Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations

Krisztian Balog
Google
London, UK
krisztianb@google.com

Filip Radlinski
Google
London, UK
filiprad@google.com

ABSTRACT

Explanations have a large effect on how people respond to recommendations. However, there are many possible intentions a system may have in generating explanations for a given recommendation—from increasing transparency, to enabling a faster decision, to persuading the recipient. As a good explanation for one goal may not be good for others, we address the questions of (1) how to robustly measure if an explanation meets a given goal and (2) how the different goals interact with each other. Specifically, this paper presents a first proposal of how to measure the quality of explanations along seven common goal dimensions catalogued in the literature. We find that the seven goals are not independent, but rather exhibit strong structure. Proposing two novel explanation evaluation designs, we identify challenges in evaluation, and provide more efficient measurement approaches of explanation quality.

CCS CONCEPTS

• **Information systems → Recommender systems**; Presentation of retrieval results; • **Human-centered computing →** *Natural language interfaces*; **HCI design and evaluation methods**.

KEYWORDS

Recommendations; explanations; evaluation

ACM Reference Format:
Krisztian Balog and Filip Radlinski. 2020. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401032

1 INTRODUCTION

Recommendations are part of everyday life. Be they made by a person, or by an automated system, the recommendations are often accompanied with an explanation, or reason, underlying the suggestions provided. Explanations are known to strongly impact how the recipient of a recommendation responds [13, 14, 23, 28], yet the effect is still not well understood.

At the same time, automated recommender systems have recently proliferated. This has increased attention on explainable and transparent AI, both from technical and ethical perspectives [1, 18]. While explainable system design is not new (dating back to rule-based expert systems of the 1980s [5]), the role of explanations has gained more attention in the past decade [29].

We study the role of the recommender's intention when generating explanations, which we refer to as explanation *goals*. Our focus is on assessing how the choice of goal affects explanations, and how the extent to which a given explanation satisfies different goals can be measured robustly.

Our work starts with seven main goals of explanations, proposed by Tintarev and Masthoff [26]: *transparency*, intended to explain how the system works; *scrutability*, allowing users to tell the system if it is wrong; *trust*, increasing users' confidence in the system; *effectiveness*, helping users to make good decisions; *efficiency*, helping users to make decisions faster; *persuasiveness*, trying to convince users to select the given item; and *satisfaction*, increasing the ease of use of a system. They argued that these goals should be identified as distinct, even if they may interact [26]. Most previous studies on *generating* explanations optimize a single goal [20], and only a handful consider multiple goals [8, 13, 26]. Yet, depending on the perspective of the explanation generator, different goals may be appropriate, and may need to be traded off.

We ask three key research questions about such goals:

RQ1 How can one robustly measure if an explanation (with a particular goal) provided with a recommendation creates the intended effect on the recipient?

RQ2 Can ordinary people write explanations that optimize a given goal, and if so, how does that target goal affect how the explanation is perceived by recipients of recommendations?

RQ3 How do different goals relate to each other, or more specifically does optimizing particular goals reduce or increase the extent to which other goals are satisfied? Can recipients of recommendations even distinguish the goals from each other?

For example, how does optimizing persuasiveness affect trust? What is the trade-off between effectiveness and efficiency? To the best of our knowledge, there have not been any holistic studies of the interaction between goals. This work aims to fill that gap.

Taking advantage of the fact that people are commonly able to explain their own recommendations, we perform a set of user studies to assess both the generation and measurement of explanations. In our work, recommendations and explanations are generated, and evaluated, by people using a crowdsourcing platform. We study if people can optimize given goals, and how to efficiently measure whether different goals are satisfied. We present and compare two alternative designs for this evaluation: an item-wise setting, and a list-wise setting, each with strengths and weaknesses.

SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8016-4/20/07.
https://doi.org/10.1145/3397271.3401032

Table 1: Explanation goals and their definitions [26].

Goal	Definition
Effectiveness	Help users make good decisions
Efficiency	Help users make decisions faster
Persuasiveness	Convince users to try or buy
Satisfaction	Increase the ease of use or enjoyment
Scrutability	Allow users to tell the system it is wrong
Transparency	Explain how the system works
Trust	Increase users' confidence in the system

Our main findings are: (1) Item-wise evaluation of explanations is more sensitive than list-wise evaluation; (2) There are strong correlations between the measured values of different goals, with some goals (such as satisfaction) being correlated with many others, while other goals (such as scrutability) being more distinct; (3) The intended goal that the person is asked to optimize has a large effect on explanation quality, yet is not always the one with highest measured rating.

In summary, our main contributions are twofold. First, we develop an experimental protocol and multiple survey designs for evaluating explanations for item recommendations. Second, we present an analysis of explanations generated for a selected domain. These lead to a number of specific recommendations for evaluation of explanations in general.

2 RELATED WORK

The ability for an artificially intelligent system to explain recommendations has been shown to be an important factor for user acceptance and satisfaction [13, 14, 23, 28]. Explanations can be characterized along a number of dimensions, including their content, form of presentation, and system's intended purpose [20]. Our interest is in the latter category, where we use the term *goal* to refer to the objective or purpose of the explanation. Specifically, our focus is on natural language explanations, the most commonly used way of presentation both historically [20] and recently [2, 6, 19].

2.1 Explanation Goals

We use the seven explanation goals identified in [26] as a basis; these are listed in Table 1. We note that there are possible refinements to these goals. For example, in [20] satisfaction is not considered as a single objective, but is split into ease to use, enjoyment, and usefulness. Nonetheless, these seven goals are regarded as the canonical categorization within explainability research for recommender systems, accurately reflecting the goals that have been studied in the past. Certain goals may be measured objectively and quantitatively. For example, *effectiveness* may be measured as the change of a user's rating of (or reported interest in) an item before and after consuming that item [3, 6], *efficiency* may be measured by time spent on rating an item [13] or reading an explanation [6], and *persuasiveness* may be measured in terms of click through rate [30]. Here, we aim to compare different goals on equal footing, and thus focus on the subjective perception of the recipient—measured at the time when a recommendation and explanation are shown.

