Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems

Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, Jürgen Ziegler
University of Duisburg-Essen, Duisburg, Germany
{firstname.lastname}@uni-due.de

ABSTRACT

Trust in a Recommender System (RS) is crucial for its overall success. However, it remains underexplored whether users trust personal recommendation sources (i.e. other humans) more than impersonal sources (i.e. conventional RS), and, if they do, whether the perceived quality of explanation provided account for the difference. We conducted an empirical study in which we compared these two sources of recommendations and explanations. Human advisors were asked to explain movies they recommended in short texts while the RS created explanations based on item similarity. Our experiment comprised two rounds of recommending. Over both rounds the quality of explanations provided by users was assessed higher than the quality of the system's explanations. Moreover, explanation quality significantly influenced perceived recommendation quality as well as trust in the recommendation source. Consequently, we suggest that RS should provide richer explanations in order to increase their perceived recommendation quality and trustworthiness.

CCS CONCEPTS

• **Information systems** → *Recommender systems*; • **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

Recommender Systems, Trust, Explanations, User Study, Structural Equation Modelling, Counterfactual Analysis

ACM Reference Format:
Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300717

1 INTRODUCTION

Contemporary online platforms typically rely on *impersonal* recommendation sources, i.e. automated Recommender Systems (RS), that automatically generate recommendations in order to faciliate users' decision making when facing a large number of alternatives. Even though recommendation algorithms have become highly accurate in terms of estimating a user's preferences [1, 15], they oftentimes appear as "black boxes" by concealing important details from their users. As a consequence, users create an unfitting mental model of the RS which may result in distrust and ultimately even in rejection of the system's recommendations [17, 46]. Hence, several researchers argue that especially the trustworthiness of a RS should be considered when assessing its quality [2, 22, 36].

RS are faceless entities lacking the human properties that are important for the development of trust, thus making it difficult for users to form bonds of any kind. One way to alleviate this, is to introduce social components into RS [6, 26]. There is a growing number of websites where automated and human-generated recommendations are combined—the latter, for example, in form of customer product reviews. For the reasons above, *personal* recommendation sources, i.e. users providing recommendations, are often associated with a higher trustworthiness [26, 44].

In the same line, designers of RS often strive to increase transparency and trustworthiness by providing textual explanatory components for recommendations [5, 46, 50]. A very common technique is to indicate similarity between recommendations and items the user is currently browsing or has expressed preferences for in the past. A well-known example for the former is Amazon's "Users who bought… also bought… " explanation. Similar kinds of explanations are applied by, for instance, Netflix and Spotify. Even though the effectiveness of such simplistic approaches utilizing similarity-based explanations has been questioned [5, 12], a thorough empirical comparison with systems using richer explanations—especially in terms of the perceived trustworthiness and its influential factors—is still missing.

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300717

We argue that overly simplistic explanations lack the expressiveness and social properties that are relevant to establish trust in a recommendation source. In order to find empirical support for this assumption, we conducted a user study in which we let participants assess recommendations that were either selected by another person or by a typical RS. Additionally, the recommended items were accompanied by individually composed explanations in the personal condition or similarity-based ones for the RS. By utilizing tools of causal statistical inference, i.e. *structural equation modeling* [33] and the *counterfactual framework* [18, 35, 40], we were able to reveal that the richness of explanations plays a pivotal role in trust-building processes. Although, as compared to the RS, humans were usually less accurate in estimating preferences, the explanations for their choice were more elaborate and comprehensible such that the overall quality of recommendations was deemed to be equal.

As a consequence, it appears reasonable to develop RS towards incorporating explanatory components that imitate more closely the way humans exchange information. Counterfactual analysis helped us answer questions about a hypothetical situation in which RS would do so: As it turns out, without any changes to the underlying algorithm, replacing similarity-based explanations with human-like ones, the quality of recommendations can be expected to improve by around 13 %.

Moreover, up to now research on trust in RS has been concentrating predominantly on an initial perception of trust and little research addresses temporal development of trust in recommendation sources [e.g. 8, 49]. Participants in our study received two recommendations over the course of two weeks, which allowed us to assess trust development over time. While trust in humans remained constant, we could observe a slight decrease for their automated counterparts. Although not statistically significant, we assume systematic effects that tackle information asymmetry and, through this, unfulfilled expectations.

