Research Article

Diana C. Hernandez-Bocanegra* and Jürgen Ziegler

Explaining Review-Based Recommendations: Effects of Profile Transparency, Presentation Style and User Characteristics

https://doi.org/10.1515/icom-2020-0021

Abstract: Providing explanations based on user reviews in recommender systems (RS) may increase users' perception of transparency or effectiveness. However, little is known about how these explanations should be presented to users, or which types of user interface components should be included in explanations, in order to increase both their comprehensibility and acceptance. To investigate such matters, we conducted two experiments and evaluated the differences in users' perception when providing information about their own profiles, in addition to a summarized view on the opinions of other customers about the recommended hotel. Additionally, we also aimed to test the effect of different display styles (bar chart and table) on the perception of review-based explanations for recommended hotels, as well as how useful users find different explanatory interface components. Our results suggest that the perception of an RS and its explanations given profile transparency and different presentation styles, may vary depending on individual differences on user characteristics, such as decision-making styles, social awareness, or visualization familiarity.

Keywords: Recommender systems, user study, explanations

1 Introduction

Providing explanations of the rationale behind a recommendation can bring several benefits to recommender systems (RS), by increasing users' perception of transparency (the system explains how it works), effectiveness (user can make good decisions), and trust [38]. Accordingly, research on explainable RS aims to establish methods and models which allow for generating relevant recommendations to users, while providing them with the reasons why an item is recommended. In this regard, various explainable recommendation methods have been proposed, mainly based on collaborative filtering (CF) and content based (CB) methods. CF explanatory models allow to generate explanations based on relevant users or items, e. g. nearest-neighbor style explanations as proposed by Herlocker et al. [21], while CB models facilitate the generation of feature-based explanations by providing users with product features that match their preferences, as proposed, for example, by Vig et al. [44]. On the other hand, matrix factorization (MF) methods, a particular case of CF, allow to generate recommendations by obtaining latent representations of items and users (latent features), which represent a major challenge when it comes to explaining to users how the algorithm works, or why the item is recommended, compared to more intuitive neighbor-style CF or CB methods. In this respect, MF explanatory methods have been proposed [53, 46], to integrate external sources of information (e. g. user generated reviews) in order to make sense – to some extent – of latent features, for example, by aligning them with explicit features drawn from reviews. In this regard, the interest in the use of user reviews in explanation methods has increased recently, given the richness of information reported on diverse aspects, which cannot be deduced from the overall item ratings, and that could be beneficial to both recommendation and explanation processes. Particularly, review-based explanatory approaches usually involve the detection and aggregation of both positive and negative opinions regarding different aspects or features of items, the selection of helpful reviews or excerpts from them that could work as explanations, or the generation of verbal summaries of items' evaluation by users. The above entails a potential for the generation of a diverse range of explanation types, consisting of arguments with different levels of detail, and portrayed in different presentation styles. Nevertheless, little is known

***Corresponding author: Diana C. Hernandez-Bocanegra,** University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science Duisburg, Germany, e-mail: diana.hernandez-bocanegra@uni-due.de, ORCID: https://orcid.org/0000-0002-1773-2633
Jürgen Ziegler, University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science Duisburg, Germany, e-mail: juergen.ziegler@@uni-due.de

about how best to convey explanatory information, in order to meet different explanatory aims like transparency, effectiveness, satisfaction or trust. This is largely due to the predominant lack of evaluation by users in works that propose new explanation methods, as noted by Nunes and Jannach [30]. In this regard, evaluating explanations from the users' perspective can contribute to better explanation design, which can significantly impact users' perception of a RS. Such perspective could contribute to answering questions that remain open, for example, to what extent the format or presentation style influences the perception of an explanation, or what are the components of an explanation that most contribute to its perceived usefulness.

As outlined by Nunes and Jannach [30], explanations may involve the following types of user interface components: input parameters, knowledge base (background or user knowledge), decision inference process (data or rationale of the inference method), and decision output. As for the knowledge base components, and depending on the method used to generate the recommendations, the explanations may reflect either the quality and the properties of the items or the matching between the recommended items and the users' preferences. In regard to the latter, a target user might benefit from knowing which of her/his performed interactions with the system have an effect on a current recommendation [3], as well as knowing how well their preferences match the justifications provided by the system, which can contribute to the acceptance of its recommendations (provided that there is actually a fit) [19]. Although the effects of providing a view on user profiles in CB or item-based CF methods has been explored before (e. g. [3, 39, 18]), such effects have not been fully addressed in review-based explanations, where information on users' profile is often omitted and used only implicitly, e. g. to filter and sort lists of relevant features, as in [28]. In consequence, we aim to address in this article the following question:

RQ1: How does including the information about *user preferences* influence the perception of a review-based RS and its explanations?

Specifically, information on user preferences refers – in the scope of this article – to a list of the relevant inferred aspects and their relevance score, which are also calculated based on the users' own reviews. Additionally, the above mentioned perception is addressed in this article in regard to explanation quality, and to the perception of the overall system in terms of: transparency, effectiveness, efficiency and trust. Likewise, we address the perception of users in regard to specific aspects of the explanations, i. e.: confidence, transparency, satisfaction, persuasiveness, effectiveness, efficiency and easiness to understand of the explanations.

In regard to the interface component "decision inference process" [30], the RS may provide details on the recommendation process, or on the data used for it. In the former case, for example, CF methods favore the generation of concise reporting of recommendation process e. g. "We suggest this option because similar users liked it.". While further algorithm details are often omitted, providing only information about the input and the output of the process might also be beneficial to users, in the case of black-box models [21]. In consequence, various explanatory methods provide information on the data used during the process, like ratings for similar items or ratings by similar users in CF models, or specifications of items in CB methods. However, when additional sources of information are taken into account, as in the case of review-based methods, users are often not informed of the type of the data utilized during the process, for example, whether the user's preferences have been calculated exclusively based on ratings, with information extracted from reviews, or based on other previous interaction with the system. Consequently, we aimed to test to what extent providing such information explicitly is considered useful by the users, more formally:

RQ2: How useful is it for users, during their evaluation of different purchase or booking options, to be informed about the origin of the data used by a review-based recommendation process?

In particular, within the scope of this article, we address how useful it is for users to read that the recommendation is based on the opinions of other customers, as well as their own comments.

The taxonomy of explanations proposed by Nunes and Jannach [30] also involves a category for presentation format, which includes: natural language (e. g. canned text, template-based, structured language), visualization, or other media formats, such as audio. While some of the existing review-based explanatory methods apply at least one of such formats, a user-centered evaluation in which the different formats are comprehensively compared is still necessary. For example, it is not yet clear whether users have a better perception of explanations consisting of aggregate information represented in tabular data, compared to those containing a graphical representation of such information. In this regard, according to Blair [4], visual arguments – defined as a combination of visual and verbal communication – may, in addition to representing propositional content, have a greater "rhetorical power potential" than verbal arguments, due (among others) to their greater immediacy. However, users with lower visual abilities might benefit less from a presentation based on images or graphics [34, 23]. Additionally, while a representation using tables has been recommended to display small data sets [16, 43], if providing accurate numerical values of proportions is not the main objective, tables seem to be less useful than graphics as a means of displaying information [36]. Nevertheless, although the findings in such direction in the field of information visualization, little is known about such effects in relation to explanations. Consequently, we aim to address in this article the following question:

RQ3: How does the display *style* of explanation (using a table or a bar chart) influence the perception of the variables of interest?

Here, the perception of the variables of interest refers to the perception of the overall system and of the specific aspects of explanations, in the same way as described for RQ1.

As it has been shown that individual user characteristics can lead to different perceptions of a RS [25, 50], we assumed that this would also be the case for explanations, as discussed by [2, 26, 22]. Consequently, and similar to Hernandez-Bocanegra et al. [22], we also aimed to test the effect that user characteristics may have on the perception of the explanations, in particular regarding decision making style (rational and intuitive) [20] and the ability of the user to take into account the views of others (social awareness) [17]. Additionally, we also aimed to test the influence that visual familiarization may have on explanations perception, as addressed by Kouki et al. [26]. Consequently:

RQ4: Do individual differences in visual familiarity, social awareness or decision making styles influence the perception of our proposed explanations design?

Here, as with previous RQs, the perception of our explanations designed is addressed in terms of system perception as well as of perception of specific aspects of explanations.

In order to address these questions, we conducted a user study to test the perception of explanations based on user opinions in the hotel domain, given different display styles and whether or not user profile information is shown. The perception was assessed regarding two levels: 1) overall system and explanation quality, and 2) perception of specific aspects of explanations.

The contributions of this paper can be summarized as follows:

- We evaluated the effect of different presentation styles, namely tabulated data or bar charts. Comparisons were conducted both between groups and within participants.
- We also evaluated the effect of providing user profile information as part of explanations, with a display that contains no information regarding user preferences.
- Furthermore, we analyzed the usefulness perceived by users of the different user interface components included in explanations.

The remainder of this paper continues as follows: We discuss related work in Section 2, and the specifics of our explanation design in Section 3. In Section 4, we present methods and results of experiment 1, while details and results of experiment 2 are provided in Section 5. Discussion of both studies and limitations are included in Section 6. Finally, we address future work in Section 7.

2 Related Work

Traditionally, many approaches to explaining the products or services suggested by an RS have been based on ratings provided by users (CF methods) or properties of the recommended items (CB methods). In the former, explanations are often provided in a nearest-neighbor style (e. g. “Your neighbors’ ratings for this movie” [21]), while the latter approach enables the generation of feature-based explanations, that inform users about item properties that may match user preferences, as in [44]. On the other hand, there has recently been increased interest in exploiting alternative sources of information to improve the performance and explainability of RS, particularly the use of user reviews, given the wealth of detailed reports on the positive and negative aspects of an item, information that is often difficult to understand from the general ratings given by users.