Most past studies are concerned with a single goal [20], and there is evidence each can be achieved individually [26]. The interactions between two or more goals, however, are much less understood. The most common explanation purpose, according to a large-scale literature review by Nunes and Jannach [20], is *transparency*, which is also considered key to building user *trust* [12]. Concerning the relationship between the two, one previous study indicates that transparency increases user trust [23], while another study finds that transparency and trust are not related [8]. The second most frequent explanation purpose is *effectiveness* [20], which can be conflicting with *persuasiveness* [7]. A systematic evaluation of explanations with respect to all goals has not been performed before.

2.2 Generating Explanations

There is an important recognized difference between *explanations* (why a certain suggestion is given) and *justifications* (why the user may be interested in the item) [19, 27]. The former consist of an honest account of the mechanism that generated the suggestion, while the latter provides a plausible reason, which may be decoupled from the underlying recommendation algorithm.

There is a growing interest in generating natural language explanations and justifications. Given a sophisticated recommendation system, justifications may often be provided by filling in natural language templates, for example, by considering simple features such as actor and director names [24] or by extracting relevant and distinguishing characteristics from reviews [19]. However, our work focuses on explanations. Justifications have in the past been created manually using crowdsourcing [6]. A main difference between that and ours, is that we ask humans to pick the recommendation as well as explain it, while [6] perform only the latter.

To summarize, by obtaining recommendations and explanations from the same person, we can focus on explanations rather than justifications; we also believe the explanations obtained are genuine. Similarly, we believe recommendations are better when accompanied by an explicit explanation. An alternative of taking items from a state of the art recommender system accompanied with human-generated justifications would introduce further limitations in any analysis.

2.3 Evaluating Explanations

Subjective perceptions of explanations are often evaluated qualitatively based on user surveys, with responses typically given on Likert scales [6, 8, 10, 17, 21–23]. Following standard practice, we design a user survey to capture the subjective perception of users regarding the seven goals.

3 EXPERIMENT DESIGN

This section presents our approach for generating and evaluating explanations accompanying item recommendations. We designed it to test the hypothesis that people are capable of writing explanations that satisfy different goals (RQ2), and that these explanations influence how the recipient of the explanation feels about the recommendation and the recommendation system (RQ1). Measuring all goals for the sample explanations allows the study of how the goals interact (RQ3). Given that this work is a first attempt to perform such a holistic evaluation of the effect of explanations on recipients, a number of the steps allow for different designs. Our experiments test a number of alternative approaches, which we anticipate to be further refined in future work.

Overall, our experiment involves three main steps, which are depicted in Figure 1. At a high level, these are:

Step 1 Test subjects are recruited and asked about their preferences. They provide an informative description of items they like/dislike, as well as specific examples, that another person can interpret to make a recommendation.

Step 2 Crowd workers are tasked with recommending items for test subjects from a pool of options. These workers also write a short explanation for why they made this particular choice, instructed to be written to serve a particular goal. Together with the item recommended, these explanations are treated as candidate explanations, and are filtered by other crowd workers in Step 2b to ensure high quality.

Step 3 The original test subjects are provided with the recommendations, and evaluate the explanations selected by completing a questionnaire. We compare two designs for this evaluation, for sensitivity and consistency.

We detail each of these steps below, although we defer a detailed presentation of the mechanics of experimental conditions, and matching crowd workers to test subject profiles, to Section 4. Also, the experiments in this study are performed on the movies domain, and hence some of the instructions have been specifically tailored to that. Nevertheless, the same experimental design is applicable to any item recommendation and explanation domain.

3.1 Step 1: Eliciting User Preferences

Test subjects are recruited via a crowdsourcing platform. To be able to generate personalized recommendations, we first need these subjects' preferences. This is done by asking them to fill out a short questionnaire, consisting of the following four questions:

(1) What sort of movies do you like?
(2) Name three of your favorite movies.
(3) What sort of movies do you dislike?
(4) Name three movies that you really disliked (or hated).

This mirrors narrative-driven recommendation, often seen extemporaneously in forums, where the recommendation is driven by both information about the user's past transactions (positive/negative examples) and a narrative describing desired items [4].

For the first and third questions, the required answer length is minimum 150 characters. Further, we manually checked the responses to ensure the user profiles are of high quality before inviting the test subject to continue as part of the experiment.

3.2 Step 2: Generating Explanations

We engage a large pool of crowd workers to generate personalized item recommendations along with explanations, based on the user interest narratives provided in Step 1. This yields a set of candidate recommendations for each test subject.

We also note that, while our design could have been to generate explanations programmatically, this would inherently reflect just one particular interpretation of each goal—that of the designers of the algorithm and introduce another important variable in the analysis. Rather, having people explain their recommendation using a goal-focused prompt provides a better insight into how a given goal should be satisfied. We now detail how this was done.

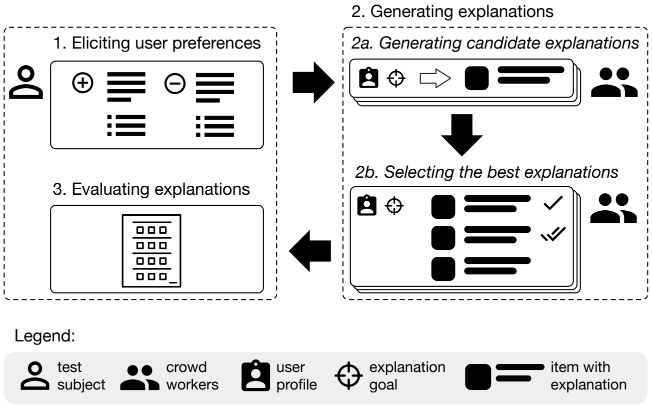
2. Generating explanations
1. Eliciting user preferences
2a. Generating candidate explanations
2b. Selecting the best explanations
3. Evaluating explanations
Legend:
test subject
crowd workers
user profile
explanation goal
item with explanation

Figure 1: Overview of our experimental design.

