

With Friends Like These, Who Needs Explanations? Evaluating User Understanding of Group Recommendations

Cedric Waterschoot
Maastricht University
Maastricht, The Netherlands
cedric.waterschoot@maastrichtuniversity.nl

Nava Tintarev
Maastricht University
Maastricht, The Netherlands
n.tintarev@maastrichtuniversity.nl

Raciel Yera Toledo
University of Jaén
Jaén, Spain
ryera@ujaen.es

Francesco Barile
Maastricht University
Maastricht, The Netherlands
f.barile@maastrichtuniversity.nl

Abstract

Group Recommender Systems (GRS) employing social choice-based aggregation strategies have previously been explored in terms of perceived consensus, fairness, and satisfaction. At the same time, the impact of textual explanations has been examined, but the results suggest a low effectiveness of these explanations. However, user understanding remains fairly unexplored, even if it can contribute positively to transparent GRS. This is particularly interesting to study in more complex or potentially unfair scenarios when user preferences diverge, such as in a minority scenario (where group members have similar preferences, except for a single member in a minority position). In this paper, we analyzed the impact of different types of explanations on user understanding of group recommendations. We present a randomized controlled trial ($n = 271$) using two between-subject factors: (i) the aggregation strategy (additive, least misery, and approval voting), and (ii) the modality of explanation (no explanation, textual explanation, or multimodal explanation). We measured both subjective (self-perceived by the user) and objective understanding (performance on model simulation, counterfactuals and error detection). In line with recent findings on explanations for machine learning models, our results indicate that more detailed explanations, whether textual or multimodal, did not increase subjective or objective understanding. However, we did find a significant effect of aggregation strategies on both subjective and objective understanding. These results imply that when constructing GRS, practitioners need to consider that the choice of aggregation strategy can influence the understanding of users. Post-hoc analysis also suggests that there is value in analyzing performance on different tasks, rather than through a single aggregated metric of understanding.

CCS Concepts

• **Human-centered computing** → **User studies**; *Social recommendation*; • **Information systems** → **Recommender systems**.

UMAP '25, New York City, NY, USA

Keywords

Group Recommender Systems, Social Choice-Based Explanations, Objective Understanding, Subjective Understanding, User Study

ACM Reference Format:

Cedric Waterschoot, Raciel Yera Toledo, Nava Tintarev, and Francesco Barile. 2025. With Friends Like These, Who Needs Explanations? Evaluating User Understanding of Group Recommendations. In *33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*, June 16–19, 2025, New York City, NY, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Group Recommender Systems (GRS) process the preferences of multiple individuals to derive a single recommendation tailored to suit the group as a whole. To achieve this, previous research has introduced social choice-based aggregation strategies [19]. These strategies present a range of options in which the preferences of individual group members are aggregated and present an opportunity to adapt GRS to different group configurations and needs. However, the implementation of these strategies on top of already complex recommender systems may result in non-transparent models, hindering user understanding. This is particularly interesting in scenarios in which user preferences diverge, leading to potentially unfair scenarios. Social choice-based explanations for group recommendations have been used as a way to intuitively explain the process behind group recommendations [13, 23]. These textual explanations have been evaluated in terms of users' fairness perception, consensus perception, and satisfaction in user studies with contrasting results [1, 30]. Hence, their effectiveness in improving user understanding of the strategies used to generate recommendations for the group still remains to be seen. In this study, we build upon these works, focusing on the understandability of explanations and in turn, group recommendations themselves. We formulate the following research questions:

RQ1. Do explanations increase users' understanding of the underlying social choice-based aggregation strategies?

RQ2. Does the effect of explanations on the understandability of GRS vary depending on the underlying social-based aggregation strategy?

Inspired by Wang and Yin [31], we constructed a randomized controlled trial with two between-subject factors (3x3 design): (i) the social choice-based aggregation strategy used for generating group recommendations, and (ii) the type of explanation provided to the participant. We included three explanation types: *no explanation* (control group), *textual explanation only* and, *a multimodal explanation* combining text with a visualization. To measure the participant’s understanding of the strategy and GRS, we made use of *objective* and *subjective understanding* as described in previous research [16, 25, 26]. We present minority scenarios, i.e. group scenarios in which user preferences are similar except for one single member, a complex and potentially unfair scenario.

Our paper makes the following contributions:

- We conduct a preregistered, randomized controlled trail ($n = 271$) to analyze the effect of different explanation types on user understanding.
- We show that more detailed explanations do not positively or negatively impact user understanding.
- We outline that the choice of aggregation strategy can impact understandability down the line.
- We find value in measuring understandability using a variety of tasks as opposed to a single metric.

The remainder of the paper is structured as follows. First, we outline the literature on GRS and understandability of Recommender Systems. Based on this body of previous work, we formulate our hypotheses. Second, we present our materials, including the measurement of understandability, explanation types, and experimental set-up. Subsequently, we describe the results of our experiment followed by a discussion of the results and their implications.

2 Background

In this section, we introduce the literature on group recommendation and social choice-based aggregation strategies. Additionally, we outline recent work combining understandability and Explainable AI (XAI) with Recommender Systems and highlight the research gap related to the understandability of GRS in particular.