Review-based methods enable the generation of the following types of explanations:

1) A verbal summarization of review findings, i. e. statements generated in natural language representing a summarized version of the original content extracted from reviews, e. g. [6, 12], who proposed methods based on natural language generation (NLG) techniques.

2) A selection of helpful reviews, or excerpts from them, that might be relevant to users, as proposed by [9], who used a deep learning model and word embeddings to jointly learn user preferences and item properties, and an attention mechanism to detect features that are of most interest to the target user.

3) A summarized view of pros and cons on specific item aspects reported by other users. Here, topic modelling and aspect-based sentiment analysis are usually used to detect the sentiment polarity towards item aspects or features addressed in reviews, as in [49, 53, 14]. Subsequently, such information can be integrated into RS algorithms such as matrix or tensor factorization, as in [53, 1, 46] in order to generate both recommendation and aspect-based explanations.

In particular, our explanation design proposal and subsequent user study is within the third approach, and is particularly related to the MF model proposed by Zhang et al. [53], since it facilitates the consolidation of statistical information on users' opinions (which can be provided using different presentation styles), as well as their alignment with the user's profile, which is fundamental to our research questions. This model allows the generation of both recommendations and explanations, based on the alignment of 1) latent representations of items and user preferences, and 2) explicit features obtained from reviews. Here, in addition to the rating matrix used in traditional MF, two additional matrices are calculated: a user preference matrix (which indicates how many times a user addressed a feature in their reviews), and an item quality matrix (which indicates how many positive and negative comments were reported in relation to an item). This input information is then used as the basis for our proposed explanation and subsequent user study.

2.1 User Profile Transparency

In regard to providing information on user profile as part of RS explanations, Bilgic and Mooney [3] proposed and tested an influence-based style for explanations in the movies domain, in which the system presents items that had the greatest impact on the recommendation, as well as the ratings that the user has given to those items. They found that such explanations enabled participants to more accurately predict user's satisfaction with the item, compared to a histogram of the user's neighbors' ratings, an explanation style that was found by Herlocker et al. [21] as the best performing among a group of explanations for CF methods, in terms of how compelling they were to study participants.

On the other hand, and using a CB method, Tintarev and Masthoff [39] compared non-personalized verbal explanations with personalized ones, in which, in addition to providing information about the properties of the articles, a sentence was included indicating how these properties related to the user's preferences. According to their findings, personalized explanations were not regarded as more effective than their counter non-personalized part. Here, and similar to [3] the effectiveness was measured based on the difference between the rating that the user would give to an item after reading the explanation, and the one given once the item has been tried. According to authors, the detrimental effect of personalized explanations on effectiveness might be due to users' expectations of preference fit that were not fulfilled once the item was tried.

Additionally, Gedikli et al. [18] compared the perception of users regarding different types of explanations provided by CF and CB recommenders in the movies domain. Their proposed personalized explanations based on clouds showed tags in different colors, depending on the sentiment previously expressed by the target user, regarding different colors (positive: blue, negative: red, neutral: gray). In line with Tintarev and Masthoff [39], they found that a non-personalized tag cloud (all tags in the same color) was slightly more effective than the personalized tag cloud. However, the personalized tag cloud was perceived better by users in terms of transparency.

Disclosure of information used during the recommendation process (e. g. user profile) as part of explanations may facilitate users in identifying and correcting erroneous inferences made by a RS [38]. In this direction, proposed work on scrutable RS seeks to enable and to leverage user control on users' own profile, which in turn may facilitate the generation of new and more accurate recommendations that fit better the real preferences of users. For example, Wasinger et al. [47] implemented a system to recommend restaurant meals based on a scrutable user model, where users could check and adjust their preferences regarding food ingredients to improve recommendations. A user study was conducted to test the application, and noted that users found it easy to understand why certain foods were recommended, by using the customization feature to adjust their preferences.

In regard to review-based methods, Chen and Wang [10] proposed a text-based explanation design that combines both summarization of item opinions as well as item specifications, and that provides a tradeoff view of properties, that allows the direct comparison of different recommended items. They found that a mixed explanatory view containing opinions and specifications was perceived more positively by users, than explanations consisting of only one of such components at the time. However, in contrast to our approach, the selected specifications correspond to explicit elicited preferences, and not to preferences detected from previous reviews written by the user.

On the other hand, Muhammad et al. [28] tested the users' perception of a series of review-based RS explanations in the hotels review. Here, item quality and user preferences are both extracted from reviews and used to generate both recommendations and explanations. However, user preferences are only used implicitly to select, show and sort a subset of the features in explanations, without any mention of such details to users.

In summary, while the effect of presenting explicit information on user profiles as part of explanations of CF and CB methods has already been addressed to some extent, the questions of how such information influences the perception of review-based SR and how such information should be presented remain open.

2.2 Decision-Making Process Transparency

In regard to informing users about the decision inference process, the RS may provide details on the recommendation process, or on the data used for it. Accordingly, Herlocker et al. [21] proposed an explanatory model based on the user's conceptual model of the recommendation process. In a white box conceptual model, users are provided with details of the different steps of the conceptual model of the system operation, e. g. user enters ratings, then system locates similar users, then neighbors' ratings are combined to provide recommendations. In a black box model, however, it may not be practical or even possible to convey details regarding the conceptual model of the system to users, which is actually the case of MF models and their latent features. Herlocker et al. [21] argues that any white box could be regarded as a black box if only information about the input and the output is provided, which could also be beneficial for users.

In regard to the source and type of input used in the process, the presentation of such elements in many of the CF and CB neighbor-style approaches is simpler and self-explanatory, compared to more complex approaches that integrate alternative sources (e. g. reviews) to latent features models as MF, where not only the steps of the process are hard to convey to users, but also the nature of the data used as input. Consequently, most current review-based explanatory approaches omit any mention of the origin of the data, particularly when explaining the inferred user profile, which may make it more difficult to understand compared to item-based explanatory information. Therefore, in addition to assessing how the different ways of presenting the input data might influence the users' perception, in this article we intend to examine also the potential usefulness of explanatory statements on the data origin, as part of review-based explanations, e. g. "based on how often you mentioned features in your own comments before".

2.3 Presentation Format

According to the taxonomy of explanations proposed by Nunes and Jannach [30], explanations could be classified by their presentation format as: natural language (e. g. canned text, template-based, structured language), visualization, or other media formats, such as audio. Regarding review-based explanations, Zhang et al. [53] proposed brief template-based statements to provide information on relevant features (e. g. "You might be interested in [feature], on which this product performs well", although the underlying method allows to generate more detailed explanations, that could also be provided visually using graphs, as elaborated in further sections of the present work. Furthermore, a distinction can also be made between verbal explanations that also provide numerical or statistical information and those that comprise strictly verbal statements. In this respect, Hernandez-Bocanegra et al. [22] compared different types of verbal explanations in the hotel domain, and found that users perceived a higher explanation quality when an aggregated view of positive and negative opinions using percentages was provided, compared with a verbal summary of the opinions that did not provide any percentage, inspired by the abstractive summarization proposed by Costa et al. [12]; furthermore, a greater perceived transparency was reported for explanations with the aggregated view using percentages of opinions, compared to explanations that only provided a useful review, as proposed by Chen et al. [9].

In regard to presentation styles based on visualization techniques applied to review-based RS, Muhammad et al. [28] proposed a summary of the positive and negative opinions on different aspects using bar charts, while Wu and Ester [49] proposed to depict such type of information as word clouds or radar charts. Although bar charts reflecting positive and negative views might be perceived as more informative and attractive than brief template-based textual explanations, easier to interpret than challenging radar charts, or quicker to process than tabulated data, it remains unclear to what extent the presentation format influences the perception of RS and its explanations. In this regard, Kouki et al. [26] proposed a series of explanations based on a hybrid RS in the music domain, and tested, among others, the influence that the presentation format could have on users' perception. In this case, the authors found that textual explanations were perceived as more persuasive than the explanations provided using a visual format; however, users with greater visual familiarity perceived one of the visual format explanations more positively (a Venn diagram). Consequently, we aimed to investigate whether such an effect is also observed in the case of review-based explanations.

Particularly, in the present work, we set our focus on two formats: bar chart and table. Bar charts are recommended to facilitate a direct and quick comparison of values between different categories or items, contrary to alternatives like pie charts, or bubbles, where additional cognitive efforts would be needed to accurately calculate the differences in values across categories, in our case, the different aspects of the items. Likewise, word clouds imply a presentation challenge, since we are willing not only to represent the amount of comments (which could be reflected by font size), but also polarity, which would require using separate clouds for the positive and negative aspects, or showing a single predominant sentiment per aspect in a single cloud, thus obscuring the information about the less predominant polarity. On the other hand, while the use of tables has been recommended to display small data sets (less than 20 data points) [16, 43], when providing exact numbers or proportions is not the main objective, tables seem to be less useful than graphics [36]. As indicated previously in the case of verbal explanations, users benefited from a view that provides percentages of positive and negative opinions, suggesting that percentages may serve as anchors to convey more compelling information in explanations, compared to purely verbal statements. In this sense, when motivation or ability is lacking, the effortless use of cues such as numerical anchors can lead to changes in attitude [32, 48], which in turn influence judgments and decision making (anchoring effect). Thus, even when the values of the proportions of the opinions included in the two types of explanations (table or chart) are the same, a different representation of them might lead to differences in explanation perception, which we set out to test in the user study.