3.2.1 Generating Candidate Explanations. There are two main design considerations in this step. One is to anchor the task in a realistic setting that workers can easily relate to (referred to as the *context of use* in [4]). Therefore crowd workers were presented with the scenario where they have to recommend a movie to a friend who will be traveling on a plane. The worker is provided with the target person's self-described preferences, i.e., descriptions and specific examples of movies liked and disliked. The pool of available movies is described as coming from the flight entertainment system's limited selection of movies.

The second design consideration is to impose a limit on the pool of items so that workers can reasonably familiarize themselves with the candidate set. We thus solicited recommendations from one of three different pools, each comprising 20 distinct movies. The pools were manually created to be disjoint yet contain a wide variety of movies to suit different tastes consisting of (1) Hollywood movies, (2) independent movies, and (3) recent movies.

In addition to selecting a movie to recommend, the crowd worker was asked explain the recommendation using a short description based on a goal-specific prompt. We required it to be at least 100 characters long. We note that the wording of the goal is key to our experiment. As we expected that the specific wording of the goal could affect the perceived adequacy of explanations, we tested two different wordings for each of the seven goals. In a preliminary evaluation (described in Section 3.5), we selected among these wordings.

3.2.2 Selecting the Best Explanations. To ensure the highest quality of recommendations and explanations, a filtering process is next employed to select the highest quality explanations from those created for the recipient. This two-step process resembles the MapReduce paradigm, which has been shown to be effective for solving complex problems through micro-task platforms [16]. Specifically, we group the explanations generated for each user and goal pair. These are shown to an independent set of crowd workers, who are asked to select the best (item, explanation) pair. Table 2 lists examples of explanations voted as best for each target goal.

Table 2: Examples of recommendations and explanations generated by crowd workers.

Goal	Recommendation	Explanation
Effectiveness	Pulp Fiction (1994)	Because you have a hard time with a lot of PG or PG 13 movies and feel that they aren't real enough for you to enjoy I think you'll find this movie very gripping and real. It will keep you aware of harsh realities of life.
Efficiency	Fantastic Beasts: The Crimes of Grindelwald (2018)	You would enjoy this movie as it is family friendly, doesn't have a lot of violence and no nudity. It has a lot of magical creatures and is just a fun film.
Persuasiveness	Love, Simon (2018)	Love Simon is a heartwarming story. It involves Simon, a young boy who is coming of age. Get ready to laugh, smile, and maybe even cry, as you follow along with Simon's daily life.
Satisfaction	A Quiet Place (2018)	I chose this movie because from your preferences I can tell that you enjoy movies which have some suspense and plot twists. This movie does this at a very high level.
Scrutability	Moneyball (2011)	I chose this awesome movie for you since you really seem to like feel-good movies that have likable characters and nice stories.
Transparency	The Lion King (1994)	Since you love love romantic comedies, comedy, mystery, some action, documentaries, biographies, Lifetime Movie Network movies, Hallmark Movies, this is a combination of them all. You will be sure to like.
Trust	Back to the Future (1985)	I think you'll like Back to the Future. It has a bit of everything including humor, drama, and sappy love. Yes, it is science fiction, but there is so much more to it that I think you will enjoy it. I think the storyline, the acting, and the dialogue supersede your possible dislike of the genre. Besides, most of the plot takes place in the 1950s and the 1980s.

Table 3: Survey questions for evaluating explanation goals.

Goal	Statement This explanation ...	Reverse statement This explanation ...
Effectiveness	helps me to determine how well I will like this movie	does not help me make a decision about this item
Efficiency	helps me to decide faster if I will like this movie	does not save me time
Persuasiveness	makes me want to watch this movie	fails to make this item appeal to me
Satisfaction	would improve how easy it is to pick a recommendation	does not satisfy me
Scrutability	would allow me to give feedback on how well my preferences have been understood	would make it difficult for me to correct the reasoning behind the recommendation
Transparency	helps me to understand what the recommendation is based on	fails to reveal the reasoning behind this recommendation
Trust	helps me to trust the recommendation	does not seem credible

3.3 Step 3: Evaluating Explanations

Finally, the test subjects are presented with the recommendations created for them, and asked to evaluate the corresponding *explanations* by filling out a questionnaire.

We develop two different experimental designs so as to evaluate the extent to which the explanations, generated to serve a specific goal, succeed in meeting that goal. The designs are illustrated in Figure 2. Our first is an *item-wise* evaluation design, which is widely used in the literature [6, 25, 27]. Here, test subjects are presented with a single recommended item along with a single explanation. We also propose an alternative *list-wise* evaluation design, which gives the test subject *three* recommended items, each from a distinct pool of movies, along with explanations for each, all for the same goal. We expect that the item-wise design has a lower cognitive load since users need to consider a single explanation. At the same time, we hypothesize the list-wise design to yield more robust observations, as responses are less likely to be influenced by the quality of a single explanation. We will therefore analyze to what degree the results obtained with the two designs align with each other.

In both cases, the test subject is asked to fill out a survey below the recommendation(s) and explanation(s). The survey consists of five parts:

(1) A summary of the preferences provided earlier, meant as a refresher, displaying the data provided by the test subject in the preference elicitation phase (i.e., interest characterizations as well as names of movies liked/disliked).

(2) An item recommendation (item-wise design) or a list of three item recommendations (list-wise design), accompanied by explanation(s).

(3) Seven statements, presented in random order, each targeting a specific goal, to be rated on a 4-point Likert scale. These are shown in the Statement column of Table 3. The Likert scale used was (1) not at all, (2) slightly, (3) moderately, and (4) a great deal.

(4) Seven reverse statements, to check the consistency of answers provided, are presented separately to avoid confusion [11], also in random order and each targeting one specific goal. These statements, listed in the Reverse statement column in Table 3, were created with a negative wording, to be rated on a 4-point Likert scale: (1) strongly disagree, (2) moderately disagree, (3) moderately agree, and (4) strongly agree.