2.1 Group Recommendation

Recommendation systems tailored to groups, rather than the preferences of a single user, are rising in demand due to applications in fields such as tourism [3] and music [23]. These Group Recommender Systems (GRS) are designed to support the decision-making process of multiple people simultaneously [20]. The preferences of the individual group members need to be processed in a way that the recommendation reflects the group as a whole. Based on *Social Choice Theory* [15], social choice-based aggregation strategies are employed for aggregating the preferences of the group members to obtain a recommendation for the group as a whole [19]. Examples of often used social-choice based approaches are summarized in Table 1. A range of different strategies have been introduced in the literature and can be categorized as follows: *consensus-based*, *borderline* and, *majority-based*. The latter makes use of only the most popular items or ratings [28]. An example of such a strategy is Approval Voting (APP), which uses a threshold to recommend items having the higher number of ratings above a certain value (Table 1). Borderline strategies do not include all ratings, and filter out a

subset of the item ratings in the group [28]. For example, the Least Misery (LMS) strategy recommends the item with the highest of all lowest per-item ratings, while the Most Pleasure (MPL) strategy looks for the highest overall rating (Table 1). On the other hand, consensus-based strategies include all ratings made by the group to derive a group recommendation [28]. An example of such a strategy is Additive Utilitarian (ADD). This strategy sums up all ratings per item and recommends the item with the highest sum (Table 1). In the current work, we make use of group recommendations derived by using a selection of strategies based on the categorization presented by Senot et al. [28].

2.2 Understandability of Recommender Systems

We put social choice-based aggregation strategies from group recommendation in the frame of Human-Centered Explainable Artificial Intelligence (HCXAI) [6, 26]. It emphasizes the importance of human stakeholders regarding the desired goals of intelligent systems, focusing on factors including trust, usability, human-AI collaboration, and understandability [26]. Specifically, understandability has attracted particular interest over the last few years [26]. Liao and Varshney [17] pointed out that understanding in the context of interacting with a machine learning (ML) model refers to the user’s grasp or mental model of how the ML model operates. This knowledge grows from using the system and from receiving clear explanations about it. More precisely, the interpretability of an AI model should be defined according to its influence over the users’ abilities in the completion of diverse tasks [5, 24].

Knijnenburg et al. [16] explored different ways in which users interact with an attribute-based recommender system. A post-experimental questionnaire is used to measure perceived understandability (using a 5-point Likert scale) of the interface. Increased domain knowledge resulted in higher understandability.

Schröder and Ghajargar [27] worked towards designing an understandable algorithmic experience in the music domain. Their small research-through-design experiment suggested that users comprehend the recommendations better when there is an easy path to accessing and understanding, and when the users are allowed to correct the system. Increased understanding could avoid a frustrated early majority.

Radensky et al. [25] analyzed how providing local, global, or explanations incorporating both, influenced user understanding of the system behavior. The results indicated that the combination of both local and global explanation types are more helpful for explaining how to improve recommendations. However, the global explanations performed better for efficiently identifying false positive and negative recommendations of the users.

Guesmi et al. [10] explored interactive explanations with a varying degree of detail using the intelligibility type categorization [18]. The authors included dimensions such as *What*, *What if*, *Why*, and *How* and conducted a user study to investigate the impact on the perception of users while presented with interactive explanations. They concluded that, while allowing users to personalize the explanation type was an integral feature, the more advanced explanation types achieved the highest levels of transparency and trust [10].

Our measurement of understandability is based on the distinction between objective and subjective understanding. Rong et al.

Table 1: Social choice-based aggregation strategies derived from [1, 8, 30]

Strategy	Type	Procedure
Additive Utilitarian (ADD)	Consensus	Recommends the item with the highest sum of all group members' ratings
Fairness (FAI)	Consensus	Ranking and recommending items according to how individuals choose them in turn
Approval Voting (APP)	Majority	Recommends the item with the highest number of ratings above a predefined threshold
Least Misery (LMS)	Borderline	Recommends the item which has the highest of all lowest per-item ratings
Most Pleasure (MPL)	Borderline	Recommends the item with the highest individual group member rating

[26] pointed out that measuring objective understanding is usually done by deploying proxy tasks to verify user comprehension of a certain computational model. Various methodologies have been explored in the literature and serve as foundation of the current work, including simulating model predictions [4, 31], evaluating counterfactual thinking [12, 31], and detecting mistakes [31]. Wang and Yin [31] showed that counterfactuals and visualizations of feature importance increased objective understanding. Cheng et al. [4] illustrated that white-box models increased the ability to simulate model behavior and concluded that interactivity is an important factor when it comes to improving objective understanding.

On the other hand, subjective understanding is usually measured by post-task questionnaires [16, 25]. These self-assessments are performed by Likert scale questions and statements such as “*I understand the decision algorithm*” and “*The explanation helps me to understand*”. Subjective understanding has been tested based on different implementations of explanations such as rule-based explanations [2], example- and counterfactual-based scenarios [31], or LIME and SHAP explanations [11]. The meta-analysis by Rong et al. [26] outlined that explanations of a computational model increased subjective understanding of users.

2.3 Social Choice-based Explanations

In the context of GRS, explanations have been used to assess a variety of factors such as consensus perception [1, 8] or privacy-preservation [21, 22]. Such explanations are natural language excerpts, typically outlining the underlying mechanism of the social choice-based aggregation strategy [13, 23]. User studies evaluated such explanations in terms of fairness perception, consensus perception and satisfaction with mixed results [1, 30]. Our control condition (no explanation) as well as the textual explanations follow the template presented in prior work.

2.4 Hypotheses

All in all, understandability of recommender systems and ML models in general has received ample attention. Additionally, explanations for GRS have been discussed and evaluated on the basis of a diverse list of factors. However, the concept of understandability of social choice-based explanations remains fairly unexplored. In this work, we aim to address this research gap by implementing the concept of understandability based on a varying degree of social choice-based explanations.

