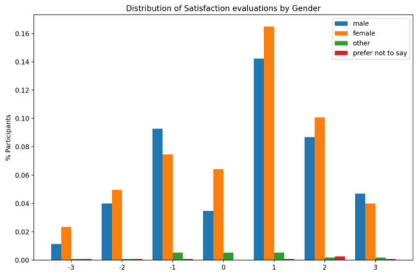
(b) Distribution of the evaluations for the Fairness perception across the different self-reported gender groups, normalized at group level. The groups with less than 10 participants are omitted.



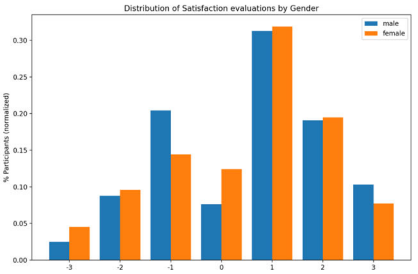
(c) Distribution of the evaluations for the Consensus perception across the different self-reported gender groups.



(d) Distribution of the evaluations for the Consensus perception across the different self-reported gender groups, normalized at group level. The groups with less than 10 participants are omitted.



(e) Distribution of the evaluations for the Satisfaction across the different self-reported gender groups.



(f) Distribution of the evaluations for the Satisfaction across the different self-reported gender groups, normalized at group level. The groups with less than 10 participants are omitted.

Fig. 8 Distributions of the evaluations for the dependent variables (fairness perception, consensus perception, and satisfaction) across the different self-reported gender groups, in the dataset 2. In the normalized charts (b, d, and f) the group with less than 10 participants are omitted

given a subgroup $C \subseteq U$, we define the graph $G[C]$ as the induced subgraph of G, hence $G[C] = (C, E[C])$ where $E[C] = \{(u_i, u_j) \in E \mid u_i \in C \wedge u_j \in C\}$.

Let $sim(u_i, u_j) \in [0, 1]$ be a similarity metric between users u_i and u_j. We assume that 0 indicates the lower similarity, while 1 indicates a perfect similarity. We associate to each edge $(u_i, u_j) \in E$ the similarity $sim(u_i, u_j)$. The similarity measure must be such that $\forall u_i, u_j \in U$:

- $sim(u_i, u_j) = sim(u_j, u_i)$
- $sim(u_i, u_i) = 1$

We aim to identify groups of interest using the sim similarity measure.

Definition 1 (Potential group of interest) A potential group of interest is a group U for which there is a partition $C_1, \ldots, C_P$ with high inter-group similarity and a low intra-group similarity. More formally:

- $\forall C_k, u_i, u_j \in C_k \implies sim(u_i, u_j) \geq \theta_1$
- $\forall C_k, C_l, u_i \in C_k \wedge u_j \in C_l \implies sim(u_i, u_j) \leq \theta_2$

Among all the possible partitions, we focus on four groups of interest:

1. Uniform: this group is defined by the trivial partition $\{U\}$. In this case, the group members' similarity is greater than the threshold θ_1. This defines a uniform group in which all group members are similar.
2. Divergent: this group is defined by the partition $\{\{u_1\}, \ldots, \{u_N\}\}$. In this case, the group members' similarity is lower than the threshold θ_2. This defines a divergent group in which all group members are dissimilar.
3. Minority: this group is defined by a partition $\{C_1, C_2\}$ such that $|C_1| = 1$ and $|C_2| = N - 1$. In this case, there is a large, uniform subgroup, but there is a group member who has a low similarity with all the other group members. This defines a minority group in which a group member is in a minority position.
4. Coalitional: this group is defined by a partition $\{C_1, C_2\}$ such that $|C_1| = N/2$ and $|C_2| = N - |C_1|$. In this case, the group can be divided into two uniform subgroups with low intra-group similarity. This defines a coalitional group composed of two coalitions.

In the following sections, we briefly introduce the similarity metrics which can be chosen to define the groups of interest and then illustrate the group generation algorithms for each group of interest.

Appendix B.2 Similarity metrics

We first considered metrics used in the scientific literature for similar purposes to choose the appropriate similarity metric to define our groups of interest. Statistical correlation measures, such as Pearson and Spearman correlation, have been used to determine the similarity between group members in the generation of synthetic groups for offline evaluation of group recommenders using datasets that do not include information about the group (Baltrunas et al. 2010). Delic et al. (2020) introduced several diversity metrics for a group, such as the Full Choice-Set Diversity metric which

Table 10 Similarity metrics comparison. For each considered metric, the quartiles are reported together with the similarity between two pairs of vectors: (i) L1 = [1, 2, 1, 2, 1, 2, 1, 2, 1, 2] and L2 = [2, 1, 2, 1, 2, 1, 2, 1, 2, 1]; (ii) L1 and H1 = [4, 5, 4, 5, 4, 5, 4, 5, 4, 5]