(5) We ask how personalized the explanation(s) felt on a 5-point Likert scale: (1) not at all, (2) a little, (3) a moderate amount, (4) a lot, and (5) a great deal. In the item-wise design, where recommendations are a single item, we also ask test subjects whether they have seen that movie or not. Depending on the answer, we ask a follow-up question "Would you watch it?" or "How did you like it?" In both cases, answers are given on a 5-point Likert scale. Subjects were further required to fill out a free-form text box describing what improvements they thought the explanation(s) needed.

For each test subject, parts (2)–(5) are repeated three times (limited to avoid survey fatigue) each time showing different items as recommendations (i.e., no movie is recommended more than once). Each iteration presented explanation(s) created by crowd workers targeting different goal(s).

The wording of survey questions is based on the definitions of the goals in [26] and is inspired by questions asked in prior work (specifically, for effectiveness [10], persuasiveness [17, 22], satisfaction [6, 22], transparency [8, 10], and trust [6]).

Table 4: **Results from goal wording calibration study. The two alternative wordings of instructions for each explanation goal are presented on the right hand side. The numbers indicate the percentage of the votes received, among all explanations generated with the given goal, by the calibration voters.**

Goal	Alternative instructions for generating the explanation				Instruction for voters
	Write the explanation such that ...		Write a short description for your friend ...		Select the explanation that...
Effectiveness	it helps your friend to make a good decision whether they will like the movie.	53%	to help them decide if this is the best movie for them among other recommendations they may receive.	47%	would help most to make a good decision.
Efficiency	it makes it faster for your friend to decide if they will like the movie.	51%	to help them quickly decide if they are likely to enjoy this movie more than other recommendations they may receive.	49%	would help most to make a quick decision.
Persuasiveness	it persuades your friend to watch this movie.	49%	to try to convince them that they should watch this movie.	51%	sounds the most convincing.
Satisfaction	your friend would find the most useful.	47%	to make it more likely that they will enjoy seeing the list of recommended movies.	53%	is the most useful.
Scrutability	it highlights the assumptions you've made, so that your friend could correct them if they're incorrect.	56%	that highlights aspects of the movie that you are unsure if they will like.	44%	makes it easiest to tell the system if it misunderstood the user's preferences.
Transparency	it shows your reasoning when deciding to recommend this movie.	58%	that explains how you decided to recommend this movie.	42%	best explains the decision behind the recommendation.
Trust	it increases your friend's trust in the recommendations you give them.	49%	so that they know that you considered their preferences when making the recommendation.	51%	is the most trustworthy.

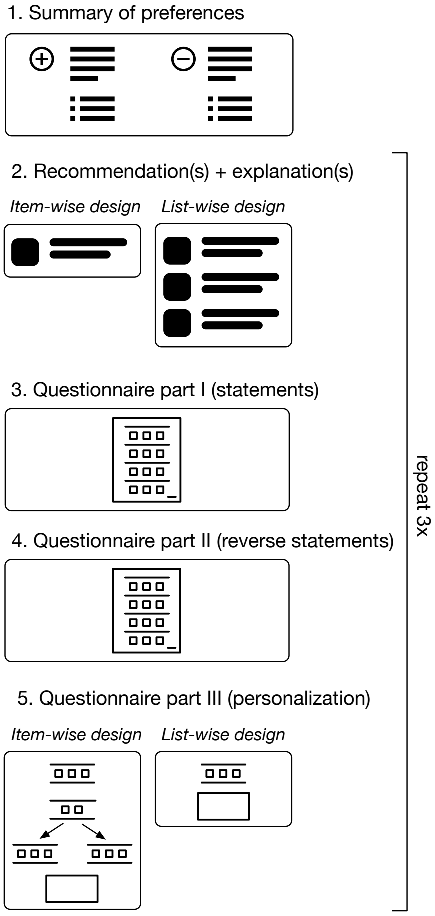
1. Summary of preferences
2. Recommendation(s) + explanation(s)
Item-wise design
List-wise design
3. Questionnaire part I (statements)
4. Questionnaire part II (reverse statements)
5. Questionnaire part III (personalization)
Item-wise design
List-wise design
repeat 3x

Figure 2: Survey designs for evaluating explanations.

3.4 Further Quality Considerations

It is known to be difficult to identify answers provided by malicious or inattentive users when collecting user opinions, and questions do not have verifiable answers [15]. Since measurements are subjective and qualitative, we take a number of precautions for quality control: (1) We restrict participation to crowd workers who are experienced and have a reputation for delivering high quality work; (2) We check for inconsistencies in ratings given to the positive and reverse statements, removing participants where the score of the positive and negative statements disagreed by more than one point at least three times; (3) We require participants to fill out a free-text box, ensuring they had attended to the task.

3.5 Goal Wording Calibration

We hypothesize that the instructions given to crowd workers writing explanations affect how the explanations are perceived. Therefore, our crowdsourcing experiment was preceded by a calibration study to select the exact wording of instructions that resulted in the best explanations for each goal.

Specifically, 5,188 recommendations and explanations were generated for 120 test subjects (for all seven goals for each test subject), where two alternative wordings of instructions for each goal were used. This resulted in 6.18 explanations on average per (test subject, goal) pair. Then, a second group of crowd workers were asked to select the single best explanation that best serves the given goal from this set. The two alternative wordings with the corresponding statistics, as well as the wording given to voters, are shown in Table 4.

We test the results for significance using a Binomial test. Statistically significant differences at the 95% confidence level are found for two of the goals, namely Scrutability and Transparency. This confirms our hypothesis that wording can have a strong impact. In our main experiment, detailed below, we used the instruction for each goal that received the higher percentage of votes as shown in the table.