In light of the literature described in Section 2.2, we hypothesize that both objective and subjective understanding will increase when the user is presented with an explanation, with a bigger effect

when presented with a multimodal explanation. More formally, we define the following *hypotheses related to RQ1*:

H1a: Explanations will lead to a higher level of objective understanding of the underlying aggregation strategy, with a bigger effect for multimodal explanations.

H1b: Explanations will lead to a higher level of subjective understanding of the underlying aggregation strategy, with a bigger effect for multimodal explanations.

In Section 2.3, we discussed the mixed results related to the inclusion of textual social choice-based explanations on factors such as fairness or satisfaction [1, 30]. However, these user studies did find differences between strategies themselves. Based on these divergent outcomes among strategies, we formulate the following hypotheses:

H2a: The effect of the explanations on the level of objective understanding is moderated by the underlying social choice-based aggregation strategy used to derive the recommendation.

H2b: The effect of the explanations on the level of subjective understanding is moderated by the underlying social choice-based aggregation strategy used to derive the recommendation.

3 Methodology

In this section, we present the methodology and procedure for our user study. Our experiment was approved by the ethical committee of Maastricht University¹. The between-subject design, including research questions, variables and hypotheses, was preregistered on Open Science Framework (OSF).²

3.1 Materials

3.1.1 Group scenarios. Inspired by Barile et al. [1], our study makes use of a series of scenarios, each containing a hypothetical group which is being recommended a restaurant. Each scenario consists of a table in which a group of five members is presented alongside their fictitious ratings of 10 restaurants (on a scale from 1 to 5). Such groups can be built using distinct configurations, representing differing degrees of agreement among group members' preferences. Configurations outlined in previous work are *divergent* (high diversity), *uniform* (low diversity), *coalitional* (two distinct sub-groups)

¹<https://www.maastrichtuniversity.nl/ethical-review-committee-inner-city-faculties-ercic>

²https://osf.io/myx7p/?view_only=597bf08540a94dbd864b420ce3351a7d

Table 2: Textual explanation presented to participants in both *text_expl* and *graph_expl* groups

Strategy	Textual explanation
ADD	i_k has been recommended to the group since it achieves the highest total rating.
LMS	i_k has been recommended to the group since no group members has a real problem with it.
APP	i_k has been recommended to the group since it achieves the highest number of ratings which are above 3.

and *minority* (low diversity with the exception of one member) [1]. While a *uniform* configuration presents a simplistic scenario which needs no explanation, *divergent* and *coalitional* configurations lead to difficulty finding recommendations that satisfy group members. Thus, in this study, we strictly constructed fictitious groups based on the *minority* configuration. Each scenario was generated using the computational procedure outlined by Barile et al. [1]. However, since that the presented scenario considers a group who already used the system three times and receives a fourth recommendation (see Section 3.3), we also imposed the constraint that each scenario generated for a specific strategy would not have ties (items with the same group score) at the fourth interaction. Finally, the scenario is completed with anonymous items (named $Rest_i$), and random names for the group members (from a list of gender neutral names, to minimize the risk of possible biases).³

3.1.2 Aggregation strategies. Participants were assigned one of three social choice-based aggregation strategy and were only presented with scenarios in which that strategy was used to derive a group recommendation. To ensure a varied selection of strategies to derive a group recommendation, we included one of each category presented in Table 1. Practically speaking, each participant was shown recommendations made by either the Additive Utilitarian (ADD), Least Misery (LMS) or Approval Voting (APP) strategy, respectively covering consensus-based, borderline and majority-based aggregation categories (Table 1). The threshold rating for the APP strategy was set at 3, equal to 60% of the rating scale.

3.1.3 Explanations. Alongside a strategy, each participant was assigned one explanation type. In total, this study included three explanation types, (two explanation types supplemented with a control condition). The first explanation level (*no_expl*), the control condition, included no explanation. The participant was only presented with the group ratings and the output. The second level (*text_expl*) provided a simple textual social choice-based explanation, adopted from previous work [1] and summarized in Table 2. The third and final explanation type supplemented the textual explanation with a graphical representation of the procedure using a bar chart, an often used visualization for traditional user interfaces crafted for explaining recommendations [9, 29]. This multimodal explanation (*Graph_expl*) visualized the ratings of each group member, as well as the already chosen restaurants and the current recommendation (Figure 1). The ratings influencing the specific outcome were colored red. A final component differed among strategies. The ADD strategy included a line indicating the sum of all ratings per restaurant (Figure 1), while graphics for the LMS strategy showcased a line corresponding to the lowest per-item

rating. Finally, graphics visualizing the APP strategy included a line indicating how many ratings are above the set threshold.

3.2 Variables

Using the materials described in Section 3.1, our study consisted of two independent, between-subject variables.

- **Exp** (categorical, between-subject): each participant is randomly assigned an explanation modality:
 $Exp_i \in [no_expl, text_expl, graph_expl]$;
- **Agg** (categorical, between-subject): each participant is randomly assigned one social choice-based aggregation strategy:
 $Agg_i \in [ADD, LMS, APP]$;

Additionally, our experiment included two main dependent variables: *objective* and *subjective* understanding. Motivated by Wang and Yin [31], we designed a set of questions based on three aspects:

- Simulate model behavior: *Giving a new scenario, choose the right recommendation.*
- Counterfactual thinking: *Giving a new scenario and a given recommendation, pick the answer that results in an alteration of that output.*
- Error detection: *Giving a new scenario and a given recommendation, identify whether the presented recommendation is correct.*

For each participant, two tasks for each of these three aspects were presented. Each task was given a 0/1 value (0 = incorrect, 1 = correct). Per participant, these scores were summed up and divided by the number of tasks to obtain their accuracy rate:

- **Objective understanding** (continuous): The accuracy rate of each participant’s answers to the objective understanding questionnaire (six questions in total).