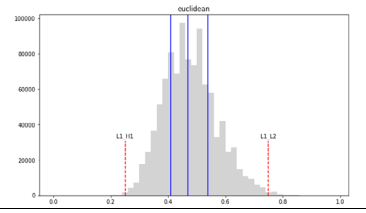
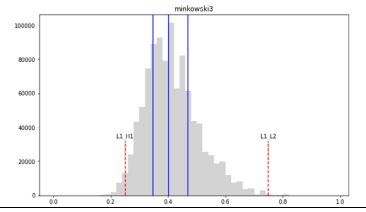
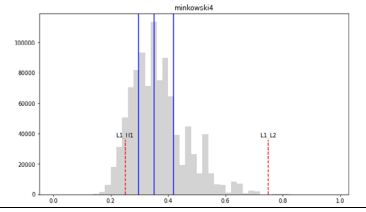
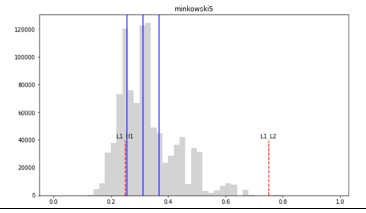
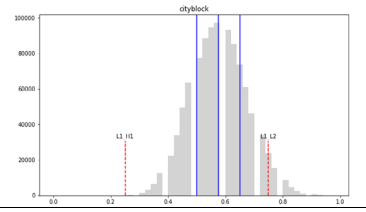
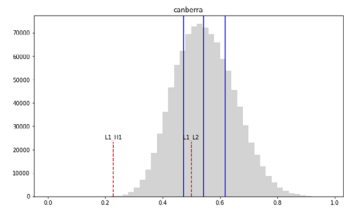
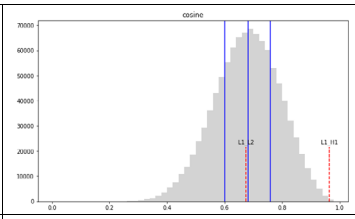
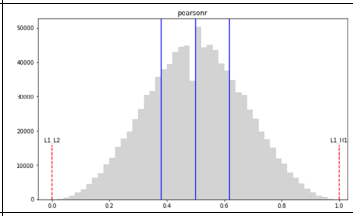
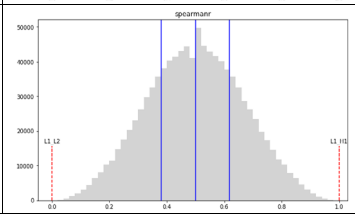
Similarity metric	Distribution plot
Euclidean 1st quartile = 0.408 2nd quartile = 0.470 3rd quartile = 0.539 $sim(L1, L2) = 0.750$ $sim(L1, H1) = 0.250$	
Minkowski (p=3) 1st quartile = 0.347 2nd quartile = 0.402 3rd quartile = 0.469 $sim(L1, L2) = 0.750$ $sim(L1, H1) = 0.250$	
Minkowski (p=4) 1st quartile = 0.297 2nd quartile = 0.352 3rd quartile = 0.418 $sim(L1, L2) = 0.750$ $sim(L1, H1) = 0.250$	
Minkowski (p=5) 1st quartile = 0.256 2nd quartile = 0.312 3rd quartile = 0.368 $sim(L1, L2) = 0.750$ $sim(L1, H1) = 0.250$	
Manhattan 1st quartile = 0.500 2nd quartile = 0.575 3rd quartile = 0.650 $sim(L1, L2) = 0.750$ $sim(L1, H1) = 0.250$	
Canberra 1st quartile = 0.473 2nd quartile = 0.543 3rd quartile = 0.618 $sim(L1, L2) = 0.750$ $sim(L1, H1) = 0.229$	

Table 10 continued

Similarity metric	Distribution plot
Cosine 1st quartile = 0.601 2nd quartile = 0.681 3rd quartile = 0.759 $sim(L1, L2) = 0.675$ $sim(L1, H1) = 0.964$	
Pearson 1st quartile = 0.381 2nd quartile = 0.500 3rd quartile = 0.619 $sim(L1, L2) = 0.000$ $sim(L1, H1) = 1.000$	
Spearman 1st quartile = 0.381 2nd quartile = 0.500 3rd quartile = 0.619 $sim(L1, L2) = 0.000$ $sim(L1, H1) = 1.000$	

Finally, the distribution is also illustrated in the associated figure, in grey color. The vertical blue lines highlight the quartiles, while the red dashed segments indicate the similarity between the two vectors L1 and L2 and the similarity between the vectors L1 and H1

measures the average pairwise Spearman foot-rule distance between group members. However, the Spearman foot-rule distance measures the difference between two rankings. In our context, we need to compare quantitative variable vectors. Other options include similarity metrics widely used for clustering (Alamuri et al. 2014; Irani et al. 2016; Lesot et al. 2009). We can identify similarity metrics defined based on a distance metric, such as Euclidean, Manhattan, Canberra, and Minkowski (Lesot et al. 2009). Furthermore, we also considered the Cosine similarity, widely used in clustering and machine learning applications (Irani et al. 2016).

All the distance metrics were first normalized to ensure that the returned values are in the [0, 1] interval, by applying a transformation from Lesot et al. (2009):

$$D_{norm}(u_1, u_2) = min\left(max\left(\frac{D(u_1, u_2) - m}{M - m}, 0\right), 1\right) \tag{1}$$

where u_1, u_2 are two users, $D(u_1, u_2)$ is the computed distance between the two vector representations of the two users, m and M are normalization values. More specifically, we considered two vectors to perform such normalization, $v_1 = [1, M, 1, M, \ldots, 1, M]$, and $v_2 = [M, 1, M, 1, \ldots, M, 1]$. Assuming that such vectors represent evaluations of items given by a group member, these two vectors

represent two people with opposing preferences (in this example, M is to highest rating a person can assign to an item, while 1 is the lowest). Hence, m is obtained by computing the distance between v_1 and itself, while M is computed by evaluating the distance between v_1 and v_2.