The contributions of this paper can be summarized as follows:

- We conducted a user study that compared personal to impersonal recommendation sources. It is shown that there exist differences between the groups in how recommendations are perceived and in how bonds are created towards the recommendation source.
- We structurally model direct and indirect effects between constructs of major interest for RS research. Concretely, we reveal complex dependencies between *explanation quality*, *recommendation quality*, *social presence*, and *trustworthiness*.
- We provide empirical evidence that simplistic explanations fall short in terms of their benefit for recommendations when compared to human explanations. We suggest that RS should be equipped with more sophisticated means of explaining their decisions. Natural language information exchange employed by humans should be the reference point.
- The contribution is also of theoretical value as we utilize profound statistical tools that allow for causal interpretation of effects. We argue that RS research will benefit substantially from this direction because it opens up an perspective that cannot be achieved with correlative studies.

The remainder of this paper is organized as follows. Section 2 examines relevant literature and puts them in relation. We describe our empirical study and the tools we used in Section 3 and present results in Section 4. Finally, implications of our findings are discussed in Section 5 and summarized in Section 6. The latter also addresses limitations and future work.

2 RELATED WORK

RS have become ubiquitous means that proactively filter information in order to help users find interesting items [38]. Providing recommendations not only helps users make decisions, thus reducing their cognitive load [19, 37], but also increase purchases and general user satisfaction [38]. Nearly all contemporary online platforms, such as Amazon, Netflix, and Facebook, make use of RS [14, 16, 43]. While for a long time research in RS focused primarily on algorithmic accuracy, it recently began to shift onto more user-centered qualities [4, 23, 27, 32] such as the degree of control [20], the transparency [46] and the trustworthiness [47] of a RS.

Trust in Recommender Systems

Trust is an important factor in human-machine interaction [28] and arguably of special interest for RS, since taking an advice is a highly trust-dependent behavior [30, 31]. Not surprisingly, increasing the trustworthiness of a RS has been shown to increase purchase volume [34, 46] and customer loyalty [46], among others.

From a cognitive science perspective, it is a non-trivial task to define what constitutes trust. Consequently, there are various definitions of trust in the literature. In this paper we follow McKnight et al. [30, 31] and their interdisciplinary model of trust. The model comprises four general constructs that are directly or indirectly influencing trust-related behavior: a user's *disposition to trust* together with their *institution-based trust*, *trusting beliefs*, and *trusting intentions*. The *disposition to trust* describes the trusting stance and trustfulness of a person, such as their general faith in humanity. In contrast to the rather constant *disposition to trust*, *institution-based trust* is ephemeral and lasts only for certain situations (e.g. visiting an online shop). *Disposition to trust* and *institution-based trust* together build the foundation for *trusting beliefs*. *Trusting* *beliefs* directly concern characteristics of the trustee, which are threefold in the model of McKnight et al.: *integrity* (the trustee's reliability and honesty), *benevolence* (the trustee's motives such as altruism and goodwill) and *competence* (the trustee's ability to fulfill the truster's needs). Before a person finally commits to a trust-related behavior (e.g. making an online purchase), *trusting intentions* need to be present. *Trusting intentions* itself consist of four subconstructs: *willingness to depend* (the general readiness to make oneself vulnerable to the trustee), *follow advise* (the intention to take an advice of the trustee), *give information* (the willingness to share some private information with the trustee) and *make purchase* (the intention to actually purchase something). Trusting intentions highly depend on disposition to trust, institution-based trust and trusting beliefs. Interestingly, such trust formation processes also seem to apply to computer systems in general [30] and to RS in particular [22].

The source of recommendation, i.e. the trustee, highly influences the acceptance of recommendations. The recommendation source, however, is not per se an automatic RS. In fact, before digitalization, recommendations were primarily provided by other humans—and often still are. The resulting two kinds of recommendation sources (i.e. human and non-human) are often termed as *personal* and *impersonal* [41, 44].

Impersonal sources that provide personalized recommendations are commonly used on contemporary online sites, but allowing other users to provide recommendations can add benefits to a service as well. Although humans have been observed to be less accurate when predicting another user's interests [25], the social cues transmitted by a personal recommendation source create social presence and can foster users' trust in a system [6, 26]. Additionally, depicting simple visual cues for trust-related attributes (e.g. expertise) of a personal recommendation source can influence trusting beliefs successfully [26].