For measuring subjective understanding, participants were asked to rate two statements previously used by Wang and Yin [31]:

- “I understand how the model works to predict the best recommendation for the group.”
- “I can predict how the model will behave.”

These questions were presented two times, once at the end of the training step, and finally at the end of the experiment. From these, we compute the following variables:

- **Preliminary subjective understanding** (continuous): The average between both subjective understanding statements asked **before** the objective understanding measurement.
- **Final subjective understanding** (continuous): The average between both subjective understanding statements asked **after** the objective understanding measurement

In the analysis described in Section 4, we strictly made use of *Final subjective understanding*, measured at the end of the survey.

³The source code for generating the group configurations and explanations is available at the following link: https://anonymous.4open.science/r/Understanding_GRS-1E7F

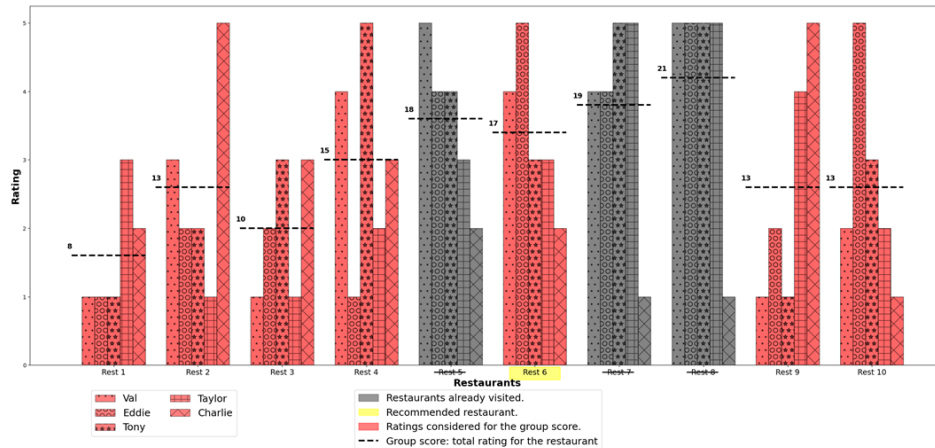


Figure 1: Graphical explanation for the Additive Utilitarian (ADD) strategy (example figure). Already visited restaurants are shown in grey. Potential recommendations are colored red. The output of the system is highlighted in yellow. Bar charts for ADD include a horizontal line indicating the sum of all per-item ratings.

In addition to the independent and dependent variables that we used for hypothesis testing, we collected age group and gender data to allow for a demographic description of our sample.

- **Age** (categorical). Participants will be able to select one of the options <18, 18-25, 26-35, 36-45, 46-55, >55, or *Prefer not to share this information*.
- **Gender** (categorical). Participants will be able to select one of the options *Female, Male, Nonbinary, A gender not listed here, or Prefer not to share this information*.

3.3 Procedure

The survey was executed with the Qualtrics tool⁴ and consisted of four stages (Figure 2).

Pre-survey. Participants were randomly assigned to one of the three social choice-based aggregation strategies and to one of the three explanation types. First, they agreed to an informed consent, and indicated their age group and gender. Afterwards, instructions were given detailing the procedure of the training phase.

Initial training. Each participant was presented with a series of six scenarios, consisting of one hypothetical group recommendation and configured on their assigned aggregation strategy and explanation type. An attention check was included in the fourth training scenario. No dependent variables are recorded during training. For each scenario, the participant followed a three-step procedure:

- (1) Make an initial recommendation decision regarding the restaurant recommendation.
- (2) Review the recommendation delivered by their assigned social choice-based strategy and, possibly, the explanation.
- (3) Make a final recommendation.

Each scenario was introduced by the following excerpt: *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*

⁴<https://www.qualtrics.com/>

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. Under the table, you also find the order of restaurants the group has already visited in the previous months. These restaurants are not an option anymore, as the group has already eaten there previously.

Understandability survey. After training, each participant was presented with the two subjective understanding statements. After rating their self-perceived understanding, six questions for testing objective understanding needed to be answered. Two questions are aimed at model simulation, two measured counterfactual thinking and finally, two questions geared towards error detection⁵. Additionally, a second attention check was presented on the page of the first counterfactual scenario. After the tasks geared towards objective understanding, the two subjective statements were presented again. For the final calculation of subjective understanding, we made use of these final two ratings. This part of the experiment ended with a final task: the participant are presented with the first training scenario, and asked to provide an explanation for the group, in their own words, on how the system derived a group recommendation. The responses gathered during this survey were used to derive the dependent variables.

Debriefing. The participants had the option to provide additional feedback in an open text field. Finally, a short debriefing message was showed, before redirecting them to the recruitment platform.

3.4 Sample Size Determination

We computed the minimum required sample size in a power analysis for the planned ANOVA tests (see Section 4.3) using *G*Power* [7]. To account for multiple hypothesis testing, we applied a Bonferroni correction by adjusting our significance threshold $\alpha = \frac{0.05}{2} = 0.025$.