Finally, to transform such normalized diversity metrics into similarity metrics, we used a simple decreasing function from Lesot et al. (2009):

$$sim(u_1, u_2) = 1 - D_{norm}(u_1, u_2) \tag{2}$$

In order to decide which similarity to use, we analyzed the behavior of the different similarity metrics in our application scenario. More specifically, we evaluated the distribution of the similarity values computed on a population of 1000 randomly generated vectors (hence, we evaluated the similarity between each pair of such vectors). Furthermore, we evaluated some edge cases of pairs of specific vectors. In Table 10, for each considered metric, we show the distribution of the values (in grey). The vertical blue lines highlight the quartiles, while the red dashed segments indicate the similarity between two pairs of vectors: (i) $L1 = [1, 2, 1, 2, 1, 2, 1, 2, 1, 2]$ and $L2 = [2, 1, 2, 1, 2, 1, 2, 1, 2, 1]$; (ii) $L1$ and $H1 = [4, 5, 4, 5, 4, 5, 4, 5, 4, 5]$. Intuitively, $L1$ and $L2$ are both vectors mapping very low evaluations for all the items; on the contrary, $H1$ represents a vector containing very high evaluations for all the items. We would like that the chosen similarity metric computes a high similarity between $L1$ and $L2$ (in the fourth quartile) and a low similarity between $L1$ and $H1$ (in the first quartile).

As we can see from the charts, statistical correlations cannot be applied in this scenario. Our borderline examples allow highlighting the problem with these metrics, which return a high similarity between $L1$ ad $H1$ and a low similarity between $L1$ ad $L2$. Something similar also happens for the cosine similarity. This makes them not ideal for our purposes. Among the others, we decided to select the Euclidean similarity since the similarity assigned to the borderline examples is the desired one and the returned values are normally distributed. We also used the quartiles boundaries as thresholds to determine if the similarity between two vectors is low or high: values belonging to the first quartile (smaller than 0.408) have a small similarity, while values in the fourth quartile (higher than 0.539) have a high similarity.

Appendix B.3 Group generation algorithm

We used a brute force approach to generate the groups representing each configuration of interest. Let's assume we want to generate a group U of N people and their corresponding ratings for M items. In this case, $N = 5$ and $M = 10$. We also set the maximum rating $max\_rate = 5$. As specified in the previous subsection, we used the Euclidean distance to define our similarity metric, using the thresholds $\theta_1 = 0.539$, $\theta_2 = 0.408$.

Note that all the generated configurations are one representative of the possible groups complying with the constraint imposed on each configuration of interest. Due to the randomness of the algorithm, the obtained group configuration will be differ-

ent at any execution. Furthermore, we only consider user vectors containing 2 or 3 evaluations with a maximum rating and 2 or 3 evaluations with a minimum rating.

In the next subsection, we illustrate the generation of each group configuration, report the resulting scenario, and show the corresponding social graph, highlighting the considered partition in different background colors. In the figures, the relationships with high similarity are in solid black, while the relationships with low similarity are in dashed grey. We also report the pseudo-codes of the generation algorithms for each group configuration.

Appendix B.3.1 Uniform group

In this case, the group is characterized by the trivial partition $\{U\}$. Considering the definition 1, we want that $\forall u_i, u_j \in U \implies sim(u_i, u_j) \geq \theta_1$. The pseudo-code of the algorithm is illustrated in Algorithm 1, while Table 11 shows the obtained group and Fig. 9 provides a graphical representation of the corresponding social graph.

Table 11 Uniform group configuration

	$Rest_1$	$Rest_2$	$Rest_3$	$Rest_4$	$Rest_5$	$Rest_6$	$Rest_7$	$Rest_8$	$Rest_9$	$Rest_{10}$
Alex	4	5	4	5	4	4	3	2	1	1
Bob	1	5	1	4	3	4	3	5	1	3
Carl	2	4	2	2	5	5	2	5	1	1
David	1	4	1	2	5	5	2	2	2	1
Elle	3	5	2	3	4	5	1	3	1	2

Recommendation lists:
MAJ: $Rest_2$, $Rest_6$, $Rest_5$, $Rest_8$, $Rest_4$, $Rest_1$, $Rest_3$, $Rest_7$, $Rest_{10}$, $Rest_9$
APP: $Rest_2$, $Rest_6$, $Rest_5$, $Rest_4$, $Rest_8$, $Rest_1$, $Rest_3$, $Rest_7$, $Rest_9$, $Rest_{10}$
ADD: $Rest_2$, $Rest_6$, $Rest_5$, $Rest_8$, $Rest_4$, $Rest_1$, $Rest_7$, $Rest_3$, $Rest_{10}$, $Rest_9$
FAI: $Rest_2$, $Rest_8$, $Rest_5$, $Rest_6$, $Rest_1$, $Rest_4$, $Rest_7$, $Rest_3$, $Rest_9$, $Rest_{10}$
LMS: $Rest_2$, $Rest_6$, $Rest_5$, $Rest_4$, $Rest_8$, $Rest_1$, $Rest_3$, $Rest_7$, $Rest_9$, $Rest_{10}$
MPL: $Rest_2$, $Rest_4$, $Rest_5$, $Rest_6$, $Rest_8$, $Rest_1$, $Rest_3$, $Rest_7$, $Rest_{10}$, $Rest_9$

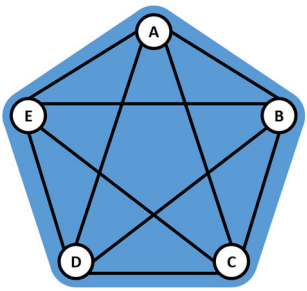


Fig. 9 Graphic representation of a Uniform group. The nodes represent the group members, solid black edges represent pairs with high similarity, grey dashed edges represent pairs with low similarity. In this case, we only have high similarity; hence, all the group members belong to the same subset

Appendix B.3.2 Divergent group

In this case, the group is characterized by the partition $\{\{u_1\}, \ldots, \{u_N\}\}$. Applying the definition 1, we want that $\forall u_i, u_j \in U \implies sim(u_i, u_j) \leq \theta_2$. The pseudo-code of the algorithm is illustrated in Algorithm 2, while Table 12 shows the obtained group and Fig. 10 provides a graphical representation of the corresponding social graph.