2.4 User Characteristics

Beyond the explanations' content itself, a number of user characteristics also contribute to differences in the overall perception of RS. Models proposed by Knijnenburg et al. [25] and Xiao and Benbasat [50] argue that perception of the interaction with the system usually depends on personal characteristics, like demographics and domain knowledge. Furthermore, Berkovsky et al. [2] evaluated how differences in the perception of trust might reflect differences in users' personality traits, given different types of explanations provided in the movies domain. To this end, they used participants' scores of the Big-Five personality traits (openness, conscientiousness, extraversion, agreeableness and neuroticism) [13, 40], and compared persuasive explanations (e. g. "highest grossing movie of all times"), personalized CF-based explanations (e. g. "because you liked X") and IMDb voting-based explanations (e. g. "Average rating *n*, Number of votes *m*"). Among their findings, authors reported that people with higher disposition to agree perceived more positively the voted-based explanations, compared to personalized explanations, seemingly to a higher disposition to accept others' opinions rather than impose their own preferences. Furthermore, they found that people with higher levels of neuroticism perceived better the voted-based explanations compared to the persuasive ones, possibly due to a perception of higher reliability of explicit voting numbers, which could presumably reduce the risk of frustration of a person with high levels of neuroticism.

Similarly, Kouki et al. [26] explored the influence of personality traits on users' explanation preferences regarding perceived accuracy and perceived novelty of recommendations, in the music domain. They compared different types of textual explanations, and found that participants with higher levels of neuroticism preferred item-based explanations (e. g. "people who listen to your profile item X also listen to Y") whereas popularity-based explanations (e. g. "X is a very popular in the last.fm database with *n* million listeners and *m* million playcounts") were preferred by users with lower levels of neuroticism, the latter in contrast to the opposite finding reported by [2], regarding trust perception.

Despite the usefulness of using the Big Five personality traits to better understand individual differences in RS perception and its explanations, we decided to address other types of user characteristics, which are more related to how users process information when making decisions, noting that supporting this process is precisely the goal of recommendation systems. Particularly, individual differences in decision-making styles are determined to a greater extent by preferences and abilities to process available information [15]. Two main aspects provide a basis to describe the differences in decision styles: information use (amount of information used during the process) and focus (alternatives addressed) [15]. "Good enough" information might be sufficient for some people, whereas others prefer to obtain and address all relevant information, in order to minimize risks or negative consequences of decisions. To the former, even when more information may be available, it is not necessary or worth taking the time to review it. Hamilton et al. [20] defines rational and intuitive decision styles similarly to the cognitive styles of Pacini and Epstein [31], with the latter having a more general scope to describe manners of solving problems. Thus, decision making styles are defined by Hamilton et al. [20] as a "habit-based propensity" to exhaustively search for information and to systematically evaluate possible alternatives (rational style), or to use of a quick process based on hunches and feelings (intuitive style).

Additionally, we were interested in another factor that may influence the way users perceive explanations: the extent to which they are able to adopt the perspective of others when making decisions. The rationale for this interest stems from the tendency of individuals to adjust their own opinions using those of others, while choosing between various alternatives [35], which may even be beneficial [51]. Particularly, individuals with greater perspective-taking skills tend to understand the views of others better [8, 5], skills that are also characterized as "social awareness" [17].

Previous work by Hernandez-Bocanegra et al. [22] evaluated the influence of decision-making styles and social awareness on the perception of review-based argumentative explanations, and suggested that social awareness might have an effect on both transparency and trust in review-based RS. Their results indicated that users with a greater willingness to listen and take into account the opinions of others valued their proposed explanations better than users who tend to listen less to others. On the other hand, contrary to the authors' expectations, the more detailed explanations summarized by the system were not preferred by the more rational users, apparently because the additional information generated by the system is not perceived as more satisfactory than the possibility of reading directly the comments written by the users.

Finally, since we aimed to compare differences in perception of explanations consisting of different visual representations, we also considered a factor that is related to visual abilities, in particular the extent to which a user is familiar with graphical or tabular representations of information. Visualization familiarity may also influence the perception of explanations provided using images or graphs, as found by Kouki et al. [26] in the music domain. Here, authors found that textual explanations were perceived as more persuasive than the explanations provided using a visual format; however, users with greater visual familiarity perceived one of the visual format explanations more positively (in particular a Venn diagram).

3 Explanation Design

In the context of RS, review-based argumentative explanations could be understood as a set of propositions, summarizing positive points reported by other users on specific aspects, that support the claim that an article can be recommended to a user. In this respect, information extracted from user reviews could be consolidated and provided as propositions, which would constitute the *backing* component according to the argumentative scheme proposed by Toulmin [42], while the conclusion (the item is recommended) constitute itself a *claim*. While this could be considered a 'shallow' structure, compared to the complete Toulmin argument scheme (which involves additional components, like rebuttal or refutation), it resembles explanation schemes based on deductive arguments, such as those widely used in the scientific field (i. e. a set of explanatory propositions is logically followed by an explanatory target, as discussed by Thagard and Litt [37]), or even more particularly, explanation schemes in RS such as the one used by Zanker and Schoberegger [52], who provides brief sentences – two facts and a claim – as explanations for content-based recommendations of hiking routes, energy and mobile phone plans.

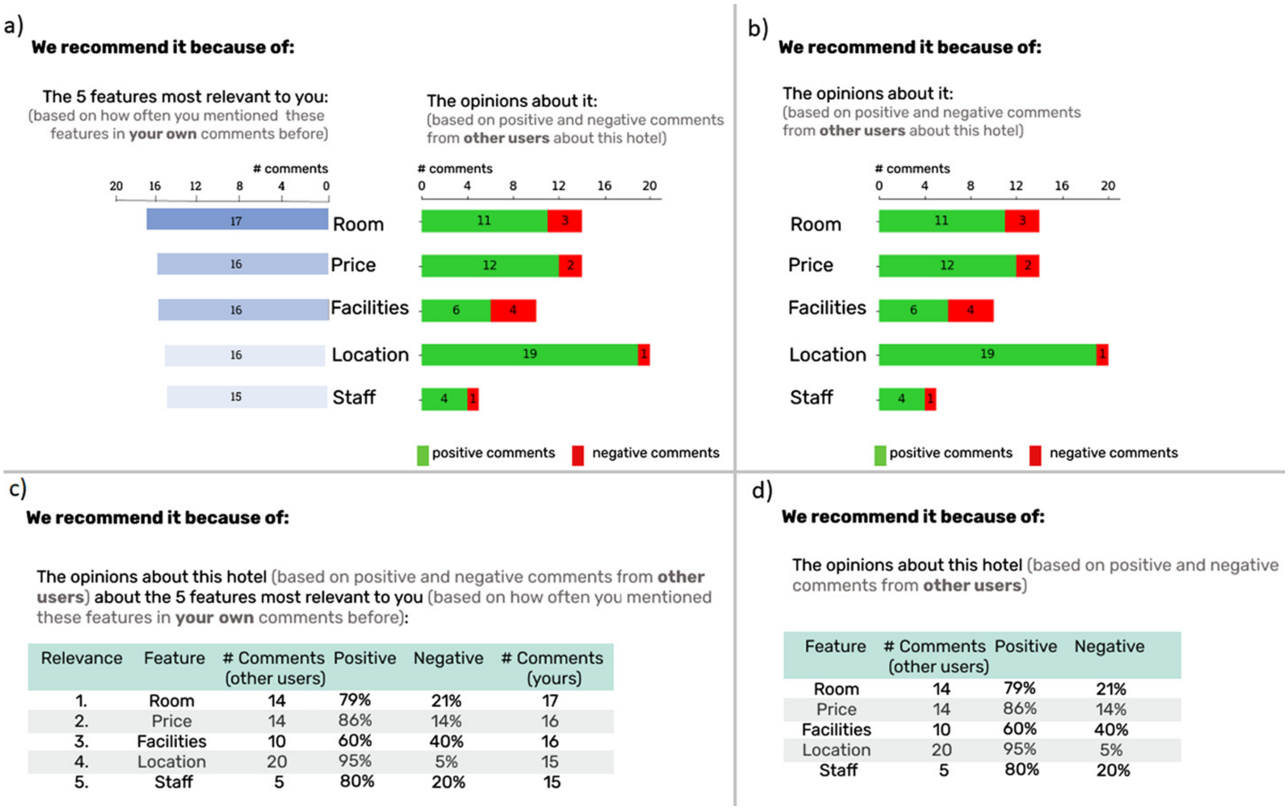
In consequence, our explanation design (see Figure 1) seeks to represent an argumentative structure, while reflecting in turn the arguments provided by other users in their reviews, in a consolidated manner. Therefore, our proposed scheme consists of a claim ("We recommend this hotel") and the propositions that support such claim, connected with the conjunction "because". We propose to provide the following pieces of information in proposition statements:

1. Item quality: A summary of comments reported by previous hotel guests for different aspects, as well as what percentage were positive and negative.

2. User preferences: what are the most important item aspects to the target user. In this regard, we aimed to make the user's own profile transparent, by showing the user's inferred importance of each aspect, together with the opinions of other users about the aspect (as shown in the examples included in Figures 1a and 1c), in order to facilitate a direct comparison of the points of view of others and their alignment with their own preferences.

3. Statements that inform how the user preferences and item quality are extracted (e.g,"based on how often you mentioned these features in your own comments before"). We believe that providing this information, in addition to the information listed above, could increase the perception of trust by users, while decreasing the perception that they are interacting with a black box.