⁵Full documentation including all questions, answer options and visualizations can be found in the OSF folder: https://osf.io/myx7p/?view_only=597bf08540a94dbd864b420ce3351a7d

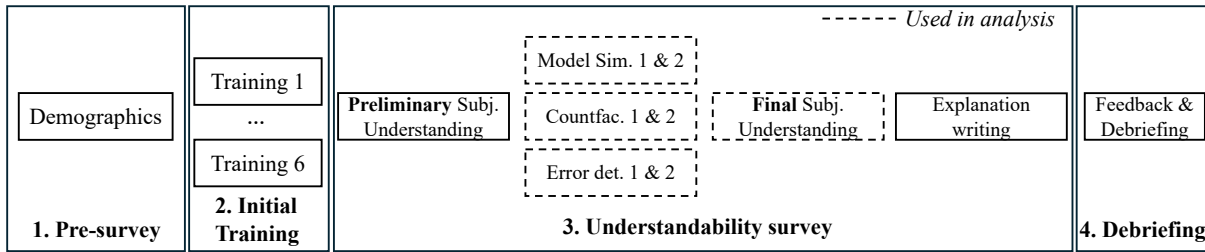


Figure 2: Survey components in the order as seen by participants. Survey questions in dotted lines were used to derive the dependent variables.

We calculated the sample size for an ANOVA (Fixed effect, special, main effects and interaction). We specified an effect size $f = 0.25$, $\alpha = 0.025$, a power of $(1 - \beta) = 0.85$, degrees of freedom numerator equal to 4, and $3 \times 3 = 9$ groups (i.e., 3 different aggregation strategies for 3 different explanation scenarios). These specifications aligned with previous works [4, 31]. These calculations resulted in a minimum required sample size of 257 participants. The analysis only included the **final** subjective understanding (one measurement per participant). As a result, we did not use a repeated measures ANOVA.

4 Results

In the following section, we report the results from our experiment. The full data is linked in the OSF folder⁶. We describe our sample, and present the overall understandability of the task, that helps us provide an indication of the complexity of the scenarios and illustrates the relationship between objective and subjective understanding. Afterwards, we address the research questions. Finally, we perform exploratory analyses regarding the different aggregation strategies and objective understanding tasks.

4.1 Participants

We recruited a total of 290 participants using the online participant pool *Prolific*.⁷ They were required to be proficient English speakers above 18 years of age. Each participant was allowed to participate in the study once, and received a reimbursement according to Prolific guidelines⁸, considering a hourly reward of £9. After removed 19 participants, due to failed attention checks, our final sample consisted of 271 participants: 30 participants for each of the nine conditions (characterized by one strategy and one explanation type), with the exception of the *APP text_expl* group ($n = 31$). Our sample was composed of 51% (139) female, 48% (131) male and 1% (1) nonbinary participants. Additionally, 42% (115) were between 26 and 35 years old, 30% (80) between 18 and 25, 16% (44) were between 36 and 45 years old, 7% (18) between 46 and 55 and 5% (14) indicated they were older than 56 years old.

⁶https://osf.io/myx7p/?view_only=597bf08540a94dbd864b420ce3351a7d

⁷<https://prolific.co>

⁸<https://www.prolific.com/resources/how-much-should-you-pay-research-participants>

4.2 Overall Understandability

Average understanding scores are visualized (by strategy and explanation type) in Figure 3. For objective understanding, we show accuracy scores which were calculated based on all six tasks. For subjective understanding, we the objective understanding survey (*Final subjective understanding*), which was recorded **after** the objective understanding tasks. Subjective understanding was not treated as repeated measure; *preliminary subjective understanding* was **not** used in this analysis. Overall, understandability was relatively high. The average objective understanding was 0.72 ($SD = 0.25$) on a scale of 0 to 1, while the average (final) subjective understanding (scale from 0 to 7) was 5.51 ($SD = 1.36$).

Table 3: Results of two two-way ANOVAs for dependent variables (DV) objective and subjective understanding. The independent variables were explanation type (Exp) and aggregation strategy (Agg)

Group	DV: Objective		DV: Subjective	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
Exp	1.490	0.227	2.274	0.105
Agg	4.081	0.018*	6.350	0.002**
Exp:Agg	0.438	0.781	0.246	0.912

4.3 Hypotheses Validation

Regarding the “RQ1: Do explanations increase users’ understanding?” we found no significant differences between the three explanation types, for both objective understanding (**H1a**; $F = 1.49$, $p = 0.23$ – see Table 3) and subjective understanding (**H1b**; $F = 2.27$, $p = 0.10$ – see Table 3). Thus, increasing detail, whether textual or multimodal, did not have an impact on user understanding as we did not find a significant difference compared to the control group (no explanation). Focusing on the RQ2: “Does the effect of explanations vary depending on the aggregation strategy?”, we also did not find significant interaction effects between aggregation strategies and explanation types regarding objective (**H2a**; $F = 0.44$, $p = 0.78$ – see Table 3) and subjective understanding (**H2b**; $F = 0.25$, $p = 0.91$ – see Table 3).