Explaining Recommendations

Another approach to enhance trust in RS is to provide the rationale behind a recommendation in the form of textual explanations [10, 17, 46]. The literature on impact of explanations is controversial, though. While explanations have been shown to have potential for increasing transparency [42, 46], this does not necessarily improve trust in RS [8]. Yet, transparency can help users in their decision making [45] and increase user satisfaction [12]. Overall, effects of explanations seem very diverse and it can be hypothesized that this is due to different types of explanation being utilized.

One of the most common types of explanations is based on similarity between items or users and is fairly simplistic. A well-established approach, for instance, brings the recommended item into relation to those for which the user has already expressed preference. Various methods for explaining recommendations based on the computed similarity between items or users have been proposed [e.g. 2, 17, 46]. Amazon's approach of explaining recommendations based on items that were bought together constitutes another well-known example of similarity-based explanations. The effectiveness of such approaches remains questionable, though. In experiments conducted by Berkovsky et al. similarity-based RS failed to convey trusting beliefs properly [2]. Especially competence and benevolence of a recommendation source appear harder to assess based on similarity only. In line with that, Bilgic and Mooney [3] found that users in conditions with similarity-based explanations tend to overestimate the quality of recommended items, which resulted in a decrease in the perceived trustworthiness of the RS—probably due to a lower perceived competence. Yet, such explanations can result in desirable effects. Berkovsky et al. observed that similarity-based explanations can successfully increase the perceived transparency of recommendations.

However, other forms of explanations can unlock further desirable qualities. For instance, explanations indicating a high average rating of a recommendation resulted in a high perceived benevolence [2]. In the same experiment, competence was rated higher for explanations that used awards and revenue of the recommended items. Qualitative comments underlined this by assigning the latter explanation style with having the most knowledge about the item domain. In another experiment, explanations that made use of content features showed potential to increase general user satisfaction [3]. Finally, first steps have been taken for generating complex explanations based on natural language [5, 9]. Besides increasing the user satisfaction, such explanations also showed potential to be perceived as more trustworthy.

In summary, explanatory complexity spreads a continuum, ranging from rather shallow, similarity-based approaches to complex explanations that leverage natural language. Research so far gives evidence that trustworthiness of a recommendation source increases along this continuum. However, it remains underexplored which attributes of a recommendation source in particular are conveyed through such complex explanations and how. Especially, investigations are missing that shed light on how aspects such as the social presence of a recommendation source, the perceived recommendation quality and the trusting beliefs relate to each other.

3 METHOD

In order to investigate differences of personal and impersonal recommendation sources and their explanation capabilities, we conducted an online study with a between-subject design. Since we were also interested in trust dynamics over time, we conducted the experiment with two measurements over the course of two weeks. The general study setup, the

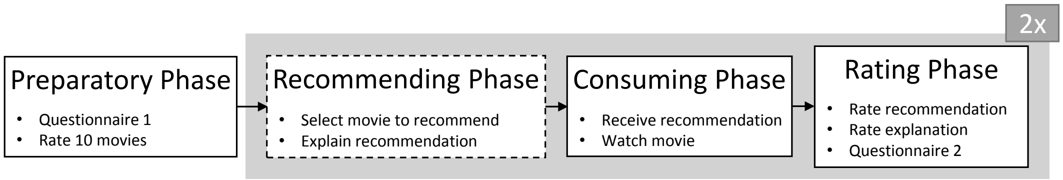

Figure 1: Study phases of the used design. Note that the phase with the dashed line (*recommending phase*) only took place in the personal condition. The phases inside the gray box were performed twice.

two conditions (personal and impersonal recommendation source), the consecutive points of measurements, as well as the tools used are described in detail below.

General Setup

As items to be recommended, we chose movies. The general rationale behind this decision was that we wanted participants to be familiar with the domain. This is a crucial point since participants had to be able to provide recommendations. A second benefit of the movie domain is the abundance of well-established datasets for automatic RS, such as the Movielens 20M rating dataset[1], which we utilized here. Since we wanted participants to be able to watch recommended items, possible recommendation candidates were restricted to those available at Amazon Prime, resulting in 393 recommendation candidates.