While arguments are usually associated with oral or written speech, arguments can also be communicated using visual representations (e. g. graphics or images). In this


a)
We recommend it because of:
The 5 features most relevant to you:
(based on how often you mentioned these features in your own comments before)
The opinions about it:
(based on positive and negative comments from other users about this hotel)
# comments
20
16
12
8
4
0
# comments
0
4
8
12
16
20
17
Room
11
3
16
Price
12
2
16
Facilities
6
4
16
Location
19
1
15
Staff
4
1
positive comments
negative comments
b)
We recommend it because of:
The opinions about it:
(based on positive and negative comments from other users about this hotel)
# comments
0
4
8
12
16
20
Room
11
3
Price
12
2
Facilities
6
4
Location
19
1
Staff
4
1
positive comments
negative comments
c)
d)

c) We recommend it because of:

The opinions about this hotel (based on positive and negative comments from other users) about the 5 features most relevant to you (based on how often you mentioned these features in your own comments before):

Relevance	Feature	# Comments (other users)	Positive	Negative	# Comments (yours)
1.	Room	14	79%	21%	17
2.	Price	14	86%	14%	16
3.	Facilities	10	60%	40%	16
4.	Location	20	95%	5%	15
5.	Staff	5	80%	20%	15

d) We recommend it because of:

The opinions about this hotel (based on positive and negative comments from other users)

Feature	# Comments (other users)	Positive	Negative
Room	14	79%	21%
Price	14	86%	14%
Facilities	10	60%	40%
Location	20	95%	5%
Staff	5	80%	20%

Figure 1: Explanations displayed in empirical study for every experimental condition, for one of the recommended hotels. a) Style 'visual', user preferences 'yes'. b) Style 'visual', user preferences 'no'. c) Style 'text', user preferences 'yes'. d) Style 'text', user preferences 'no'.

regard, according to Blair [4], visual arguments (a combination of visual and verbal communication) may, in addition to representing propositional content, have a greater "rhetorical power potential" than verbal arguments, due (among others) to their greater immediacy.

In consequence, we aimed to test the effect of the two factors: display *style* and display of the *user preferences*. An example of each condition is provided in Figure 1.

'Bar chart' style: Provides a view of the number of comments per aspect and percentages of positive and negative opinions using bar charts.

'Table' style: Provides the same information used in the visual condition, but instead of using bar charts, presents the information within a table.

Additionally, every display style involves two variations:

User preferences 'yes'. The information about the user preferences is provided.

User preferences 'no'. No information about the user preferences is displayed.

4 Experiment 1: System and Explanation Quality Perception, between Subjects

We implemented a prototype of a hotel recommender system that provides both recommendations and explanations, based on the design discussed in Section 3, and conducted an experiment where we compared users' perception of the overall system in terms of transparency, effectiveness, efficiency and trust. In this regard, we aimed to test our hypothesis that users would report a more positive perception of the RS when information about their user preferences is provided (*H1*). Additionally, we hypothesized that users with greater visual abilities would find explanations better when these are provided using visual aids, such as a bar chart, in comparison to tabulated information (*H2*). In particular, the aim of experiment 1 was to compare the overall perception of the prototype and its explanation quality, in a between groups manner (participants were assigned to conditions that reflect the different types of explanations designed), while in a subsequent experiment (see Section 5) we addressed the perception of specific aspects of explanations within subjects, as well as the usefulness of individual explanation components.

4.1 Methods

Participants

We recruited 150 participants (66 female, mean age 39.08 and range between 23 and 73) through Amazon Mechanical Turk. We restricted the execution of the task to workers located in the U.S, with a HIT (Human Intelligence Task) approval rate greater than 95 %, and number of HIT's approved greater than 500. We applied a quality check in order to select participants with quality survey responses, i. e. at least 5 of the 6 high priority validation questions were answered correctly, more than 30s were spent on the recommendation step and more than 50s on the evaluation questionnaire. The responses of 46 subjects were discarded due to this quality check (from an initial number of 195 workers), so only the responses of 150 subjects were used for the analysis (statistical power of 85 %, $\alpha = 0.05$). Participants were rewarded with $1 plus a bonus up to $0.4 depending on the quality of their response to the question "Why did you choose this hotel?" set at the end of the survey. Time devoted to the task by participants (in minutes): M = 8.04, SD = 1.62.

Study Design

The study follows a 2x2 between-subjects design, and each participant was assigned randomly to one of four conditions that represent the combination of the two factors: display *style* and *user preferences* provided or not. We presented participants with a fixed list of 5 hotels that represented the recommendations for a hypothetical hotel search, and a detailed view including an explanation of why every item was recommended. Then, participants were asked to choose the hotel they considered the best, to report their reasons to it, and to rate their perception of both recommender and its explanations. The explanations and recommendations were generated using the EFM algorithm [53] and the dataset of hotels' reviews, ArguAna [45], although they were presented to the participants only through a prototype, i. e. no real system was implemented to allow the interactions.

Given that we had no access to previously written participants' reviews (which is not only important for the optimal functioning of the algorithm, but also constitutes a base to test the condition "user preferences"), we calculated the top 5 of the most important aspects to all users within the dataset, namely: room, price, facilities, location and staff. Then, a random user was chosen from the dataset with those same preferences, and 5 of her top-ranked options according to the EFM algorithm were selected to be presented to participants, alongside their explanations. Additionally, we presented the users with a cover story, in which we told the users to pretend that their most important aspect was the "room" and the "price".

Questionnaires

Evaluation: We utilized items from [33] to evaluate the perception of system transparency (construct *transparency*, user understands why items were recommended), from [25] to evaluate the perception of system effectiveness (construct *perceived system effectiveness*, system is useful and helps the user to make better choices) and efficiency (user can save time with the recommender), and items from [27] to assess the perception of trust in the system (constructs *trusting beliefs*, user considers the system to be honest and trusts its recommendations, and *trusting intentions*, user willing to share information). In addition, we also adapted 3 items from [25] to address explanation quality (construct *perceived recommendation quality*, user likes explanations, considers them relevant). All items were measured with a 1–5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

User characteristics: We used all the items of the Rational and Intuitive Decision Styles Scale [20] as well as the scale of the social awareness competency [17]. Additionally, We used the visualization familiarity items as proposed by [26]. All items were measured with a 1–5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

Procedure

First, participants were asked to answer demographic questions and the questionnaire on user characteristics. We indicated in the instructions step that a 5 hotels list reflecting the results of a hypothetical hotels' search would be presented. We asked them to click the "View Details" button for each hotel, and to read carefully the explanations provided in each case (examples of explanations for the different experimental conditions are provided in Figure 1). Additionally, we provided a cover story, as an attempt to establish a common starting point in terms of reasons to travel (a business trip), and the supposedly most interesting aspects for the user (room and facilities).

The list of hotels, their names, photos, prices and locations, as well as their ratings and the numbers of reviews and positive and negative opinions, remained constant to all users. Variations focused only on display style and the presentation of user preferences, depending on the condition to which each participant was assigned. After the interaction with the prototype, subjects were asked to choose the hotel that best suited their purpose, as well as an open question about their reasons for choosing that hotel. Then, subjects answered the evaluation questionnaire. In addition, we included an open-ended question, so that participants could indicate in their own words their general opinion about the explanations provided. We included 11 validation questions to check attentiveness and the effective completion of the task.

Data Analysis

We evaluated the effect that display style and the display of user preferences (independent variables IVs) may have on the perception of the prototype and its explanations, and to what extent user characteristics (regarded as moderators or covariates) could influence such perception (rational and intuitive decision-making style, social awareness and visualization familiarity). Here, the dependent variables (DVs) are evaluation scores on: system transparency, effectiveness, efficiency, trust and explanation quality. Here, evaluation scores were calculated as the average of the individual values reported for the questionnaire items related to each DV. Regarding the covariates, we calculated the scores of the rational and the intuitive decision making styles, social awareness and visualization familiarity for each individual as the average of the reported values for the items of every scale.

Given that our DVs are continuous (scores are the averages of reported answers of questionnaire items of each construct) and correlated (correlation coefficients in Table 1), and that we address also the effect of covariates, a MANCOVA analysis was performed, to assess the simultaneous effect of presentation styles and interactivity on the overall system perception, as well as the influence of user characteristics on it. Subsequent ANCOVA analyses were performed to test main effects of IVs and covariates, as well as the effect of interactions between them. Q-Q plots of residuals were checked to validate the adequacy of the analysis.

4.2 Results

Evaluation and User Characteristics Scores

We found that explanations including information of user preferences are perceived slightly better than explanations without this information in terms of explanation quality and system transparency, effectiveness, efficiency and trust. On the other hand, explanations including a bar chart were perceived slightly better than explanations with a table, in regard to explanation quality and trust, while the opposite was observed in relation to transparency, effectiveness and efficiency. However, as discussed in detail below, such differences are not statistically significant. The average evaluation scores by presentation style and display of user preferences are shown in Table 1.

In regard to user characteristics, distributions of the scores of rational (M = 4.24, SD = 0.56) and the intuitive (M = 2.65, SD = 1.01) decision making styles, social awareness (M = 3.92, SD = 0.59) and visual familiarity (M = 3.03, SD = 1.02) are depicted in Figure 2a. Here, we observed a skewed right distribution of rational decision making-style and social awareness scores, not being that the case for the intuitive decision-making style and visual familiarity, i. e., most users consider themselves to be predominantly rational decision makers who are able to listen to others and take into account the opinions of others; however, a more balanced distribution is observed in the remaining user characteristics: only a minority recognize themselves as very (or not at all) familiar with visual representations of information, and as very (or not at all) intuitive decision makers. In addition, results suggest an influence of some of the user characteristics on the perception of the system by users, which we describe in detail below.

Table 1: Experiment 1, mean values and standard deviations of perception on explanation aims, per *display style* and *display of user preferences* (n = 150); values reported with a 5-Likert scale; high mean values correspond to a positive perception of recommender and its explanations. Pearson correlation matrix, p<0.001 for all correlation coefficients.