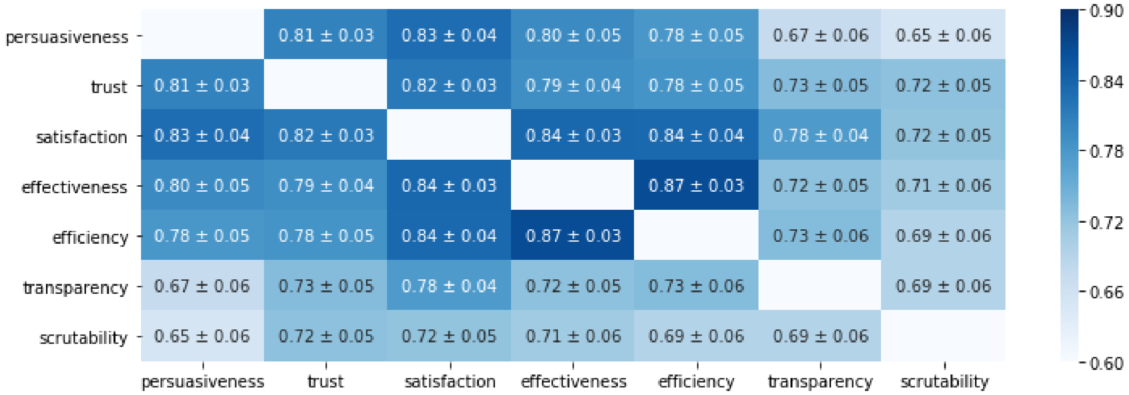

Figure 3: Pearson Correlation between different metrics across all goals, with 95% confidence interval using bootstrap sampling. (The order of goals is rearranged for improved readability.)

4 DATA COLLECTION

We performed our user study on a large crowdsourcing platform. Our worker pool consisted of workers in the United States with a high (over 98%) task approval rate. Personal worker information, including demographics, was not judged essential hence was not collected. However, the worker pool is known to consist of diverse crowd workers.

4.1 Mechanics and Experimental Conditions

4.1.1 Step 1: Eliciting User Preferences. In Step 1 of our study, we collected preferences from 240 test subjects (after filtering out low quality responses). There were no experimental conditions, as all test subjects experienced the same survey.

4.1.2 Step 2: Generating Explanations. The second step consisted of two sub-parts. In Step 2a, crowd workers were tasked with generating personalized recommendations and corresponding explanations with a given target goal and movie pool as the experimental conditions. Each micro-task asked for recommendations for 5 different user profiles with the same target goal. We solicited recommendations from three separate movie pools (Hollywood, independent, and recent movies), each consisting of 20 highly rated movies. To ensure diversity, workers were restricted to doing at most 10 micro-tasks. A total of 7,340 personalized explanations were generated by 702 crowd workers.[1] The average task completion time was 534s, that is, 107s for making a single recommendation and writing the corresponding explanation.

In a subsequent filtering step (2b), other workers were asked to vote on the best explanation for each test subject and goal pair. Each micro-task contained explanations targeting a specific goal (3.4 explanations on average) for 5 test subjects.

Table 5 displays descriptive statistics of the data collected (after filtering participants with three or more inconsistent responses).

Baseline. As a baseline, “neutral” explanations were also collected. Specifically, the description of movies (limited to 500 characters) was extracted from the information panel of a major web search engine. We note that these are not personalized, and are usually written by human experts per movie. As such they are likely to have high quality.

Table 5: Statistics of collected data.

Design:	**Item-wise**	**List-wise**
Test subjects	84	88
Median survey completion time (sec)	325	349
Number of observations	239	254

4.1.3 Step 3: Evaluating Explanations. In Step 3, each test subject was presented with explanations for three different goals. The experimental conditions were the combination of goals and the experiment design (item-wise or list-wise). Each test subject was assigned a combination of three goals in a Round Robin fashion, such that the number of observations for each combination of goals is approximately the same. The order of the three goals was randomized for each survey. For each target goal, the best (i.e., highest voted) explanation was chosen across the three movie pools, with the corresponding movie shown as the recommendation.

To compare to the baseline explanation, each test subject was also presented with a fourth case: We selected the movie that was recommended by the most workers, but that had not been selected for any of the three goals evaluated for this particular subject. This movie was then shown to the test subject along with the neutral (baseline) explanation.

We made sure that the same movie is not repeated for a given subject in case of the item-wise design, and that the list of suggestions always had three distinct movies for the list-wise design. To avoid low quality responses, those with a too short completion time (below 60 seconds) were removed.

A total of 107 study subjects provided valid responses in this final step (amounting to a 44.5% return from Step 1). We obtained 84 responses for the item-wise design and 88 for the list-wise design, with 65 participants completing both.[2] Each target goal evaluated by a study subject constitutes an observation. Thus, the total number of observations is three times the number of responses.

[1] Since our surveys evaluate three goals for each study subject, we only generate explanations for those goals that will actually be shown in the surveys, and not for all possible subject-goal pairs.

[2] For those that participated in both designs, at least one week elapsed between the two surveys.

5 RESULTS

The aggregate results of the perceived quality of the recommendations generated for each goal are presented in Table 6. In this section we present a detailed analysis of these results. As a high level summary, we found that different goals have a strong effect on how users reacted to explanations. We found that some pairs of goals are significantly more correlated than others. We also found that the item-wise design was most sensitive.

5.1 Metric Correlation

We start the analysis by considering the different goals as metrics for the quality of explanations, independently of how they are generated. Thus, for the moment, the approaches used to *create* the explanations are not relevant. Rather, we consider how test subjects respond to recommendations along the seven goal dimensions, and how the different goal metrics interact. For example, does higher Persuasiveness correlate with lower Trust?

Figure 3 shows the Pearson Correlation between each pair of metrics, by analyzing the responses of each user who was presented with a recommendation across both experimental conditions. For each correlation, the figure also shows a 95% confidence interval for the correlation measured using bootstrap sampling. Note also that we compared the correlations in the item-wise and list-wise designs, as well as comparing the correlations taking only seen versus unseen movies, or highly rated versus low rated recommendations. In all cases, we found that the correlations do not differ substantially, nor with statistical significance.

We find that the explanation metrics are strongly correlated on average, yet exhibit clear structure. All seven goals have at least moderate correlation. Effectiveness and Efficiency are particularly strongly correlated. We see that five goals, namely Persuasiveness, Trust, Satisfaction, Effectiveness and Efficiency, appear to move together. It is particularly interesting to see that for instance Persuasiveness and Trust are strongly *correlated*. On the other hand, Transparency and Scrutability stand out as less correlated with each other, and with all the other five goals.