Table 12 Divergent group configuration

	$Rest_1$	$Rest_2$	$Rest_3$	$Rest_4$	$Rest_5$	$Rest_6$	$Rest_7$	$Rest_8$	$Rest_9$	$Rest_{10}$
Fran	5	4	3	2	2	1	2	3	5	1
Gene	1	4	5	4	5	1	4	3	1	5
Hilary	2	4	1	5	1	5	3	1	3	4
Izzy	4	2	2	1	4	5	1	5	1	4
Jess	1	1	4	4	5	5	4	4	5	1

Recommendation lists:
MAJ: $Rest_6$, $Rest_5$, $Rest_9$, $Rest_3$, $Rest_4$, $Rest_8$, $Rest_{10}$, $Rest_1$, $Rest_2$, $Rest_7$
APP: $Rest_2$, $Rest_4$, $Rest_5$, $Rest_6$, $Rest_{10}$, $Rest_1$, $Rest_3$, $Rest_7$, $Rest_8$, $Rest_9$
ADD: $Rest_5$, $Rest_6$, $Rest_4$, $Rest_8$, $Rest_2$, $Rest_3$, $Rest_9$, $Rest_{10}$, $Rest_7$, $Rest_1$
FAI: $Rest_1$, $Rest_3$, $Rest_4$, $Rest_6$, $Rest_5$, $Rest_9$, $Rest_{10}$, $Rest_2$, $Rest_8$, $Rest_7$
LMS: $Rest_1$, $Rest_2$, $Rest_3$, $Rest_4$, $Rest_5$, $Rest_6$, $Rest_7$, $Rest_8$, $Rest_9$, $Rest_{10}$
MPL: $Rest_1$, $Rest_3$, $Rest_4$, $Rest_5$, $Rest_6$, $Rest_8$, $Rest_9$, $Rest_{10}$, $Rest_2$, $Rest_7$

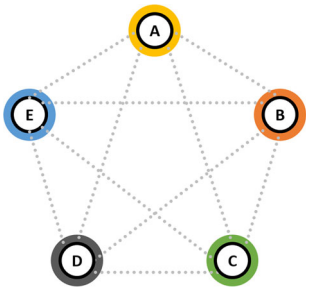


Fig. 10 Graphic representation of a Divergent group. The nodes represent the group members, solid black edges represent pairs with high similarity, grey dashed edges represent pairs with low similarity. In this case, we only have pairs with low similarity; hence, each group member belongs to a different subset

Appendix B.3.3 Minority group

In this case, the group is characterized by a partition $\{C_1, C_2\}$ such that $|C_1| = 1$ and $|C_2| = N - 1$. Applying the definition 1, we want that (i) $\forall u_i, u_j \in C_2 \implies sim(u_i, u_j) \geq \theta_1$, and (ii) being $C_1 = \{u_k\}$, $\forall u_i \in C_2 \implies sim(u_i, u_k) \leq \theta_2$. The pseudo-code of the algorithm is illustrated in Algorithm 3, while Table 13 shows the obtained group and Fig. 11 provides a graphical representation of the corresponding social graph.

Table 13 Minority group configuration

	$Rest_1$	$Rest_2$	$Rest_3$	$Rest_4$	$Rest_5$	$Rest_6$	$Rest_7$	$Rest_8$	$Rest_9$	$Rest_{10}$
Kris	3	3	5	5	3	1	4	5	1	3
Leslie	2	5	5	1	3	1	5	3	3	4
Max	4	4	3	1	3	1	4	5	2	5
Noel	4	4	5	4	2	1	3	2	1	5
Pat	4	2	1	3	2	5	4	2	5	1

Recommendation lists:
MAJ: $Rest_3$, $Rest_8$, $Rest_{10}$, $Rest_2$, $Rest_1$, $Rest_4$, $Rest_7$, $Rest_5$, $Rest_9$, $Rest_6$
APP: $Rest_7$, $Rest_1$, $Rest_2$, $Rest_3$, $Rest_{10}$, $Rest_4$, $Rest_8$, $Rest_6$, $Rest_9$, $Rest_5$
ADD: $Rest_7$, $Rest_3$, $Rest_2$, $Rest_{10}$, $Rest_1$, $Rest_8$, $Rest_4$, $Rest_5$, $Rest_9$, $Rest_6$
FAI: $Rest_3$, $Rest_2$, $Rest_8$, $Rest_{10}$, $Rest_6$, $Rest_4$, $Rest_7$, $Rest_1$, $Rest_5$, $Rest_9$
LMS: $Rest_7$, $Rest_1$, $Rest_2$, $Rest_5$, $Rest_8$, $Rest_3$, $Rest_4$, $Rest_6$, $Rest_9$, $Rest_{10}$
MPL: $Rest_2$, $Rest_3$, $Rest_4$, $Rest_6$, $Rest_7$, $Rest_8$, $Rest_9$, $Rest_{10}$, $Rest_1$, $Rest_5$

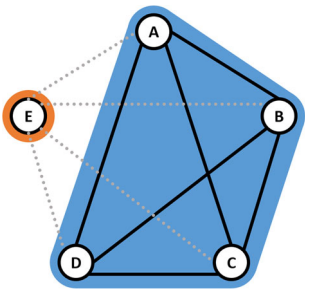


Fig. 11 Graphic representation of a Minority group. The nodes represent the group members, solid black edges represent pairs with high similarity, grey dashed edges represent pairs with low similarity. In this case, we have pairs with high similarity between the nodes A, B, C, and D. All such nodes have a low similarity with the node E, which represents a group member in a minority position in the group