Style	Table		Bar chart		*User Pref.*	Yes		No		*Corr.*	Variable			
Variable	M	SD	M	SD		M	SD	M	SD		1	2	3	4
1. Expl. Quality	3.83	0.65	3.86	0.67		3.88	0.67	3.81	0.65					
2. Transparency	3.96	0.73	3.87	0.85		3.99	0.74	3.84	0.84		0.37	—		
3. Effectiveness	3.88	0.61	3.75	0.75		3.84	0.76	3.79	0.61		0.60	0.47	—	
4. Efficiency	3.96	0.73	3.89	0.92		4.00	0.78	3.86	0.87		0.36	0.39	0.52	—
5. Trust	3.75	0.60	3.89	0.63		3.84	0.65	3.81	0.59		0.66	0.40	0.67	0.58

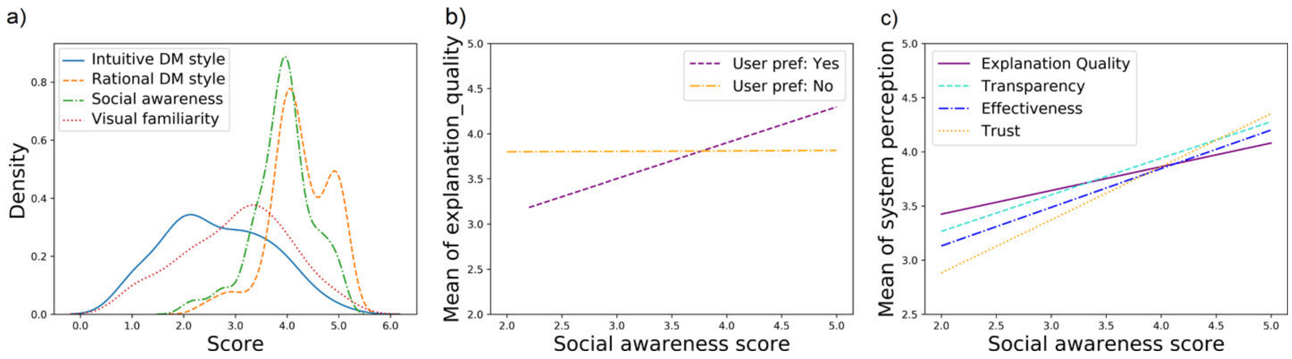


Figure 2: Plots Experiment 1. a) Kernel density estimate of user characteristics scores: rational and intuitive decision making styles, social awareness and visual familiarity. b) Interaction plot for explanation quality (fitted means of individual scores) between display of user preferences and social awareness. c) Effect of social awareness on all explanation aims (fitted means of individual scores). All scores within the range [1,5].

System and Explanation Quality Perception

Presentation style and display of user preferences: We found no main significant effect of the combination of these factors.

Display of user preferences: No significant multivariate effect was found for display of user preferences.

Presentation style: No significant multivariate effect was found for presentation style.

Rational decision-making style: We found a significant main effect of rational style $F(5, 138) = 4.50$, $p < .001$. Univariate tests revealed a significant effect of this variable on: effectiveness $F(1, 142) = 9.12$, $p = .003$), efficiency ($F(1, 142) = 10.98$, $p = .001$) and trust ($F(1, 142) = 18.82$, $p < .001$). Here, a positive trend was observed between rational decision-making score and the above mentioned DVs, i. e. the higher the rational decision making score, the higher the perception scores of these DVs.

Intuitive decision-making style: We found a significant main effect of intuitive style $F(5, 138) = 3.25$, $p = .008$. Univariate tests revealed a significant effect of this variable on: explanation quality ($F(1, 142) = 16.37$, $p < .001$). Here, a positive trend was observed between this variable and the score of intuitive decision-making style.

Social awareness: We found a significant main effect of social awareness $F(5, 138) = 6.72$, $p <.001$. Univariate tests revealed a significant effect of this variable on: explanation quality ($F(1, 142) = 5.62$, $p = .019$), transparency ($F(1, 142) = 7.93$, $p = .006$), effectiveness ($F(1, 142) = 8.79$, $p = .004$) and trust ($F(1, 142) = 26.56$, $p < .001$). Here, we observed a positive trend in the relationship between social awareness and these DVs (see Figure 2d).

Additionally, a significant interaction effect between social awareness and the display of user preferences on explanation quality was found $F(1, 146) = 4.79$, $p = .030$, with the "yes" condition having a steeper slope than the "no" condition (showing a positive relationship between social awareness and displaying user preferences), the latter remaining constant regardless of the social awareness score (Figure 2b).

5 Experiment 2: Perception on Specific Aspects of Explanations, within Subjects

We used screenshots of the prototype implemented for experiment 1 (see Section 4), reflecting the design discussed in Section 3, and conducted a second experiment aiming to compare users' perception of specific aspects of explanations, when presented to all the four possible explanations (see Figure 1). In experiment 2, differences were addressed within subjects, while in experiment 1 we evaluated the perception of the overall system in a between subjects manner. Likewise to experiment 1, we also aimed to test our hypothesis that users would report a more positive perception when information about user preferences is provided (*H1*), and also that users with greater visual abilities would find explanations better when these are provided using visual aids, such as a bar chart, in comparison to tabulated information (*H2*).

Additionally, the experiment 2 also involved the assessment of the usefulness of individual components of explanations, by participants. In this regard, for example, we hypothesised that most users would find useful the information regarding the origin of the explanatory information provided (*H3*).

5.1 Methods

Participants

We recruited 35 participants (14 female, mean age 42.77 and range between 24 and 65) through Amazon Mechanical Turk. We restricted the execution of the task to workers located in the U.S, with a HIT (Human Intelligence Task) approval rate greater than 95 %, and a number of HIT's approved greater than 500. We applied a quality check in order to select participants with quality survey responses, i. e. at least 5 of the 7 validation questions were answered correctly. The responses of 7 subjects were discarded due to this quality check (from an initial number of 42 workers), so only the responses of 35 subjects were used for the analysis, a value consistent to our within subjects design (statistical power of 95 %, $\alpha = 0.05$). Participants were rewarded with $1. Time devoted to the task by participants (in minutes): M = 6.70, SD = 1.07.

Study Design

The study follows a within-subjects design, and each participant was presented sequentially with an example of each of the 4 types explanations, that represent the combination of the two factors: display *style* and *user preferences* provided or not. The order of presentation of the 4 types of explanation was counterbalanced.

Questionnaires

We used the user experience items (UXP) proposed by [26] to address the explanations reception, comprising: explanation confidence (explanation makes user confident that she/he would like the recommended item), explanation transparency (explanation makes the recommendation process clear), explanation satisfaction (user would enjoy a system if recommendations are presented this way), and explanation persuasiveness (explanations are convincing). Finally, we included additional elements to assess explanation effectiveness (user can make better decisions if explanation presented this way), explanation efficiency (user can save time if system provides this type of explanation), and explanation easiness (explanation is easy to understand). All items were measured with a 1–5 Likert-scale (1: Strongly disagree, 5: Strongly agree). Users were asked to respond to the same user characteristics questionnaire we used in experiment 1.

Additionally, participants were requested to provide their opinions on how helpful they considered the different components of the explanations: the bar plots, the tables, the information about others' opinions, the information about their supposed own comments, and the information on where the bar plots and tables come from. All items were measured with a 1–5 Likert-scale (information is helpful, 1: Strongly disagree, 5: Strongly agree).

Procedure

First, participants were asked to answer demographic questions and the questionnaire on user characteristics. We indicated in the instructions that they will be presented with information about the pros and cons of different hotel features that might be relevant to you, using 4 different display options, and that they would then indicate their opinion about each option. Additionally, we provided a cover story, as an attempt to establish a common starting point in terms of reasons to travel (a business trip), and the supposedly most interesting aspects for the user (room and facilities). After the assessment of all types of explanations, participants were asked to reply questions about the usefulness of specific components of explanations. At the end, they were asked to report their comments and suggestions about the explanations with an open-ended question.

Data Analysis

We evaluated the effect that display style and the display of user preferences (independent variables IVs) may have on the perception of specific aspects regarding the proposed explanations, and to what extent user characteristics (regarded as moderators or covariates) could influence such perception (rational and intuitive decision-making style, social awareness and visualization familiarity). Here, the dependent variables (DVs) are evaluation scores on the following aspects: explanation confidence, explanation transparency, explanation satisfaction, explanation persuasiveness, explanation effectiveness, explanation efficiency and explanation easiness to understand. Regarding the covariates, we calculated the scores of user characteristicas the same way as in study 1 (average of the reported values for the items of every user characteristics scale).

Given that our DVs are ordinal (scores are the reported answers to single questionnaire items) we performed a Friedman test, the non-parametric alternative to the repeated measures ANOVA. Given that our variables are correlated, the significant tests were conducted using Bonferroni adjusted alpha levels of .007 (.05/7).

Additionally, we calculated the average evaluation scores for each possible value of the two factors: presentation style (bar chart and table), and display of user preferences (yes and no). Using these continuous and correlated evaluation scores, we then perform a repeated measures MANCOVA, to assess the simultaneous effect of presentation styles and display of user preferences on explanations perception, as well as the influence of user characteristics on it. Subsequent ANCOVA analyses were performed to test main effects of IVs and covariates, as well as the effect of interactions between them.

Usefulness of explanations components: We performed a series of ordinal logistic regressions to test influence on scores of usefulness of components – DVs (bar chart, table, others' opinion view, own preferences view, information source) by predictor variables, in this case the user characteristics (rational and intuitive decision-make style, social awareness and visualization familiarity), which were tested a priori to verify there was no violation of the assumption of no multicollinearity. Q-Q plots of residuals were also checked to validate the adequacy of the analysis.