One possible reason is that there is only a limited amount of information that can be conveyed in a few hundred characters, and hence it is difficult to tailor an explanation to a particular goal. Further, we hypothesize that a different result may be observed if explanations were longer and more detailed. On the other hand, for typical use-cases, this is about the desired explanation length because of limited screen real-estate. As argued by Tintarev and Masthoff [25], "brevity is important in a context where the user has to review many possible options." An alternative possible reason considered was that crowd workers may not be very good at creating explanations for any given goal, and that the explanations rather optimize a goal that depends more on the worker writing the explanation than on the instructions provided. However, our results in Table 6 show this not to be the case, as the different goals presented to workers generating explanations result in significantly different *values* of the scores for each goal.

We conclude that Satisfaction is most correlated metric with all goals for explanations, and is the single most predictive goal. The most distinct secondary metric is Scrutability, which is also consistent with intuition that an invitation for feedback may create a different user experience. Finally, Transparency is third most distinct metric.

An important limitation that must be considered in our design is whether the questions proposed in Table 3 actually measure the goals intended to measure. Even though there exist ways to measure individual explanation goals, there is no established evaluation approach for comparing and contrasting multiple goals systematically and holistically, thus a direct comparison to other approaches is not possible. Following DeVellis [11] recommendations for scale development, we measure the internal consistency of each metric using Cronbach's α [9] for each goal. As shown in Table 7, we find that most metrics have suitable internal consistency, with Scrutability and Trust somewhat lower.[3] Thus we see that the answers to the positive and negative questions mirror each other as designed, despite the questions intentionally using significantly different wording. While refinements to the questions are likely possible, we are confident that the goal labels correspond to the intention of the questions.

5.2 Goal-Metric Agreement

Next, taking the goals targeted into account, we start with a surface level analysis of the agreement between the goal targeted by the crowd worker, and the ratings provided by the test subject for explanations.

Table 6 shows each goal targeted as one row, with the number of test subjects who received recommendations for this goal indicated. We show the mean score for each metric, recalling that these were collected on a Likert scale of 1 to 4, taking the average of the positive and reverse statement scores from Table 3 for each test subject. An ↓ symbol indicates that the score for that metric is statistically significantly worse than the maximum for the goal optimized (using a two sample t-test, with 95% confidence).

We may expect a diagonal form for the matrix: the maximum of each metric may be expected to coincide with the same goal being targeted by the people who produced the explanations. This is not the result seen. Rather, we observe that some target goals are much more effective at producing high scoring explanations across all metrics. This tells us that the wording of the goal provided to the crowd workers is critical in determining the quality of the explanations. Even if some of the goals are considered more important to a system designer, it may be that better explanations are obtained by instructing crowd workers to optimize a *different* goal.

A second key observation is that the neutral explanation performs particularly well. Recall that this is a synopsis taken from the information panel of a major web search engine. Thus it is not personalized to the user, and usually consists of basic metadata about the movie recommended. One possible explanation that must be excluded is that crowd workers may be particularly poor at generating explanations for a given goal, especially in comparison to a neutral sentence that is likely to have been reviewed and refined by a number of experts. However, the crowd workers are always capable of creating explanations that are equally as good as the neutral explanation, and the explanations have been filtered in Step

[3] We also tested the Cronbach's α *without* filtering the 29 (6%) raters with many inconsistent responses. This results in Cronbach's α values around 0.07 lower. However, as crowd workers are known to produce answers with variable quality, we consider the filtering appropriate for our setting.

Table 6: Quality of explanations produced by humans optimizing different goals. Each row corresponds to one goal given to workers (i.e. this should be considered as an algorithm). For each metric, the highest performing goal is in bold. All goals with performance statistically significantly worse from the best are marked with ↓ (p<0.05, two sample t-test), with all others shaded.

Item-wise Experiment Design

Goal	**Evaluation Metric**						
Targeted	Persuasiveness	Trust	Satisfaction	Effectiveness	Efficiency	Transparency	Scrutability
Persuasiveness (n=33)	2.86	2.94	2.71	2.68↓	2.67↓	2.83↓	2.70↓
Trust (n=29)	2.64↓	2.76↓	2.67	2.64↓	2.69↓	2.90↓	2.81
Satisfaction (n=29)	2.97	3.03	2.95	3.00	3.00	3.02↓	2.74↓
Effectiveness (n=26)	**3.27**	**3.25**	2.98	3.12	3.06	**3.37**	3.13
Efficiency (n=30)	2.73↓	2.95	2.77	2.77↓	2.98	2.85↓	2.73↓
Transparency (n=30)	3.13	3.20	**3.07**	3.07	**3.20**	3.20	**3.15**
Scrutability (n=33)	2.77↓	2.97↓	2.73	2.82	2.85	2.98↓	2.77↓
Neutral (n=29)	3.12	3.24	3.00	**3.16**	3.05	2.76↓	2.98

List-wise Experiment Design

Goal	**Evaluation Metric**						
Targeted	Persuasiveness	Trust	Satisfaction	Effectiveness	Efficiency	Transparency	Scrutability
Persuasiveness (n=32)	3.14	2.97	2.97	3.08	3.02	3.05	2.92
Trust (n=30)	3.07	2.97	3.07	2.95	3.00	**3.20**	3.12
Satisfaction (n=31)	3.11	3.10	3.00	3.06	**3.16**	3.19	**3.15**
Effectiveness (n=30)	3.10	3.12	3.07	3.05	3.15	3.20	3.02
Efficiency (n=29)	2.79↓	2.91↓	2.88	2.81↓	2.97	2.97	2.93
Transparency (n=38)	2.99	2.92↓	2.87	2.95	3.00	3.12	3.07
Scrutability (n=33)	2.95↓	2.92↓	2.92	3.00	3.00	2.98	2.85↓
Neutral (n=31)	**3.29**	**3.26**	**3.11**	**3.23**	3.15	3.00	3.11

Table 7: Internal consistency (Cronbach's α) of the pair of questions asked of raters for each goal measured (n=493).