For the experiment we recruited 93 participants (55 female) with an average age of ($M = 25.75$, $SD = 9.00$) years. Most participants were students (68 %) or employees (24 %). A requirement for participating in the experiment was that the candidates had an Amazon Prime account so that they could actually consume recommended items. Consequently, participants were used to online streaming providers, using them on a daily (41 %) or weekly (31 %) basis. Participants were randomly assigned to conditions, resulting in sample sizes of $N = 49$ for the personal and $N = 44$ for the impersonal condition.

In a preparation step, all participants—independent of the assigned condition—were asked to rate 10 movies they already knew on a 5-point rating scale in order to elicit preferences. Afterwards, they were asked to follow the scheduled interaction cycle (see Figure 1) that was slightly varied between conditions.

Impersonal Condition

We designed the system inspired by typical online RS: after rating the items (see above), participants immediately received a recommendation. Recommendations were generated using the well-established technique of *Matrix Factorization* [24]. Specifically, we used the *Java* implementation of the *ParallelSGDFactorizer* made available by *Apache Mahout*[2]. In tandem with the recommended item, a similarity-based explanation for the recommendation was presented. Supposing that *Fight Club* was recommended and *Pulp Fiction* was highly rated by the user, the explanation had the following form:

> *Fight Club is recommended to you because it is very similar to Pulp Fiction.*

After receiving recommendation and explanation, participants were asked to watch the movie and subsequently rate movie, recommendation and explanation on a 5-point rating scale. Some days later, a new recommendation and explanation was calculated and presented. Again, participants were asked to watch the recommended movie and rate recommendation, movie and explanation afterwards.

Personal Condition

Overall, the personal condition followed the study design of the impersonal condition with one exception: All participants were assigned a *buddy*[3] and—in order to estimate preferences—were presented with the buddy's 10 rated movies. At the same interface, a searchable list of all 393 recommendation candidates, being available at Amazon Prime, was shown. Out of these candidates, participants should pick one as recommendation and compose an explanation for why they recommended it. This explanation was restricted to 255 characters in order to be comparable to the explanations from the impersonal condition in terms of length.

Instruments

We set up a website in order to deliver automatic recommendations to participants in the impersonal condition and to connect participants in the personal condition to each other. We used the same layout for both to control for confounding stimuli.

[1] https://grouplens.org/datasets/movielens/20m/; the dataset comprises 20 million ratings for 27,000 movies by 138,000 users

[2] https://mahout.apache.org/

[3] In general this assignment was random but we controlled it for avoiding reciprocal relations. Participants received recommendations from a different person as they were providing recommendations to.

Several times over the course of the two weeks (see Figure 1), participants were asked to fill in questionnaires. After the first login into the system and before preference elicitation (i.e. *Preparatory Phase*), participants were asked to complete the first questionnaire on general demographics. Additionally, prior domain knowledge, the frequency of using online streaming providers and general trust in technology Knijnenburg et al. [21] were measured. Furthermore *disposition to trust* and *institution-based trust* [30, 31] were assessed. All items were measured using a 7-point Likert scale.

The second questionnaire was presented after participants had watched the first and second recommended movie respectively (*Rating Phase*). We used items from McKnight et al. [30, 31] to measure *trusting beliefs* and *trusting intentions*. For measuring *social presence* we relied on items from Gefen [13]. All items were assessed on a 7-point Likert scale. In addition, participants were asked to rate recommendations and explanations on a 5-point rating scale. We decided to incorporate post- instead of pre-consumption assessments because we assume participants can more resonably evaluate recommendations and explanations after consuming the item[4] [29].

4 RESULTS

Descriptive results of our study can be found in Table 1. They are split subject according to the experimental *condition* and the *point in time*, i.e. measurement. In order to unravel how *social presence*, *explanation quality*, and *recommendation quality* relate to each other and how they affect trust in the source of recommendation we hypothesized a structural model (see Figure 2) that we will describe in the following.