DVs were initially rated using a 5-likert scale, but additionally we grouped answers as Yes (agree and strongly agree that element is helpful), and No / Neutral (disagree, strongly disagree and neutral that element is helpful) for subsequent analysis. We then calculated the percentages of Yes and No/Neutral responses regarding the different explanation components, and performed a binomial test, to check whether the proportions of Yes and No/Neutral answers were different from a proportion that assumes that the percentages are equal (50 % of Yes and 50 % of No/Neutral).

Finally, we used a Wilcoxon rank t-test to compare the average responses of the perception of usefulness of a view of others' opinion with that of a view of their own preferences, as well as the average responses of perceived usefulness of tables compared to bar charts in explanations.

5.2 Results

Evaluation and User Characteristics Scores

We observed only small differences between table and bar chart explanations, and between explanations including or not user preferences, in regard to most of the specific aspects of explanations evaluated, with the exception of easiness to understand. As discussed in detail below, explanations without display of user preferences were perceived easier to understand, this difference being statistically significant. The average evaluation scores by presentation style and display of user preferences are shown in Table 2.

In regard to user characteristics scores, we observed similar distributions of such scores: a skewed right distribution of rational decision making-style and social awareness scores, not being that the case for the intuitive decision-making style and visual familiarity. Distributions of the scores of rational (M = 4.34, SD = 0.7) and intuitive (M = 2.13, SD = 0.83) decision making styles, social awareness (M = 3.55, SD = 0.53) and visualization familiarity (M = 2.82, SD = 1.17) are depicted in Figure 3a. Additionally, we observed a main effect of some of these user characteristics on the perception of specific aspects of explanations, as well as interaction effects involving these variables. Such findings are described below.

Perception of Explanations

Presentation style and display of user preferences: We found no main significant effect of the combination of these factors after Bonferroni correction.

Display of user preferences: We found a multivariate effect of display of user preferences, $F(7,28) = 2.41$, $p = .046$. Univariate tests revealed a main effect of display of user preferences on explanation easiness to understand $F(1,34) = 6.42$, $p = .016$, so that explanations that do not include information on user preferences are significantly easier to understand (M=3.76, SD=0.91), compared to those showing such information (M=4.01, SD=0.72).

Presentation style: No multivariate main effect of presentation style was found.

Table 2: Experiment 2, mean values and standard deviations of perception on explanation aspects, per *display style* and *display of user preferences* (n=35); values reported with a 5-Likert scale; high mean values correspond to a positive perception of explanations aspects.

Style	Table		Bar chart		*User Pref.*	Yes		No	
Variable	M	SD	M	SD		M	SD	M	SD
1. Expl. Confidence	3.30	0.92	3.33	0.98		3.33	0.86	3.30	0.96
2. Expl. Transparency	3.50	0.88	3.64	0.97		3.51	0.88	3.63	0.74
3. Expl. Satisfaction	3.41	1.08	3.27	1.05		3.24	1.03	3.44	0.93
4. Expl. Persuasiveness	3.30	0.92	3.29	1.02		3.34	0.94	3.24	1.06
5. Expl. Effectiveness	3.33	0.97	3.37	0.95		3.29	0.93	3.41	0.85
6. Expl. Efficiency	3.31	1.13	3.34	1.09		3.21	1.05	3.44	0.93
7. Expl. Easiness to understand	3.76	0.92	3.84	0.86		3.59	0.93	4.01	0.72

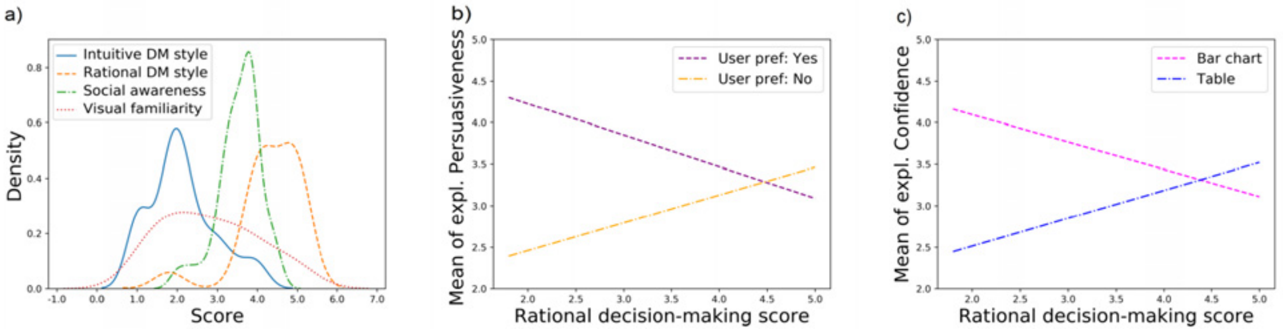


Figure 3: Plots Experiment 2. a) Kernel density estimate of user characteristics scores: rational and intuitive decision making styles, social awareness and visual familiarity. b) Interaction plot for explanation persuasiveness (fitted means of individual scores) between display of user preferences and rational decision-making style. c) Interaction plot for explanation confidence (fitted means of individual scores) between presentation style and rational decision-making style. All scores within the range [1,5].

Display of user preferences and rational decision-making style: We found a multivariate interaction effect between these two variables, $F(7{,}27)$ =, $p = .002$. Univariate tests revealed the significant interaction effect ot these variables on: explanation transparency ($F(1{,}33) = 7.79$, $p = .009$), explanation satisfaction ($F(1{,}33) = 5.62$, $p = .024$), explanation persuasiveness ($F(1{,}33) = 20.67$, $p < .001$), explanation easiness to understand ($F(1{,}33) = 7.36$, $p = .011$) and explanation effectiveness ($F(1{,}33) = 5.34$, $p = .027$). For all these DVs, the same trend was observed: the higher the reported rational decision-making style, the higher the scores on the different DVs when the user profile was not shown, while the opposite trend was observed when it was shown. An example of this trend is observed in Figure 3b.

Presentation style and rational decision-making style: We found a multivariate interaction effect between these two variables, $F(7{,}27)$ =, $p = .006$. Univariate tests revealed the significant interaction effect ot these variables on: explanation confidence ($F(1{,}33) = 14.09$, $p = .001$), explanation satisfaction ($F(1{,}33) = 5.78$, $p = .022$), explanation easiness to understand ($F(1{,}33) = 7.36$, $p = .011$), explanation effectiveness ($F(1{,}33) = 4.34$, $p = .045$) and explanation efficiency ($F(1{,}33) = 7.15$, $p = .012$). For all these DVs, the same trend was observed: the higher the reported rational decision-making style, the higher the scores on the different DVs when the table was provided, while the opposite trend was observed when the bar chart was shown. An example of this trend is observed in Figure 3c.

Usefulness of Explanation Components

Effect of user characteristics on usefulness.

An increase in intuitive decision-making score was significantly associated with an increase in the odds of participants reporting higher values of: usefulness of bar charts in explanations, with an odds ratio of 3.16 (95 % CI, 1.12 to 9.81), Wald $\chi 2(1) = 4.76$, $p = .029$, and usefulness of a view of others' opinions in explanations, with an odds ratio of 3.21 (95 % CI, 1.11 to 11.19), Wald $\chi 2(1) = 4.69$, $p = .030$.

An increase in social awareness score was significantly associated with an increase in the odds of participants reporting higher values of: usefulness of information origin in explanations, with an odds ratio of 5.77 (95 % CI, 1.38 to 27.20), Wald $\chi 2(1) = 5.82$, $p = .016$.

An increase in visualization familiarity score was significantly associated with an increase in the odds of participants reporting higher values of: usefulness of bar charts in explanations, with an odds ratio of 3.79 (95 % CI, 1.76 to 9.47), Wald $\chi 2(1) = 12.33$, $p < .001$, usefulness of a view of own comments in explanations, with an odds ratio of 2.54 (95 % CI, 1.35 to 5.14), Wald $\chi 2(1) = 8.62$, $p = .003$, and usefulness of information origin in explanations, with an odds ratio of 3.21 (95 % CI, 1.62 to 6.96), Wald $\chi 2(1) = 11.77$, $p < .001$.

Participants who found components helpful.

We found then that a significant majority found the information about others' opinion helpful ($\chi 2(1, N = 35) = 6.40$, $p = .011$), so that both tables ($\chi 2(1, N = 35) = 6.40$, $p = .011$) and bar charts ($\chi 2(1, N = 35) = 8.75$, $p = .003$), whereas only a significant minority found the display of user preferences helpful ($\chi 2(1, N = 35) = 6.40$, $p = .011$). On the other hand, providing details about where the information used for the recommendation comes seems to be regarded as helpful by most people, but the difference with the proportion of people that found it non helpful / neutral is not significant. Proportions are depicted in Figure 4.

Comparison of usefulness scores.

We found that the average usefulness of the components view of others' opinions and view of own prefer-

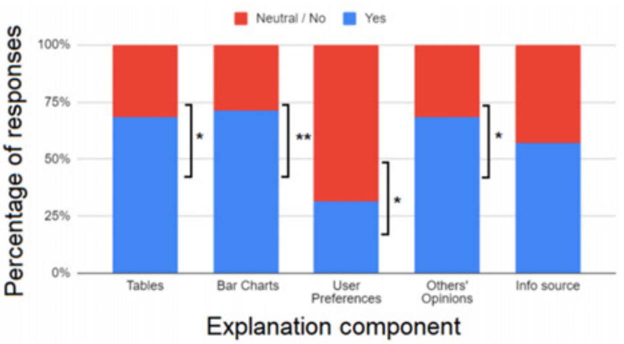


Figure 4: Proportion of participants who found the different explanation components helpful (Y) or non helpful neutral (Neutral/No). * p <0.5, **p<0.01.

ences are significantly different ($W = 0.89$, $p < .001$), with the display of others' opinions having a higher mean (M = 3.74, SD = 1.07) than the display of users' preferences (M = 2.63, SD = 1.29). On the other hand, we found no significant difference when comparing the mean responses of usefulness of tables (M = 3.60, SD = 1.19), and bar charts (M = 3.71, SD = 1.20), although bar charts are perceived slightly more helpful than tables.