Goal Measured	**Cronbach's α**	**Interpretation [11]**
Effectiveness	0.81	Very Good
Efficiency	0.83	Very Good
Persuasiveness	0.86	Very Good
Satisfaction	0.81	Very Good
Scrutabilty	0.67	Minimally Acceptable
Transparency	0.80	Very Good
Trust	0.74	Respectable

2b of our experiment. This raises a question of potential headroom, whether it is possible to outperform a neutral summary with an explanation consisting of 100-200 characters on average.

5.3 Experimental Design Analysis

Next, we compare the two experiment designs. We can see in Table 6 that the item-wise experiment design is much more sensitive to differences among the different goals. There are almost no statistically significant differences in metric performance in the list-wise design across all goals, given very similar sample sizes across the two designs. This surprising result suggests a further analysis of the limitations of the two designs. In particular, we expected that by presenting a single item, the perceived quality of the explanation would be strongly influenced by the specific item recommended. Recipients may react to the *explanations* differently based on whether the *item* recommended appears relevant or not, especially if they have already watched the movie. We assess this next.

Figure 4 presents a more fine-grained view of the results, breaking down the item-wise design split by these two effects. The top row splits the explanation ratings by whether or not the recommendation was considered *good*. Specifically, if the test subject had seen the movie, they were asked if they liked the movie. Otherwise, the test subject was asked if they would watch it. Ratings of 4 or 5 were considered "rated highly", the others were considered as "rated low". We find that most (69%) of recommendations were considered good. We also see that there is a large difference in mean score across all values, supporting our hypothesis.

Finally, the bottom row in Figure 4 splits the recommendations by whether the test subject had already seen the movie or not. This is roughly an even split, and we see a similar effect: Explanations for movies the user has already seen receive substantially different scores.

Returning to the motivating task for this paper, namely that of making and explaining recommendations, we believe that the results on unseen items are the most important condition — such recommendations provide the most value to the recipients. Combining with the observation that the item-wise approach is more sensitive, this suggests that future work on evaluating the impact of explanations should filter evaluations for items that have not already been seen (or otherwise consumed by) the test subject to avoid seen items affecting the measured performance metrics. It is important to note, however, that despite these differences by item rating and consumption seen here, the *correlation* between metrics is *not* affected by these conditions.

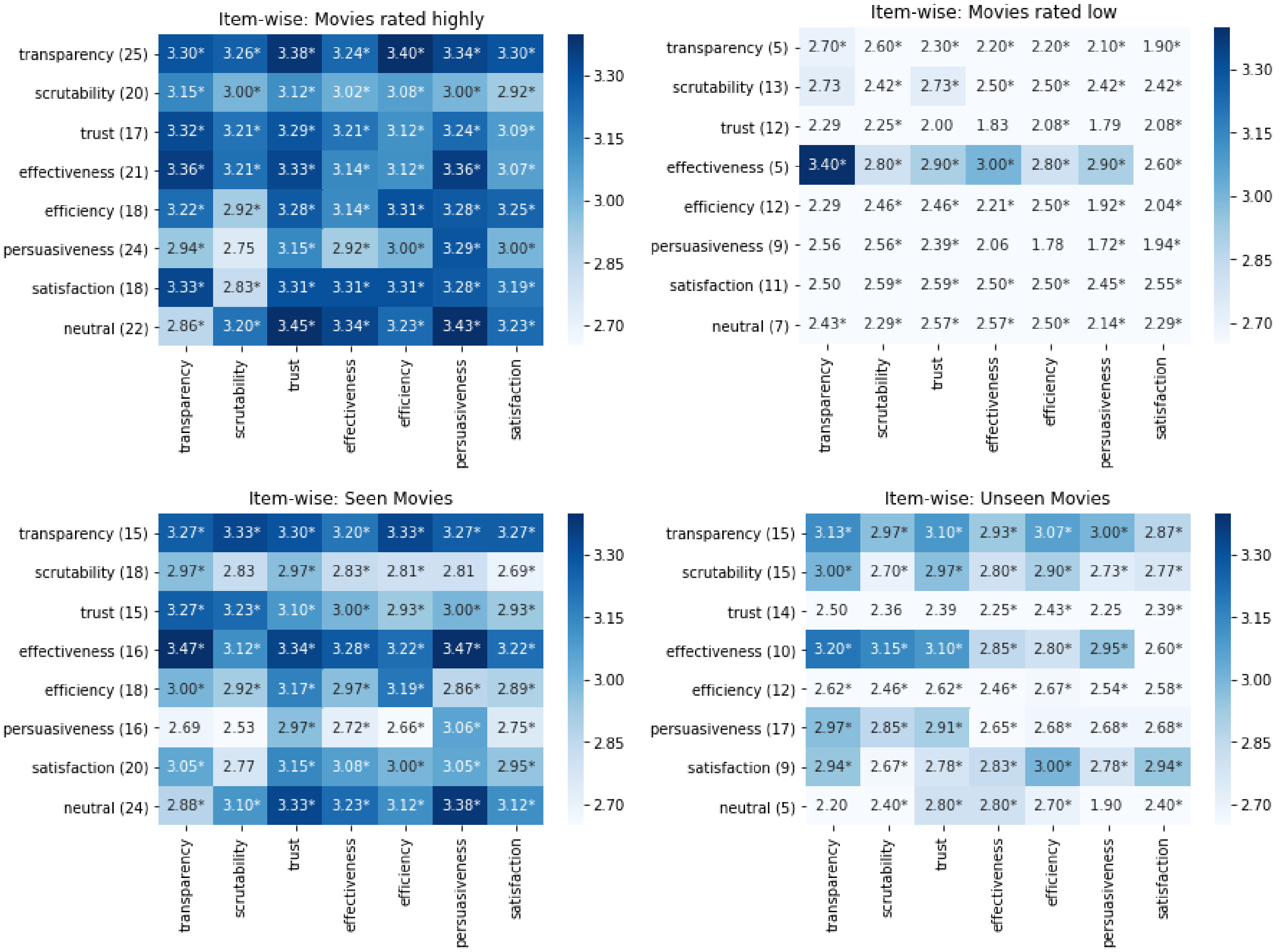

Figure 4: Performance of item-wise design on different subsets. For each metric, all goals with performance that is not statistically distinguishable from the best, are marked with * (p<0.05, two sample t-test).