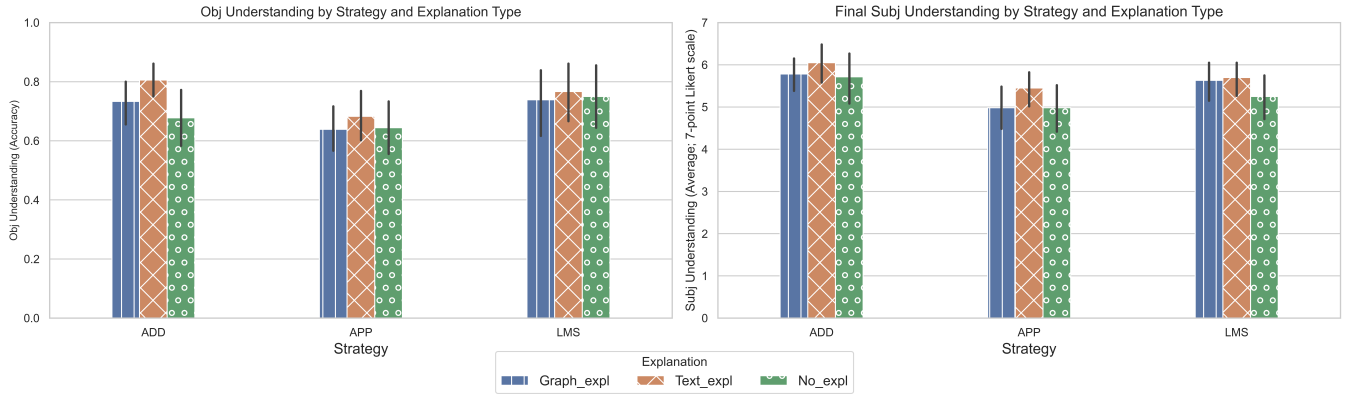


Figure 3: Objective (left, 0-1 scale) and subjective (right, 0-7 scale) understanding clustered by aggregation strategy (ADD = additive utilitarian, APP = approval voting, LMS = Least misery) and explanation modality (no_expl = control group, text_expl = textual, Graph_expl = multimodal)

4.4 Exploratory Analysis

Additional to testing our hypotheses, we performed some exploratory analyses to understand the impact of social choice-based aggregation strategies included in the experiment. Additionally, we attempted to explain some of the variance in objective understanding. As shown in Table 3, our analysis shows significant differences between aggregation strategies. Tukey pairwise post-hoc analysis revealed that participants assigned to LMS achieved higher objective understanding ($p_{adj} = 0.024$) compared to the participants assigned to APP. Additionally, participants exposed to the ADD strategy indicated a higher subjective understanding compared to those assigned to APP ($p_{adj} = 0.001$). Finally, we analyzed the different tasks making up our measurement of objective understanding: *model simulation*, *counterfactuals* and *error detection* (see Table 4). Overall, counterfactuals resulted in a lower average accuracy, implying a rather difficult task. Explanations, however, did not increase average counterfactual accuracy rates. On the other hand, error detection assignments resulted in a relatively high accuracy across the board.

Table 4: Average accuracy (and standard deviation) for objective understanding tasks; range between 0 and 1

	Model sim.	Counterfactual	Error detection
No_expl	0.73 (0.37)	0.52 (0.43)	0.82 (0.31)
Text_expl	0.81 (0.31)	0.54 (0.43)	0.91 (0.21)
Graph_expl	0.74 (0.39)	0.53 (0.41)	0.84 (0.25)

5 Discussion

In this study, we evaluated the effect of explanations for group recommendations on user understandability. Additionally, we looked at the outcomes based on the different social choice-based aggregation strategies used, and compared the accuracy rates of the three distinct assignment types making up our objective understanding measurement. In the following section, we discuss the results from

our experiments and formulate potential causes. Additionally, we argue in favor of measuring understanding using diverse assignments and formulate the implications of our study.

5.1 The Lack of Impact of Explanations

We did not find any significant differences between the three explanation types included in our experiment. This was the case for both objective and subjective understandability (Table 3). We did not find a negative impact of more complex explanations, which was reported in Kaur et al. [14]. This divergent outcome may be caused by the fact that the more complex explanation types analyzed by Kaur et al. [14] involved interactive elements, while our more complex explanation category did not. However, we can observe a trend in which textual explanations (*text_expl*) achieved slightly higher understanding (both objective and subjective) compared to the more complex, multimodal explanation (Figure 3).

Our results are in line with previous work analyzing overall effect of explanations in group recommendation scenarios: Barile et al. [1] did not find an impact of explanations on perceived consensus, perceived fairness and satisfaction. Furthermore, our outcomes are in line with a recent analysis on the understanding of machine learning models by Rong et al. [26], which reported mixed results on the impact of explanations on objective understanding.

The lack of significant effects between the different explanation conditions may depend on several reasons. A first possible motivation might be related to the tasks, that could have been too hard or abstract. However, the participants' understanding is relatively high across the board, including the control group (*no_expl*) – see Figure 3. Additionally, the total time spent by participants was higher than expected; if participants disengaged due to complexity, we would expect a lower accuracy and a shorter duration.

On the other hand, the used scenarios might have been too easy, leading to high understandability regardless of the presence of an explanation. We argue against this argument by comparing our subjective understanding measurement used in the analysis with the *preliminary subjective understanding*, measured between the training phase and objective understanding tasks: we observed a

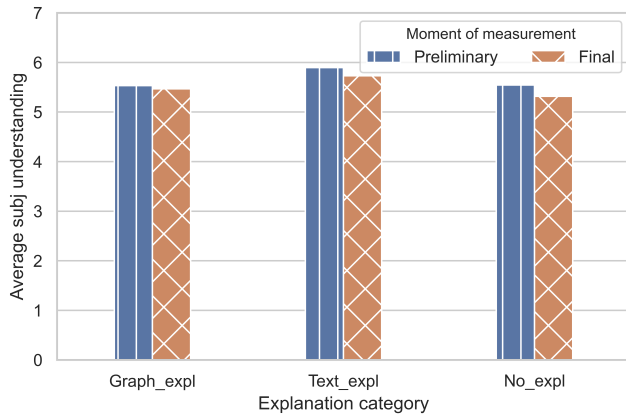


Figure 4: Subjective understanding (7-point Likert scale) measured before (preliminary) and after (final) the objective understanding tasks; by explanation modality (no_expl = control group, text_expl = textual, Graph_expl = multimodal)

decrease in self-perceived understanding after participants completed the six objective understanding tasks (Figure 4). This trend, combined with the fact that we did not receive participant feedback that implied a lack of difficulty, indicates that the group recommendation scenarios might not have been too simple.