6 Discussion

6.1 Effect of Profile Transparency

In regard to our **H1**, we found no main effect of the display of user preferences on the perception of the system or its explanations. Although contrary to our expectations, the lack of a significant influence of disclosing user preferences seems to be somehow in line with the results reported by Tintarev and Masthoff [39], Gedikli et al. [18], who observed that providing personalized explanations (in which preferences were presented along with item properties), while potentially beneficial in terms of satisfaction with the explanations, did not necessarily result in a better perception of effectiveness (helping the user to make better decisions). The authors suggested that a possible reason could be a mismatch between the expectation generated by the explanation and the actual evaluation after trying the item. In our case, however, this could be related to how easy it was for the participants to understand the explanations. In particular, we observed that explanations without information on user preferences were significantly easier to understand compared to those that included such information. In addition, we observed that users with less visualization familiarity reported lower usefulness scores of the user preference section, suggesting that the proposed presentation of this section still needs to be improved to benefit users who do not have sufficient experience with information visualization techniques as well.

Although the display does not have a main effect on the perception of the system and its explanations, we observed a mediating effect of social awareness, such that individual differences in this characteristic were reflected in differences in the perception of the explanation quality. Here, our findings suggest that people who tend to listen more to others tend to perceive better the explanations that include information about their own profile. On the other hand, when user preferences are not displayed, the perception of explanation quality remains pretty much the same, despite the extent of users' social awareness. At this respect, we believe that users with greater abilities to take into account the opinion of others might appreciate the chance to see the alignment of their own preferences with the opinions of others, in an effortless manner, given that a metric of aspect relevance was placed right next to the metric of other users' opinions regarding such aspect.

Additionally, although no significant interaction effect was found between the rational decision-making style and the display of user preferences on the perception of the system in general, we found that this interaction had a significant effect on the perception of most of the specific aspects of the explanations. In this case, users who reported higher scores for rational decision-making style reported less preference for explanations that provided information on the user's profile. In this regard, we believe that more skeptical users might think that the system hides additional information about the user's profile that could be used to generate recommendations, so showing only the frequencies of the mentions of the user's aspects may not be enough to satisfy their curiosity and need for further information.

Overall, while most users reported they found the information about others' opinions in explanations useful, the opposite was the case for the information about own preferences, with only a minority of users reporting they found this section useful, and reporting comments in this sense, e. g. "It makes sense that a program would analyse my past comments to find out about my preference...", or "It could be more useful if there was an explanation of how my preferences are used in the calculation" (the latter by a user assigned to a non user preferences condition in Experiment 1). While the difficulty in understanding this information seems to play an important role in this regard, as discussed above, we believe that in the face of a lack of motivation or "feeling of personal relevance" to perform the task, and the need for greater cognitive effort to do so, the user may simply choose not to attend this section, as discussed by [7, 41].

Overall, the results suggest that users seemed to be much more interested in other people's opinion and their weight in the recommendation, rather than how these recommendations fit their own preferences. The reasons for this could be twofold: 1) domain under study is an experience good, where the search for information is characterized by a greater reliance on word-of-mouth [29, 24], and where users might be interested, for example, in finding aspects that had a prominent negative opinion, even when the aspect is not necessarily the most important for them. 2) user models enabled by methods like EFM might not accurately reflect users' real preferences.

As for the explanatory model chosen as inspiration for our study, we believe that the user profile obtained using methods such as the Explicit Factors Model (EFM) [53], may not fully reflect the true preferences of the user, as addressing an aspect in a review, other than reflecting one's own preference, may be motivated by other factors. On the one hand, customers report on the aspects they consider satisfactory or unsatisfactory, but the nature of these aspects may define the satisfaction report on them, as discussed by Chowdhary and Prakas [11]: the presence of some aspects that are taken for granted (cleanliness, for example) may not lead to customers' satisfaction, while their absence leads to dissatisfaction and subsequent reporting. Similarly, motivational factors (e. g., proximity to the beach) can lead customers to satisfaction, but their absence does not necessarily cause a negative report.

On the other hand, when inspecting the data we aimed to provide to our participants in the experimental set-up, we observed that in many cases, users in dataset had fairly homogeneous frequencies of reporting aspects in their reviews i. e., many of them tend to talk about general aspects (e. g., "room", "facilities") in similar proportions. This makes it difficult, in some cases, to detect compelling preferences, which can be prominently represented in an explanation. Thus, we believe that if all aspects have a very similar assessment of relevance (and thus the bars or numbers in the chart look almost the same) the preference information in explanations might be perceived as irrelevant, unnecessary, and even confusing to users. This seemed to be the case for one of the study participants, assigned to the condition bar chart – user preferences displayed, who reported: "I did not understand the left side of the graph which was consistent across about the features relevant to me (seems weird and confusing to include that)". In this regard, however, further evidence is needed to confirm that this is actually the case.

6.2 Effect of Presentation Style

In regard to presentation style, we compared users' perception of explanations consisting of tables or bar charts, that provided an aggregated view of positive and negative opinions given by users to every hotel. Here, we did not find a salient preference of one style over the other. Additionally, despite no significant interaction effect between visualization familiarity and display style was found, we observed, in line with our **H2**, that visualization familiarity might play a role in this perception, since users with higher scores in relation to this user characteristic, gave higher usefulness scores to the bar charts as part of the explanations.

Additionally, our results suggest an interaction effect between rational decision-making style and presentation style on the perception of explanations,so that users with a more rational decision-making style reported higher confidence scores for explanations consisting of tables, while the opposite trend was observed for bar chart explanations. This could be explained by the tendency to seek more detailed information when making decisions, which characterizes individuals with a predominantly rational decision-making style, who may be more interested in evaluating explicit and accurate numbers (such as those presented in the table), compared to less rational users, who may benefit more from representations that allow faster comparisons (such as the bar chart). In this respect, according to Spence and Lewandowsky [36], a presentation of data by means of a table may be more beneficial than the use of a graphical representation, when the objective is the evaluation of exact numbers, and provided that the number of data points presented remains low (in our case it is 5, the number of aspects for which information is provided). The above is also consistent with another of our findings, in which users with a predominantly intuitive decision-making style reported significantly higher scores on the usefulness of bar charts, which seems to be a consequence of the rapid processing of information enabled by graphical representations.

Overall, and despite the differences in perception between tables and bar charts in terms of user characteristics, most users found the two types of explanatory components to be useful, and although the perception of usefulness of the bar chart is slightly more positive than that of the table, this difference is not significant, so we can conclude that both types of presentation are useful to users.

6.3 Main Effect of User Characteristics

So far we have discussed how user characteristics mediate the effect of user preferences or display style on the perception of both the system and its explanations. However, it is important to note that we also observed main effects of decision-making style and social awareness on participants' perception. In particular, we observed that users with a predominant rational style seemed to perceive a greater benefit of the explanation in helping them make faster and better decisions, and as a good means to believe that the recommender is honest, while more intuitive users reported a more positive perception on the quality of explanations, i. e. they like it better and found them more relevant. We believe that the reason why more rational users did not necessarily like our explanations much more could be the lack of additional and detailed arguments addressing the causes of the positive and negative evaluation by customers, given their tendency to examine the information in depth when making decisions, while more intuitive users do not need to go into such detail, and can be satisfied with the aggregate view of opinions we provide in our proposed design. In fact, we received several observations in this regard: "Written reviews from others could be helpful. Rather than just the amount of positive or negative opinions, if you could see specificaly (sic) why they rated the hotel that way it would help personalize your experience even more.", "I think specific comments and reviews would've been helpful in making a final decision. I prefer to read other users' comments about their hotel stay to make a more informed decision".

Furthermore, our results also suggest that social awareness seems to play a significant role in the perception of review-based RS, since we found significant main effects of social awareness on almost all variables evaluated, which seems to be a natural consequence of using users' opinions as a basis for generating explanations, which seems to benefit greatly people with a more pronounced tendency to listen to others and take their opinions into account.

6.4 Usefulness of Origin of Information

Finally, with respect to our **H3**, we found that participants did not find indications of the origin of the information significantly more useful, unless user characteristics such as social awareness were taken into account. In this case, users who were more willing to consider other opinions found more useful the explanatory component reporting the explanations' source of information (i. e. the reviews written by users). In this regard, it is possible that users with less social awareness, being less interested in others' opinions, might have been disappointed because of the expectation that other sources of information would be taken into account when generating recommendations. We believe that this mismatch between the user's conceptual model and the transmitted system's conceptual [21] model could have resulted in a lower usefulness score for this section. However, an alternative explanation could be that users found that information redundant, which could be the case for users who felt that the information on the origin of the information represented in the explanation was already sufficiently self-explanatory.

6.5 Limitations

An important limitation of our study is the fact that user's preferred aspects were fixed and participants were instructed to pretend that those aspects were the ones that mattered most to them, aiming to give a practical work around to the cold-start problem in the user study design. However, we acknowledge that this might interfere with the real perceived benefit of providing the user preferences as part of the explanations.

Additionally, we acknowledge that the use of the Amazon Mechanical Turk implies an important challenge in regard to high quality responses. Here, despite our implemented quality control and the bonus offered, further actions might be still evaluated, aiming to encourage users to genuinely make a decision. In this case, a game strategy could be used, in which users are asked to solve a specific task, for example, to choose the hotel that fits certain conditions using the information provided in the explanations, and to receive a bonus only if the task is solved successfully.