Appendix B.3.4 Coalitional group

In this case, the group is characterized by a partition $\{C_1, C_2\}$ such that $|C_1| = N/2$ and $|C_2| = N - |C_1|$. Applying the definition 1, we want that (i) $\forall u_i, u_j \in C_1 \implies sim(u_i, u_j) \geq \theta_1$, (ii) $\forall u_i, u_j \in C_2 \implies sim(u_i, u_j) \geq \theta_1$, and (iii) $\forall u_i \in C_1 \wedge u_j \in C_2 \implies sim(u_i, u_j) \leq \theta_2$. The pseudo-code of the algorithm is illustrated in Algorithm 4, while Table 14 shows the obtained group and Fig. 12 provides a graphical representation of the corresponding social graph.

Table 14 Coalitional group configuration

	$Rest_1$	$Rest_2$	$Rest_3$	$Rest_4$	$Rest_5$	$Rest_6$	$Rest_7$	$Rest_8$	$Rest_9$	$Rest_{10}$
Robin	2	4	2	5	2	1	1	4	5	5
Sandy	4	4	1	5	3	4	5	3	1	2
Terry	5	4	3	5	5	4	4	3	1	1
Vic	4	5	4	4	3	4	3	5	1	1
Willie	1	4	1	3	3	2	1	4	5	5

Recommendation lists:
MAJ: $Rest_4$, $Rest_9$, $Rest_{10}$, $Rest_2$, $Rest_8$, $Rest_1$, $Rest_5$, $Rest_6$, $Rest_3$, $Rest_7$
APP: $Rest_2$, $Rest_4$, $Rest_1$, $Rest_6$, $Rest_8$, $Rest_7$, $Rest_9$, $Rest_{10}$, $Rest_3$, $Rest_5$
ADD: $Rest_4$, $Rest_2$, $Rest_8$, $Rest_1$, $Rest_5$, $Rest_6$, $Rest_7$, $Rest_{10}$, $Rest_9$, $Rest_3$
FAI: $Rest_4$, $Rest_9$, $Rest_1$, $Rest_2$, $Rest_7$, $Rest_{10}$, $Rest_8$, $Rest_5$, $Rest_3$, $Rest_6$
LMS: $Rest_2$, $Rest_4$, $Rest_8$, $Rest_5$, $Rest_1$, $Rest_3$, $Rest_6$, $Rest_7$, $Rest_9$, $Rest_{10}$
MPL: $Rest_1$, $Rest_2$, $Rest_4$, $Rest_5$, $Rest_7$, $Rest_8$, $Rest_9$, $Rest_{10}$, $Rest_3$, $Rest_6$

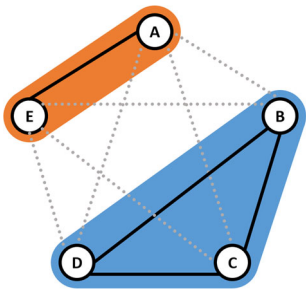


Fig. 12 Graphic representation of a Coalitional group. The nodes represent the group members, solid black edges represent pairs with high similarity, grey dashed edges represent pairs with low similarity. In this case, we can see two subgroups with high inter-group similarity and low intra-group similarity. This results in a division of the group into two coalitions with different preferences

Algorithm 1: Generation of the *Uniform* group

Input: $N > 0; M > 0; \theta_1 > 0; max\_rate > 0$
Output: U
$U = \{\}$;
while $|U| < N$ **do**
  $u_{new} = generate\_candidate(M, max\_rate)$; /* returns a random vector of size M */
  if $check\_candidate(u_{new})$ **then**
    $good\_candidate = True$
    forall the $u_k \in U$ **do**
      if $sim(u_{new}, u_k) < \theta_1$ **then**
        $good\_candidate = False$
      end
    end
    if $good\_candidate = True$ **then**
      $U.add(u_k)$; /* adds the u_k vector to the set U */
    end
  end
end
return U

Algorithm 2: Generation of the *Divergent* group

Input: $N > 0; M > 0; \theta_2 > 0; max\_rate > 0$
Output: U
$U = \{\}$;
while $|U| < N$ **do**
  $u_{new} = generate\_candidate(M, max\_rate)$; /* returns a random vector of size M */
  if $check\_candidate(u_{new})$ **then**
    $good\_candidate = True$
    forall the $u_k \in U$ **do**
      if $sim(u_{new}, u_k) > \theta_2$ **then**
        $good\_candidate = False$
      end
    end
    if $good\_candidate = True$ **then**
      $U.add(u_k)$; /* adds the u_k vector to the set U */
    end
  end
end
return U

Algorithm 3: Generation of the *Minority* group

Input: $N > 0; M > 0; \theta_1 > 0; \theta_2 > 0; max\_rate > 0$
Output: U
$U = \{\}$;
while $|U| < N - 1$ **do**
  $u_{new} = generate\_candidate(M, max\_rate)$; /* returns a random vector of size M */
  if $check\_candidate(u_{new})$ **then**
    $good\_candidate = True$
    forall the $u_k \in U$ **do**
      if $sim(u_{new}, u_k) < \theta_1$ **then**
        $good\_candidate = False$
      end
    end
    if $good\_candidate = True$ **then**
      $U.add(u_k)$; /* adds the u_k vector to the set U */
    end
  end
end
while $|U| < N$ **do**
  $u_{new} = generate\_candidate(M, max\_rate)$; /* returns a random vector of size M */
  if $check\_candidate(u_{new})$ **then**
    $good\_candidate = True$
    forall the $u_k \in U$ **do**
      if $sim(u_{new}, u_k) > \theta_2$ **then**
        $good\_candidate = False$
      end
    end
    if $good\_candidate = True$ **then**
      $U.add(u_k)$; /* adds the u_k vector to the set U */
    end
  end
end
return U