A more likely motivation for our results can be found within *bounded rationality*. As discussed by Kaur et al. [14] in the context of interpreting machine learning models, humans are inclined to achieve *good enough* understanding (*satisficing*) as opposed to optimizing their decisions. In our study, it is possible that participants were satisfied with the group ratings and previously recommended restaurants, disregarding additional explanations. The relatively high understanding for the control group presents evidence in favor of bounded rationality (Figure 3). However, future work needs to unpack the individual degree of satisficing for group recommendation further, for example by asking participants to indicate which information they used to make their decision. Additionally, the spectrum of presented information, both within the scenario and explanation, can be expanded to pinpoint when participants are provided enough information for a satisficing outcome.

5.2 Performance on Objective Understanding Tasks

In Section 4.4, we separately presented the average accuracy rates for each of the three objective understanding tasks to investigate the variance within the objective understanding measurement. Error detection resulted in the highest overall accuracy (Table 4). However, these questions were binary (correct/incorrect), while the other two comprised four or more options. Additionally, we found that, compared to both model simulation and error detection, counterfactuals turned out to be a difficult task. A potential cause for this lower accuracy may be the fact that our counterfactual assignment required participants to perform multiple model

simulations in sequence.⁹ This is reflected in the timestamp of the assignments. On average, it took participants longer to complete a counterfactual assignment compared to the other two tasks (Table 5). This post-hoc analysis suggests the value of measuring objective understanding based on the performance on a variety of tasks, as different types of tasks lead to divergent results.

Table 5: Average time (and standard deviation) in seconds for completing a single assignment; per objective assignment task

	Model sim.	Counterfactual	Error detection
No_expl	51.5 (40.1)	113.2 (85.9)	36.8 (26.4)
Text_expl	58.5 (114.2)	114.3 (74.0)	39.0 (25.5)
Graph_expl	48.4 (35.7)	111.6 (94.9)	39.5 (32.2)

5.3 Implications

Our findings have clear implications which need to be considered by system designers and other practitioners. First, the results presented in Section 4.4 show that the social choice-based aggregation strategy implemented within a GRS may influence user understandability down the line. When designing a GRS, one should be mindful that these methodological decisions should be weighed not only in terms of fairness, consensus and satisfaction, but also in terms of user understanding. If user understanding and explainability are primary concerns, opting for a more straightforward aggregation strategy in the design phase is beneficial.

Second, our experiment provides an indication to system designers looking to improve user understandability by adapting certain elements in the interface. Against intuition, textual explanations were not useful in improving user understanding. Similarly, adding visual elements to the textual explanation did not increase user understanding either. Thus, practitioners need to look towards other procedures (other than explanations) to improve user understanding of GRS. The relatively high understanding achieved by the control group implies that the presentation of previous output alongside group context might already provide sufficient information to users.

All in all, our results imply that increased detail in explanations might become redundant when users receive a satisficing amount of information already. This information and explainability can be derived from the implemented methodology (e.g. strategy) and context clues such as group ratings and previous output. System designers and other practitioners should keep in mind that simply increasing the details presented in explanations will not necessarily translate into improved user understanding. For GRS specifically, opting for social choice-based aggregation strategies such as *Least Misery* or *Additive Utilitarian*, as opposed to *Approval Voting*, might be beneficial to ensure user understanding.

⁹All scenarios and corresponding tasks are found in the OSF folder: https://osf.io/myx7p/?view_only=597bf08540a94dbd864b420ce3351a7d

5.4 Limitations

We identified several limitations that may have had an impact on our results. First, we focused on the minority group configuration; this limits the generalizability of our results. As discussed in Barile et al. [1], different group configurations may lead to different evaluations in terms of satisfaction and perception of fairness and consensus. Future works are necessary to evaluate the possibility of different impact of explanations for different group configuration.

Additionally, we did not ask participants to evaluate their AI literacy. Kaur et al. [14] highlighted that Machine Learning practitioners were faster but less accurate when having access to interpretability tools. Future research could include self-rated measurements of AI literacy and compare practitioners and novices in terms of understandability of GRS.

Finally, we presented group scenarios involving the recommendation of unnamed restaurants. However, high investment domains such as tourism could have a higher need for explainability. Future work could make the comparison between low and high investment domains in group scenarios to investigate whether domain-specific factors influence the impact of explanations.

6 Conclusion

In this study, we presented a randomized controlled trial to analyze whether increasing detail of social choice-based explanations for GRS improved understandability. Our setup consisted of two between-subject factors: (i) the explanation modality, and (ii) the social choice-based aggregation strategy used to generate the group recommendations. We constructed two ANOVA models using objective and subjective understanding as dependent variables. However, we did not find significant differences between explanation types (no explanation, textual and multimodal explanation). Hence we outlined potential causes for this result rooted in bounded rationality: the group context and previous recommendations may suffice to provide participants with a satisficing level of understanding.