7 Conclusion and Future Work

In this paper, we have presented the design of argumentative explanations based on reviews, in display styles that involve visual representations like tabulated data and bar charts, as well as information about the user preferences. We also addressed the role that individual differences regarding decision making styles, social awareness and visual familiarity play in such perception. Although we found no main differences in perception between the regarded display styles, nor the presence or absence of user preferences in explanations, we found that, when taking into account user characteristics, i. e. social awareness, rational or intuitive decision making style, we are able to do detect differences in explanations' perception between users.

Given the variability of perception of explanatory components when taking into account user characteristics, and given the difficulties (even impossibility) posed by a request or automatic inference of them, we suggest explanation designers to consider a more flexible approach, that allow users to interactively request for different presentation styles and explanatory components whenever it is needed. For example, the system could offer an initial view of explanatory information using a chart, and provide an option to visualize the same data as explicit numbers in a table, or within verbal sentences, to ensure that users who require more support to interpret the explanations have the opportunity to do so.

As part of our future work, and in order to mitigate our limitation regarding the use of real user preferences, we plan to provide a mechanism that allows participants to read explanations that fit better to their real preferences, e. g. to request participants preferences and calculate similarity with users within the dataset, so that we obtain the most similar user in terms of preferences, and use them as a proxy to calculate rating predictions.

Furthermore, we plan to work on and test improvements of the explanatory component of user profile, in order to rule out the difficulty of understanding this type of information as the main cause of its lack of usefulness, so that we can further explore the convenience of using reviews as the primary source for modeling user preferences in review-based explanatory methods.

Funding: German Research Foundation (DFG), Grant Number: GRK 2167, Research Training Group "User-Centred Social Media".

References

[1] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 717–725.

[2] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to Recommend?: User Trust Factors in Movie Recommender Systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 287–300.

[3] Mustafa Bilgic and Raymond J Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *Proceedings of the Workshop on the Next Stage of Recommender Systems Research, Beyond Personalization IUI 05*. 13–18.

[4] J. Anthony Blair. 2012. The Possibility and Actuality of Visual Arguments. in: Tindale C. (eds), *Groundwork in the Theory of Argumentation 21*, 205–223.

[5] Jacob A. Burack, Tara Flanagan, Terry Peled, Hazel M. Sutton, Catherine Zygmuntowicz, and Jody T. Manly. 2006. Social Perspective-Taking Skills in Maltreated Children and Adolescents. *Developmental Psychology* 42, 2, 207–217.

[6] Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. 2013. Multi document summarization of evaluative text. In *Computational Intelligence*, Vol. 29. 545–574.

[7] Richard L. Celsi and Jerry C. Olson. 1988. The Role of Involvement in Attention and Comprehension Processes. *Journal of Consumer Research* 15, 2, 210–224.

[8] Michael Chandler. 1973. Egocentrism and Antisocial Behavior: The Assessment and Training of Social Perspective-Taking Skills. *Developmental Psychology* 9, 3, 326–332.

[9] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web. International World Wide Web Conferences Steering Committee*. 1583–1592.

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

[11] Nimit Chowdhary and Monika Prakas. 2005. Service Quality: Revisiting the two factors theory. *Journal of Services Research* 5, 1, 61–75.

[12] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic Generation of Natural Language Explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. 57:1–57:2.

[13] Paul T. Costa and Robert R. McCrae. 1992. Four ways five factors are basic. *Personality and Individual Differences* 13, 6, 653–665.

[14] Ruihai Dong, Michael P. O Mahony, and Barry Smyth. 2014. Further Experiments in Opinionated Product Recommendation. In *Case Based Reasoning Research and Development*. Springer International Publishing, 110–124.

[15] Michael J. Driver, Kenneth E. Brousseau, and Phil L. Hunsaker. 1990. The dynamic decision maker.

[16] Andrew S. C. Ehrenberg. 1975. *Data reduction: Analyzing and interpreting statistical data*. Wiley, New York.

[17] Collaborative for Academic Social and Emotional Learning. 2013. 2013 CASEL guide: Effective social and emotional learning programs – Preschool and elementary school edition.

[18] 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, 367–382.

[19] Justin Scott Giboney, Susan A Brown, Paul Benjamin Lowry, and Jay F Nunamaker Jr. 2015. User Acceptance of Knowledge-Based System Recommendations: Explanations, Arguments, and Fit. *Decis. Support Syst* 72, 1–10.

[20] Katherine Hamilton, Shin-I Shih, and Susan Mohammed. 2016. The Development and Validation of the Rational and Intuitive Decision Styles Scale. *Journal of Personality Assessment* 98, 5, 523–535.

[21] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[22] Diana C. Hernandez-Bocanegra, Tim Donkers, and Jürgen Ziegler. 2020. Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*.

[23] John R Kirby, Phillip J Moore, and Neville J Schofield. 1988. Verbal and visual learning styles. *Contemporary Educational Psychology* 12, 2, 169–184.

[24] Lisa Klein. 1998. Evaluating the Potential of InteractiveMedia through a New Lens: Search versus Experience Goods. In *Journal of Business Research*, Vol. 41. 195–203.

[25] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. In *User Modeling and User-Adapted Interaction*. 441–504.

[26] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of 24th International Conference on Intelligent User Interfaces (IUI 19)*. ACM, 379–390.

[27] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. In *Information Systems Research*, Vol. 13.

[28] Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. In *Intelligent User Interfaces (IUI 16)*, Vol. 2. 256–260.

[29] Philip J. Nelson. 1981. Consumer Information and Advertising. In *Economics of Information*. 42–77.

[30] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model User-Adap* 27, 393–444.

[31] Rosemary Pacini and Seymour Epstein. 1999. The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology* 76, 972–987.

[32] Richard E. Petty and John T. Cacioppo. 1986. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer-Verlag, New York.

[33] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems – RecSys 11*. 157–164.

[34] Wolfgang Schnotz. 2014. Integrated Model of Text and Picture Comprehension. In *The Cambridge Handbook of Multimedia Learning (2nd ed.)*. 72–103.

[35] Janet A. Sniezek and Timothy Buckley. 1995. Cueing and Cognitive Conflict in Judge Advisor Decision Making. *Organizational Behavior and Human Decision Processes* 62, 2, 159–174.

[36] Ian Spence and Stephan Lewandowsky. 1991. Displaying proportions and percentages. 5, 1, 61–77.

[37] Paul Thagard and Abninder Litt. 2000. Models of Scientific Explanation. *The Cambridge handbook of computational cognitive modeling*.

[38] Nava Tintarev. 2007. Explanations of recommendations. *Proceedings of the 2007 ACM conference on Recommender systems, RecSys 07*, 203–206.

[39] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 399–439.

[40] Marko Tkalcic and Li Chen. 2015. Personality and Recommender Systems. In *Recommender Systems Handbook*, 715–739.

[41] Peter Todd and Izak Benbasat. 1999. Evaluating the Impact of DSS, Cognitive Effort, and Incentives on Strategy Selection. *Information Systems Research* 10, 4, 356–374.

[42] Stephen E. Toulmin. 1958. The Uses of Argument.

[43] Edward R. Tufte. 1983. *The visual display of quantitative information*. Graphics Press, Cheshire.

[44] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent User Interfaces*. ACM, 47–56.

[45] Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *15th International Conference on Intelligent Text Processing and Computational Linguistics*. 115–127.

[46] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 18*. 165–174.

[47] Rainer Wasinger, James Wallbank, Luiz Pizzato, Judy Kay, Bob Kummerfeld, Matthias Böhmer, and Antonio Krüger. 2013. Scrutable User Models and Personalised Item Recommendation in Mobile Lifestyle Applications. In *User Modeling, Adaptation, and Personalization, UMAP*. 77–88.

[48] Duane T. Wegener, Richard E. Petty, Kevin L. Blankenship, and Brian Detweiler-Bedell. 2010. Elaboration and numerical anchoring: Implications of attitude theories for consumer judgment and decision making. *Consumer Psychology* 20, 5–16.

[49] Yao Wu and Martin Ester. 2015. Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Eighth ACM International Conference on Web Search and Data Mining*. ACM, 153–162.

[50] Bo Xiao and Izak Benbasat. 2007. ECommerce product recommendation agents: use, characteristics, and impact. *MIS Quarterly* 31, 1, 137–209.

[51] Ilan Yaniv and Maxim Milyavsky. 2007. Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes* 103, 104–120.

[52] Markus Zanker and Martin Schoberegger. 2014. An empirical study on the persuasiveness of fact-based explanations for recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*. 33–36.

[53] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval*. 83–92.

Bionotes

Diana C. Hernandez-Bocanegra
University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science Duisburg, Germany
diana.hernandez-bocanegra@uni-due.de

Diana C. Hernandez-Bocanegra is a research associate and PhD student in the Department of Computer Science and Applied Cognitive Science at the University of Duisburg-Essen, and a member of the Interactive Systems Research Group. Her research interests are focused on recommender systems, explainable AI and the exploration of deep learning and NLProc techniques to improve human-computer interaction.

Jürgen Ziegler
University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science Duisburg, Germany
juergen.ziegler@@uni-due.de

Jürgen Ziegler is a full professor in the Department of Computer Science and Applied Cognitive Science at the University of Duisburg-Essen where he directs the Interactive Systems Research Group. His main research interests lie in the areas of human-computer interaction, human-AI cooperation, recommender systems, information visualization, and health applications. Among other scientific functions he is currently editor-in-chief of i-com - Journal of Interactive Media and chair of the German special interest group on user-centred artificial intelligence.