Structural Equation Modeling

Based on the number of latent constructs and observed variables we estimated the lower-bound for the sample size. With the probability level set to $\alpha = 0.05$ and a desired statistical power level of 0.8, the sample is required to be comprised of 184 observations to, at least, detect medium effects (0.3) [7, 48]. Since measurements of our experiment were taken at two points in time, we had access to 186 observations[5] in total for our analysis and are thus matching the required threshold.

We were interested in identifying whether the interaction led to differences in the assessment of trust subject to our experimental condition, i.e. a personal vs. impersonal recommendation source. *Condition* was defined as an exogenous categorical variable. We hypothesized the recommendation source not only to have an impact on trust towards the source itself (*trusting beliefs*) but also on the willingness to perform trust-related behavior (*trusting intentions*). We further assumed that this effect was mediated by systematic differences between the recommendations provided by the two sources, e.g. the nature of the explanations. Since the interaction stretched across two phases, we additionally considered whether trust would change over time. Just like *condition*, *point in time* was defined as an exogenous dummy variable. Structural equation modeling was applied to trace causal paths that lead to the development of trust or a lack thereof. For this, we utilized the *R* package *lavaan*, version 0.6-2 [39].

	1. Measurement				2. Measurement			
	Imp.		*Per.*		*Imp.*		*Per.*	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Trusting Beliefs	4.64	1.32	4.98	1.01	4.3	1.51	4.92	1.12
-Benevolence	4.45	1.48	5.0	1.18	4.2	1.6	4.88	1.21
-Integrity	4.53	1.53	5.1	1.11	4.27	1.55	5.08	1.15
-Competence	4.94	1.42	4.8	1.45	4.41	1.71	4.81	1.56
Trusting Intentions	4.43	1.29	4.39	1.18	4.12	1.41	4.41	1.2
-Willingness t. D.	4.56	1.44	4.27	1.47	4.14	1.59	4.36	1.6
-Follow Advice	4.75	1.54	4.71	1.48	4.28	1.58	4.65	1.5
-Give Information	3.99	1.43	4.18	1.34	3.95	1.65	4.23	1.36
Social Presence	2.38	1.43	3.8	1.67	2.39	1.5	3.73	1.61
Expl. Quality	3.23	1.25	3.76	1.12	2.84	1.33	3.88	1.18
Rec. Quality	3.91	1.21	3.67	1.18	3.51	1.33	3.92	1.08

Table 1: Mean Values and Standard Deviations for dependent variables. All variables were assessed using a 7-point Likert scale. Only explanation and recommendation quality were elicited on 5-point rating scales.

We conducted missing data analysis, outlier detection, a test for normality, and the selection of an appropriate estimator as preparation steps. Missing columns were observed for two participants. Little's MCAR test turned out to be non-significant ($\chi^2 = 78.01, df = 64, p = 0.11$). Therefore, we can safely assume that the data was missing completely at random and we were allowed to use maximum likelihood parameter estimation. Outlier detection based on Cook's distance revealed three rows to be outliers which were subsequently dropped leaving us with a final sample size of 184. Shapiro's test for normality indicated that several variables of interest significantly deviated from normal distributions. As a result, we conducted the analysis with an estimator that allows for robust standard errors and scaled test statistics. Together with the requirement to handle missing data, we settled with the MLR estimator [11].

Since we found no evidence that *point in time* had an influence on any constructs of interest, we decided to omit

[4] We, nonetheless, tested for possible differences between pre- and post-consumption and did not find any significant differences which is in line with [29] for the movie domain.

[5] Due to combining observations from two points in time we cannot assume mutual independence. Separate structural models for each point in time, however, revealed effects identical to the combined model. Therefore, we assume that the influence of dependence is neglegible.