6 DISCUSSION

We now revisit our research questions and answer them based on the experimental results obtained. We further suggest some points of advice for evaluation of the impact of recommendation explanations on people who receive such explanations.

(RQ1) How can one robustly measure if an explanation provided with a recommendation creates the intended effect on the recipient? We compared two alternative survey designs, consisting the same series of question and reverse pairs for each explanation goal. These questions showed good internal consistency. The item-wise experimental design appears statistically more powerful than the list-wise one, although it is likely to be beneficial to filter for items that are novel recommendations to the recipient (i.e., the user has not seen/consumed them).

(RQ2) Can ordinary people write explanations that optimize a given goal, and if so, how does that target goal affect how the explanation is perceived by recipients of recommendations? We showed that crowd workers can indeed generate explanations that are equally as good as neutral explanations written by experts. The precise wording of the goal of the explanations is found to be key to obtaining high quality explanations. However, the wording need not necessarily align with the goals considered most important by a system designer.

(RQ3) How do different goals relate to each other, or more specifically does optimizing particular goals reduce or increase the extent to which other goals are satisfied? Can recipients of recommendations even distinguish the goals from each other? We found that all seven goals are moderately correlated, while some pairs are particularly strongly correlated with each other. As measurement of the impact of explanations on recipients is expensive, it appears that Satisfaction, Scrutability and Transparency — if they are desirable properties for a given system — may provide the most complete assessment of explanation quality across the seven established goals from [26].

6.1 Limitations

As a first analysis of the interaction of different goals for explanations, this study is not without limitations. First, we performed this analysis in a single domain, with a single length limit to explanations. It must be verified that our findings transfer to domains beyond movies. It must also be assessed whether the results generalize to recommendations made by a state-of-the-art algorithm rather than crowd workers. Differences among individuals generating explanations, as well as individuals receiving explanations, may affect results—although this effect would be better teased apart at larger scales where fewer experimental design variants are considered. Expert copy-writers may also be able to generate more compelling explanations.

From the perspective of measurement, it is likely that there exist refinements to the questionnaire that would increase sensitivity and internal consistency. It is also likely that further refinements to the experiment design would increase sensitivity. It should be noted that the inherent correlation and dependencies of some of the goals makes this particularly difficult to control for in any design.

7 CONCLUSION

We have presented a first analysis of how different goals, or intentions behind recommendation explanations, can be measured. We found that when treated as metrics, the seven goals are often highly correlated, yet exhibit clear structure and interact. This suggests that it may not be necessary to separately consider so many distinct goals. In asking crowd workers to generate explanations, we also found that the wording of the goal has a strong impact on the quality of the explanations as perceived by the test subjects. Two experiment designs were presented, and found to exhibit similar patterns yet different sensitivity and some biases depending on the *items* recommended rather than just the explanations.

This study represents a step towards the development of appropriate evaluation methodology for explainable recommender systems. Future directions concern the generalization of findings to other domains and further refinements to the survey and experimental designs. While our main focus has been on the measurement aspects, our methodology also facilitates the large-scale collection of high-quality explanations for a given goal. The collected data may be utilized for generating explanations automatically (for example, as training data for generative neural models).

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Article 582, 18 pages.

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

[3] Mustafa Bilgic and Raymond J. Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces.* 153–160.

[4] Toine Bogers and Marijn Koolen. 2017. Defining and Supporting Narrative-driven Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. 238–242.

[5] Bruce G. Buchanan and Edward H. Shortliffe. 1984. *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley Series in Artificial Intelligence)*. Addison-Wesley.

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

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

[8] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The Effects of Transparency on Trust in and Acceptance of a Content-based Art Recommender. *User Modeling and User-Adapted Interaction* 18, 5 (Nov. 2008), 455–496.

[9] Lee J. Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16, 3 (September 1951).

[10] Marek Czarkowski. 2006. *A scrutable adaptive hypertext*. Ph.D. Dissertation. University of Sydney.

[11] Robert F. DeVellis. 1991. *Scale development: Theory and applications*. Applied Social Research Methods, Vol. 26. SAGE Publications.

[12] Fatih Gedikli, Mouzhi Ge, and Dietmar Jannach. 2011. Understanding Recommendations by Reading the Clouds. In *E-Commerce and Web Technologies*. 196–208.

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

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

[15] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. 453–456.

[16] Aniket Kittur, Boris Smus, and Robert Kraut. 2011. CrowdForge: Crowdsourcing Complex Work. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. 1801–1806.

[17] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2017. User Preferences for Hybrid Explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. 84–88.

[18] Don Monroe. 2018. AI, Explain Yourself. *Communications of the ACM* 61, 11 (November 2018), 11–13.

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

[20] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Model. User-Adap. Inter.* 27, 3 (01 Dec 2017), 393–444.

[21] Pearl Pu and Li Chen. 2006. Trust Building with Explanation Interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI '06)*. 93–100.

[22] Masahiro Sato, Budrul Ahsan, Koki Nagatani, Takashi Sonoda, Qian Zhang, and Tomoko Ohkuma. 2018. Explaining Recommendations Using Contexts. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. 659–664.

[23] Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. 830–831.

[24] Panagiotis Symeonidis, Ros Nanopoulos, and Yannis Manolopoulos. 2008. Justified Recommendations based on Content and Rating Data. In *WebKDD workshop on Web Mining and Web Usage Analysis*.

[25] Nava Tintarev and Judith Masthoff. 2012. Evaluating the Effectiveness of Explanations for Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4 (01 Oct 2012), 399–439.

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

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

[28] L. Richard Ye and Paul E. Johnson. 1995. The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice. *MIS Quarterly* 19, 2 (June 1995), 157–172.

[29] Yongfeng Zhang and Xu Chen. 2018. Explainable Recommendation: A Survey and New Perspectives. *arXiv preprint* arXiv:1804.11192 (2018). arXiv:1804.11192

[30] 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 & Development in Information Retrieval (SIGIR '14)*. 83–92.