Algorithm 4: Generation of the *Coalitional* group

Input: $N > 0; M > 0; \theta_1 > 0; \theta_2 > 0; max\_rate > 0$
Output: U
$G_1 = \{\}$;
while $|G_1| < N/2$ **do**
  $u_{new} = generate\_candidate(M, max\_rate)$; /* returns a random vector of size M */
  if $check\_candidate(u_{new})$ **then**
    $good\_candidate = True$
    forall the $u_k \in G_1$ **do**
      if $sim(u_{new}, u_k) < \theta_1$ **then**
        $good\_candidate = False$
      end
    end
    if $good\_candidate = True$ **then**
      $U.add(u_k)$; /* adds the u_k vector to the set G_1 */
    end
  end
end
$G_2 = \{\}$;
while $|G_2| < N - |G_1|$ **do**
  $u_{new} = generate\_candidate(M, max\_rate)$; /* returns a random vector of size M */
  if $check\_candidate(u_{new})$ **then**
    $good\_candidate = True$
    forall the $u_k \in G_2$ **do**
      if $sim(u_{new}, u_k) < \theta_1$ **then**
        $good\_candidate = False$
      end
    end
    forall the $u_k \in G_1$ **do**
      if $sim(u_{new}, u_k) > \theta_2$ **then**
        $good\_candidate = False$
      end
    end
    if $good\_candidate = True$ **then**
      $U.add(u_k)$; /* adds the u_k vector to the set G_2 */
    end
  end
end
$U = G_1 \cup G_2$;
return U

References

Alamuri, M., Surampudi, B.R., Negi, A.: A survey of distance/similarity measures for categorical data. In: 2014 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1907–1914 (2014)

Ardissono, L., Goy, A., Petrone, G., et al.: Intrigue: personalized recommendation of tourist attractions for desktop and hand held devices. Appl. Artif. Intell. **17**(8–9), 687–714 (2003)

Arrow, K.J.: A difficulty in the concept of social welfare. J. Polit. Econ. **58**(4), 328–346 (1950)

Baltrunas, L., Makcinskas, T., Ricci, F.: Group recommendations with rank aggregation and collaborative filtering. In: Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 119–126 (2010)

Barile, F., Najafian, S., Draws, T., et al.: Toward benchmarking group explanations: Evaluating the effect of aggregation strategies versus explanation. In: Proceedings of Perspectives@ RecSys (2021)

Berkovsky, S., Freyne, J.: Group-based recipe recommendations: analysis of data aggregation strategies. In: Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 111–118 (2010)
Cao, D., He, X., Miao, L., et al.: Attentive group recommendation. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 645–654 (2018)
Chen, L., De Gemmis, M., Felfernig, A., et al.: Human decision making and recommender systems. ACM Trans. Interactive Intell. Syst. (TiiS) **3**(3), 1–7 (2013)
Chen, Y.L., Cheng, L.C., Chuang, C.N.: A group recommendation system with consideration of interactions among group members. Expert Syst. Appl. **34**(3), 2082–2090 (2008)
Delic, A., Masthoff, J., Neidhardt, J., et al.: How to use social relationships in group recommenders: empirical evidence. In: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, pp 121–129 (2018)
Delic, A., Masthoff, J., Werthner, H.: 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, p 117–129 (2020)
Faul, F., Erdfelder, E., Lang, A.G., et al.: G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods **39**(2), 175–191 (2007). https://doi.org/10.3758/BF03193146
Felfernig, A., Boratto, L., Stettinger, M., et al.: Explanations for groups. In: Group Recommender Systems. Springer, pp. 105–126 (2018)
Gartrell M, Xing X, Lv Q, et al.: Enhancing group recommendation by incorporating social relationship interactions. In: Proceedings of the 16th ACM International Conference on Supporting Group Work, pp. 97–106 (2010)
Gedikli, F., Jannach, D., Ge, M.: How should i explain? A comparison of different explanation types for recommender systems. Int. J. Hum Comput Stud. **72**(4), 367–382 (2014)
Herzog, D., Wörndl, W.: User-centered evaluation of strategies for recommending sequences of points of interest to groups. In: Proceedings of the 13th ACM Conference on Recommender Systems, pp. 96–100 (2019)
Irani, J., Pise, N., Phatak, M.: Clustering techniques and the similarity measures used in clustering: A survey. Int. J. Comput. Appl. **134**(7), 9–14 (2016)
Jannach, D., Zanker, M., Felfernig, A., et al.: Recommender Systems: An Introduction. Cambridge University Press, Cambridge (2010)
Kapcak, Ö., Spagnoli, S., Robbemond, V., et al: 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, pp 33–36 (2018)
Kaya, M., Bridge, D., Tintarev, N.: Ensuring fairness in group recommendations by rank-sensitive balancing of relevance. In: Fourteenth ACM Conference on Recommender Systems, pp. 101–110 (2020)
Kelly, J.S.: Social choice theory: An introduction. Springer Science & Business Media (2013)
Kim, J.K., Kim, H.K., Oh, H.Y., et al.: A group recommendation system for online communities. Int. J. Inf. Manage. **30**(3), 212–219 (2010)
Knijnenburg, B.P., Reijmer, N.J., Willemsen, M.C.: 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, pp 141–148 (2011)
Lesot, M.J., Rifqi, M., Benhadda, H.: Similarity measures for binary and numerical data: A survey. Int. J. Knowl. Eng. Soft Data Paradigms **1**(1), 63–84 (2009)
Malecek, L., Peska, L.: Fairness-preserving group recommendations with user weighting. In: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, pp. 4–9 (2021)
Masthoff, J.: Group modeling: Selecting a sequence of television items to suit a group of viewers. In: Personalized Digital Television. Springer, pp. 93–141 (2004)
Masthoff, J.: Group recommender systems: aggregation, satisfaction and group attributes. In: Recommender Systems Handbook. Springer, pp. 743–776 (2015)
Masthoff, J., Delić, A.: Group recommender systems: Beyond preference aggregation. In: Recommender Systems Handbook. Springer, pp. 381–420 (2022)
Masthoff, J., Gatt, A.: In pursuit of satisfaction and the prevention of embarrassment: Affective state in group recommender systems. User Model. User-Adap. Inter. **16**(3–4), 281–319 (2006)