Furthermore, we conducted several post-hoc analyses to explain some of the variance found within our understanding measurements. We found that the methodological choice regarding aggregation strategy can impact understandability down the line. Additionally, our results indicate that counterfactuals tasks were more difficult compared to model simulation and error detection. We conclude that there is value in measuring understanding based on a multitude of assignments.

Finally, we discussed some implications of our work. Besides factors such as fairness and satisfaction, the methodological choice of aggregation strategy needs to be weighed in terms of understandability. Additionally, system designers looking to adapt user interfaces need to be mindful of the fact that additional elements do not necessarily translate to improved understandability.

Acknowledgments

This work was supported by the Faculty of Science and Engineering at Maastricht University (2024 FSE Starter Grant). R. Yera Toledo is supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement number 101106164.

References

- [1] Francesco Barile, Tim Draws, Oana Inel, Alisa Rieger, Shabnam Najafian, Amir Ebrahimi Fard, Rishav Hada, and Nava Tintarev. 2023. Evaluating explainable social choice-based aggregation strategies for group recommendation. *User Modeling and User-Adapted Interaction* (2023), 1–58.
- [2] Zana Bućinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.
- [3] Lei Chen, Jie Cao, Huanhuan Chen, Weichao Liang, Haicheng Tao, and Guixiang Zhu. 2021. Attentive multi-task learning for group itinerary recommendation. *Knowl. Inf. Syst.* 63, 7 (2021), 1687–1716. doi:10.1007/s10115-021-01567-3
- [4] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
- [5] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [6] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*. Springer, 449–466.
- [7] Franz Faul, Edgar Erdfelder, Albert Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (2007), 175–191. doi:10.3758/BF03193146
- [8] Alexander Felfernig, Ludovico Boratto, Martin Stettinger, and Marko Tkalčić. 2018. Explanations for Groups. In *Group Recommender Systems*. Springer, 105–126.
- [9] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.
- [10] Mouadh Guesmi, Mohamed Amine Chatti, Shobeh Joarder, Qurat Ul Ain, Rawaa Alatrash, Clara Siepman, and Tannaz Vahidi. 2024. Interactive explanation with varying level of details in an explainable scientific literature recommender system. *International Journal of Human-Computer Interaction* 40, 22 (2024), 7248–7269.
- [11] Sophia Hadash, Martijn C Willemsen, Chris Snijders, and Wijnand A IJsselstein. 2022. Improving understandability of feature contributions in model-agnostic explainable AI tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–9.
- [12] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
- [13] Öykü Kapcak, Simone Spagnoli, Vincent Robbmond, Soumitri Vadali, Shabnam Najafian, and Nava Tintarev. 2018. Tourexplain: A crowdsourcing pipeline for generating explanations for groups of tourists. In *Workshop on Recommenders in Tourism-co-located with the 12th ACM Conference on Recommender Systems (RecSys 2018)*. CEUR, 33–36.
- [14] Harmanpreet Kaur, Matthew R Conrad, Davis Rule, Cliff Lampe, and Eric Gilbert. 2024. Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–34.
- [15] J.S. Kelly. 2013. *Social Choice Theory: An Introduction*. Springer Berlin Heidelberg.
- [16] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. 2011. 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*. 141–148.
- [17] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).
- [18] Brian Y Lim and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*. 195–204.
- [19] Judith Masthoff. 2015. Group recommender systems: aggregation, satisfaction and group attributes. In *recommender systems handbook*. Springer, 743–776.
- [20] Judith Masthoff and Amra Delić. 2022. *Group Recommender Systems: Beyond Preference Aggregation*. Springer US, New York, NY, 381–420. doi:10.1007/978-1-0716-2197-4_10
- [21] Shabnam Najafian, Amra Delic, Marko Tkalčić, and Nava Tintarev. 2021. Factors Influencing Privacy Concern for Explanations of Group Recommendation. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 14–23.
- [22] Shabnam Najafian, Tim Draws, Francesco Barile, Marko Tkalčić, Jie Yang, and Nava Tintarev. 2021. Exploring User Concerns about Disclosing Location and Emotion Information in Group Recommendations. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. 155–164.

- [23] Shabnam Najafian and Nava Tintarev. 2018. Generating Consensus Explanations for Group Recommendations: an exploratory study. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) (UMAP '18). Association for Computing Machinery, 6 pages. doi:10.1145/3213586.3225231
- [24] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.
- [25] Marissa Radensky, Doug Downey, Kyle Lo, Zoran Popovic, and Daniel S Weld. 2022. Exploring the role of local and global explanations in recommender systems. In *Chi conference on human factors in computing systems extended abstracts*. 1–7.
- [26] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2024. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE transactions on pattern analysis and machine intelligence* 46, 4 (2024), 2104–2122.
- [27] Anna Marie Schröder and Maliheh Ghajargar. 2021. Unboxing the algorithm: Designing an understandable algorithmic experience in music recommender systems. In *Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2021)*, September 25th, 2021, co-located with the 15th ACM Conference on Recommender Systems.
- [28] Christophe Senot, Dimitre Kostadinov, Makram Bouzid, Jérôme Picault, Armen Aghasaryan, and Cédric Bernier. 2010. Analysis of strategies for building group profiles. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 40–51.
- [29] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.
- [30] Thi Ngoc Trang Tran, Müslüm Atas, Alexander Felfernig, Viet Man Le, Ralph Samer, and Martin Stettinger. 2019. Towards social choice-based explanations in group recommender systems. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 13–21.
- [31] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 318–328.