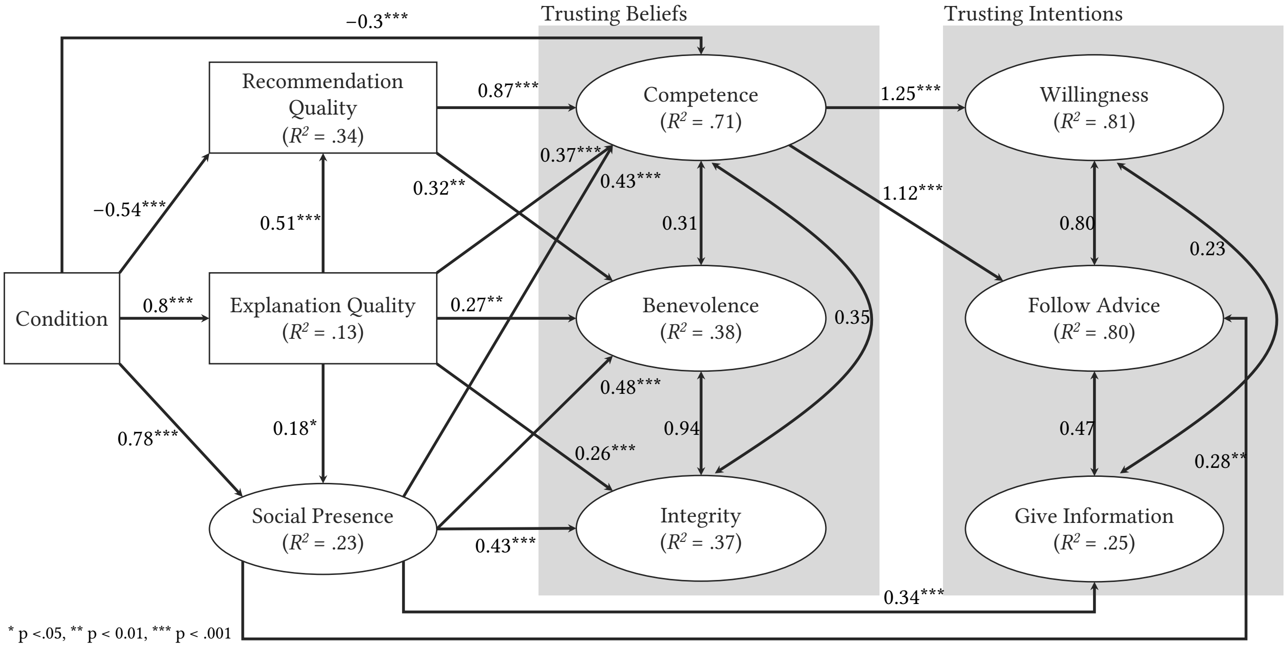
Trusting Beliefs
Trusting Intentions
−0.3***
Recommendation Quality (R^2 = .34)
0.87***
Competence (R^2 = .71)
1.25***
Willingness (R^2 = .81)
0.37***
0.43***
0.32**
−0.54***
0.51***
0.31
1.12***
0.80
0.23
Condition
0.8***
Explanation Quality (R^2 = .13)
0.27**
Benevolence (R^2 = .38)
0.35
Follow Advice (R^2 = .80)
0.48***
0.78***
0.18*
0.94
0.47
0.26***
0.28**
Social Presence (R^2 = .23)
0.43***
Integrity (R^2 = .37)
Give Information (R^2 = .25)
0.34***
* p <.05, ** p < 0.01, *** p < .001

Figure 2: Structural Equation Model comparing the influence of an algorithmic recommendation source with a human. Manifest (observed) variables are depicted as rectangles and latent (unobserved) constructs as ellipses. To prevent overloading the graph, the observed questionnaire items corresponding to latent variables are omitted. The edges show standardized parameter weights and the amount of explained variance for endogenous variables is displayed inside the nodes.

it. The remainder of our hypothesized model appears to be a good fit for the data (CFI = .970, TLI = 0.963, $RMSEA$ = 0.052). For the sake of clarity, we will report significant direct effects successively from left to right. Along these paths we will trace back mediated influences from *condition* on the endogenous variables.

Direct Effects & Mediation via Explanation Quality. The positive direct effect from *condition* onto *explanation quality* (see Figure 2) suggests that the explanations formulated by human buddies attain higher quality than the generic similarity-based ones. *Explanation quality* acts as a mediator between *condition* and *recommendation quality* as well as between *condition* and *social presence.*

While *condition* has a negative direct effect on *recommendation quality*, suggesting that human buddies provide recommendations of lower quality, the mediation [*condition* → *explanation quality* → *recommendation quality*] yields a competing impact of .44 (p < .001). Hence, although recommended movies from a personal source are perceived as worse if examined in isolation, this effect is antagonized by the significant positive influence exhibited by the explanations provided. When put together, both effects cancel each other resulting in a total effect of 0.02 (p = .919).