Najafian, S., Tintarev, N.: Generating consensus explanations for group recommendations: an exploratory study. In: Adjunct Publication of the 26th Conference on User Modeling, pp. 245–250. ACM, Adaptation and Personalization (2018)
Najafian, S., Herzog, D., Qiu, S., et al.: 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, pp 187–196 (2020a)
Najafian, S., Inel, O., Tintarev, N.: 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, pp 85–86 (2020b)
Najafian, S., Delic, A., Tkalcic, M., et al.: Factors influencing privacy concern for explanations of group recommendation. In: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, pp 14–23 (2021a)
Najafian, S., Draws, T., Barile, F., et al.: Exploring user concerns about disclosing location and emotion information in group recommendations. In: Proceedings of the 32st ACM Conference on Hypertext and Social Media, pp. 155–164 (2021b)
Napierala, M. A.: What Is the Bonferroni correction? http://www.aaos.org/news/aaosnow/apr12/research7.asp (2012)
Nguyen, T.N., Ricci, F., Delic, A., et al.: Conflict resolution in group decision making: Insights from a simulation study. User Model. User-Adap. Inter. **29**(5), 895–941 (2019)
Norman, G.: Likert scales, levels of measurement and the "laws" of statistics. Adv. Health Sci. Educ. **15**(5), 625–632 (2010). https://doi.org/10.1007/s10459-010-9222-y
Ntoutsi, E., Stefanidis, K., Nørvåg, K., et al.: Fast group recommendations by applying user clustering. In: International Conference on Conceptual Modeling, Springer, pp. 126–140 (2012)
O'connor, M., Cosley, D., Konstan, J.A., et al.: Polylens: A recommender system for groups of users. In: ECSCW 2001, Springer, pp 199–218 (2001)
Quijano-Sanchez, L., Sauer, C., Recio-Garcia, J.A., et al.: Make it personal: a social explanation system applied to group recommendations. Expert Syst. Appl. **76**, 36–48 (2017)
Rossi, S., Caso, A., Barile, F.: Combining users and items rankings for group decision support. In: Bajo, J., Hernández, J.Z., Mathieu, P., et al. (eds.) Trends in Practical Applications of Agents, Multi-Agent Systems and Sustainability, pp. 151–158. Springer International Publishing, Cham (2015)
Rossi, S., Barile, F., Caso, A., et al.: Pre-trip ratings and social networks user behaviors for recommendations in touristic web portals. In: Monfort, V., Krempels, K.H., Majchrzak, T.A., et al. (eds.) Web Information Systems and Technologies, pp. 297–317. Springer International Publishing, Cham (2016)
Rossi, S., Cervone, F., Barile, F.: An altruistic-based utility function for group recommendation. In: Transactions on Computational Collective Intelligence XXVIII. Springer, pp. 25–47 (2018)
Sankar, A., Wu, Y., Wu, Y., et al: Groupim: A mutual information maximization framework for neural group recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1279–1288 (2020)
Senot, C., Kostadinov, D., Bouzid. M., et al: Analysis of strategies for building group profiles. In: International Conference on User Modeling, Adaptation, and Personalization, Springer, pp 40–51 (2010)
Sinha, R., Swearingen, K.: The role of transparency in recommender systems. In: CHI'02 Extended Abstracts on Human Factors in Computing Systems, pp. 830–831 (2002)
Tintarev, N., Masthoff, J.: Over-and underestimation in different product domains. In: Workshop on Recommender Systems associated with ECAI, pp. 14–19. Springer, Boston (2008)
Tintarev, N., Masthoff, J.: Beyond explaining single item recommendations. In: Recommender Systems Handbook. Springer, pp. 711–756 (2022)
Tran, T.N.T., Atas, M., Felfernig, A., et al.: Towards social choice-based explanations in group recommender systems. In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, pp. 13–21 (2019)
Vinh Tran, L., Nguyen Pham, T.A., Tay, Y., et al.: Interact and decide: Medley of sub-attention networks for effective group recommendation. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 255–264 (2019)
Wang, X., Yin, M.: Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In: 26th International Conference on Intelligent User Interfaces, pp 318–328 (2021)
Zhang, J.S., Gartrell, M., Han, R., et al.: Gevr: An event venue recommendation system for groups of mobile users. Proc. ACM Interactive Mobile Wearable Ubiquit. Technol. **3**(1), 1–25 (2019)

Francesco Barile is an Assistant Professor of Explainable Recommender System at Maastricht University, Faculty of Science and Engineering, Department of Advanced Computing Sciences. His research focuses on Group Recommender Systems, Multi-stakeholder Recommender Systems, and Explainable AI. More specifically, he is currently investigating the influence of the group context on the individual satisfaction, defining novel aggregation strategies to support the group decision-making process, and strategies for generating explanations for group recommendations. Furthermore, he is working on designing explanations tailored for each specific stakeholder for job recommenders. He is involved in the program committee of conferences like Recsys, IUI, UMAP, XAI, and he is organizer of the workshop on Group Modeling, Adaptation and Personalization (GMAP) co-located with the UMAP conference. He published over 30 papers in conferences and journals such as UMUAI, TiiS, TCCI, Recsys, ECAI, Hypertext, and UMAP.

Tim Draws is a PhD Candidate under the supervision of Prof. Nava Tintarev in the Web Information Systems group at Delft University of Technology in the Netherlands. Based on a background in psychology, research methods, and statistics, his work now focuses on viewpoint biases in web search results and their effects on user behavior and opinions. He has co-authored 22 published research papers, 10 of which as first author, and four of which have won best paper awards. Furthermore, he has served on the program committee for academic conferences such as SIGIR, CHIIR, and The Web Conference.

Oana Inel is a Postdoctoral Researcher in the Dynamic and Distributed Information Systems group at the University of Zurich. Her research focuses on measuring the quality of human-annotated and human-generated data and their impact on the development of decision-support systems. She is also investigating the use and role of explanations as a means to support humans in decision-making. Previously, Oana was a Postdoctoral Researcher at the Delft University of Technology and she received her PhD at the Vrije Universiteit Amsterdam, where her research focused on detecting, representing, and exploiting events and their explicit semantics for understanding, organizing, and representing knowledge on the Web.

Alisa Rieger is a PhD candidate in the Web Information Systems group at TU Delft and an Early Stage Researcher in the NL4XAI project. Before coming to TU Delft, she obtained her BSc and MSc degree in cognitive science and human-machine interaction from Chemnitz University of Technology. Currently, her research is focused on understanding and mitigating users' cognitive biases during web search. Specifically, she researches interventions to support thorough and unbiased search on debated topics, for example through boosting or nudging approaches.

Shabnam Najafian is a researcher in the field of group recommendation systems. She received her Ph.D. in Computer Science from the Technical University of Delft, in 2023. Since then, she has been working as a data specialist and risk model developer at the DLL company. Her research interests include human-centered computing, privacy-preserving natural language explanations, user modeling, and group recommender systems. Her research has been published in leading conferences and journals in related domains (e.g., ACM RecSys, ACM UMAP, ACM Hypertext, IUI, and the UMUAI Journal). As a reviewer for several conferences, she served on the reviews for UMUAI, ACM RecSys, and ACM UMAP. In her free time, Shabnam enjoys hiking and playing tennis.

Amir Ebrahimi Fard is a Postdoctoral researcher at the Department of Advanced Computing Sciences in Maastricht University, studying decision making in online media sphere. He completed his PhD at Delft University of Technology with a thesis on countering rumors in online social media. Amir's research interests lies in the intersection of data, algorithm, and society.

Rishav Hada is a Research Fellow at Microsoft Research. He is interested in the intersection of Natural Language Processing, Computational Social Science, and Fairness and Transparency in AI. Specifically, he is interested in understanding how language use can reveal information about individuals and communities, and how to integrate this knowledge into neural models to develop socially inclusive AI applications. He has conducted research on various topics, including offensive language detection, measuring viewpoint diversity, and identifying gender biases in datasets and models. Rishav is motivated to address

the limitations of existing methods and develop new strategies for dataset evaluation that promote careful curation and help mitigate social biases.

Nava Tintarev is a Full Professor of Explainable Artificial Intelligence at the University of Maastricht, and a guest professor at TU Delft. She leads or contributes to several projects in the field of human-computer interaction in artificial advice-giving systems, such as recommender systems; specifically developing the state-of-the-art for automatically generated explanations (transparency) and explanation interfaces (recourse and control). She currently participates in a Marie-Curie Training Network on Natural Language for Explainable AI (October 2019–October 2024). She is also representing Maastricht University as a Co-Investigator in the ROBUST consortium, selected for a national (NWO) grant with a total budget of 95M (25M from NWO) to carry out long-term (10-years) research into trustworthy artificial intelligence, and is a co-director of the TAIM lab on trustworthy media. She regularly shapes international scientific research programs (e.g., on steering committees of journals, or as program chair of conferences), and actively organizes and contributes to high-level strategic workshops relating to responsible data science, both in the Netherlands and internationally. She has published over 80 peer-reviewed papers in top human-computer interaction and artificial intelligence journals and conferences such as UMUAI, TiiS, ECAI, ECIR, IUI, Recsys, and UMAP. These include best paper awards at Hypertext, CHI, HCOMP, and CHIIR.

Authors and Affiliations

Francesco Barile[1] · Tim Draws[2] · Oana Inel[3] · Alisa Rieger[2] · Shabnam Najafian[2] · Amir Ebrahimi Fard[1] · Rishav Hada[1,4] · Nava Tintarev[1]

Tim Draws
t.a.draws@tudelft.nl

Oana Inel
inel@ifi.uzh.ch

Alisa Rieger
a.rieger@tudelft.nl

Shabnam Najafian
s.najafian@tudelft.nl

Amir Ebrahimi Fard
a.ebrahimifard@maastrichtuniversity.nl

Rishav Hada
rishavhada@gmail.com

Nava Tintarev
n.tintarev@maastrichtuniversity.nl

1 Department of Advanced Computing Sciences (DACS), Maastricht University, Maastricht, The Netherlands

2 Department of Software Technology (ST), Delft University of Technology, Delft, The Netherlands

3 Department of Informatics, University of Zurich, Zurich, Switzerland

4 Microsoft Research, Bangalore, India