Concerning *social presence*, the direct as well as the indirect effect [*condition* → *explanation quality* → *social presence*] (standardized coefficient = 0.12, p = 0.03) assume positive polarity. The combined total effect is 0.9 (p < .001) indicating that having a personal recommendation source is related to higher levels of *social presence.*

Direct & Indirect Effects on Trusting Beliefs. We can observe four direct effects on *competence*: The negative impact from *condition* suggests that, per se, the human buddy is perceived as less competent. The remaining three effects from *recommendation quality*, *explanation quality*, and *social presence* are all positive with, according to its parameter weight, *recommendation quality* having the strongest influence.

Both *recommendation quality* and *social presence* thereby become mediators themselves carrying some of the explanatory power of *condition* and *explanation quality.* For instance, the path [*condition* → *social presence* → *competence*] yields an indirect effect of 0.21 (p < .001). Please note that we now also have to consider paths with two mediators such as [*condition* → *explanation quality* → *recommendation quality* → *competence*] with an effect of 0.21 (p = .001). Combining all these effects leads to a non-significant total effect of .11 (p = .475) from *condition* on *competence.*

There exist similar causal patterns for *benevolence* except for the insignificant direct effect from *condition.* As a result, all explanatory power can be distributed to the mediators. The total effect of 0.453 (p = .002) tells us that personal recommenders cause participants to develop higher levels of benevolence via better *explanations* and increased *social presence* despite the negative impact of lower *recommendation quality.*

The strongly positive total effect from *condition* on *integrity* with 0.51 ($p < .001$) suggests that personal recommenders appear more honest and genuine than their automated counterparts. Again, we cannot identify a significant direct influence from *condition* such that all causal effects can be explained by means of mediators. While the indirect paths via *explanation quality* and *social presence* depict similar patterns as the ones discussed for *competence* previously, the effect from *recommendation quality* turned out to be non-significant.

Direct & Indirect Effects on Trusting Intentions. The causal influence on *willingness to depend* can exclusively be reduced to *competence* as it is the only significant predictor. Therefore, it is sufficient to only analyze the paths to that point as the implications are equivalent. By combining the previous non-significant effect from *condition* on *competence* with the direct impact from the latter, we obtain a total effect of 0.26 ($p = 0.161$).

Since *competence* is also an influencing factor for *follow advice*, the same relationships as for *willingness to depend* are of importance again. Additionally to the effects via the route [*condition* → *(mediators)* → *competence*], there is a significant direct influence from *social presence* of 0.39 ($p < .001$) this time. Elevated *social presence* therefore leads to a greater tendency to follow the recommender's advice. Put together, the total effect is 0.41 ($p = .028$).

Give information is completely independent of any paths that tackle recommendations or even the recommendation source. Exclusively by an increased *social presence* is it possible to predict a higher probability of a person sharing information (standardized coefficient = 0.39, $p = .005$).

Counterfactual Analysis

The structural model described in the previous section has already provided some insights into causal effects exhibited by the exogenous exposure variable *condition.* By decomposing its total effect into direct and indirect parts, we have exposed *explanation quality* as the pivotal discriminating factor between personal and impersonal recommendation sources. Due to the generic nature of the explanations generated by the RS, its trustworthiness and *recommendation quality* as well as perceived *social presence* were obviously confined.

On the basis of these findings, we can now hypothesize that RS performance is likely to be substantially improved if better explanations could be provided. The counterfactual mediation framework allows us to investigate questions about such hypothetical situations with outcomes we cannot observe in reality. Specifically, counterfactual analysis lets us express the potential change induced by the *condition* when keeping *explanation quality* fixed at the value that had naturally been observed. In other words, we can estimate the degree to which, for instance, *recommendation quality* would change if the RS was capable of generating explanations of the same quality as humans.

We can achieve this in terms of composite or nested counterfactuals. Let $Y_i(x, M(x))$ be the outcome for individual i when exposed to *condition* x under consideration of the mediator's M influence. For binary exposures, the composite counterfactual is then the outcome for *condition* x subject to the intermediate outcome for the alternate exposure level x^*, i.e. $M(x^*)$. Generalizing to population level is done by taking the expected value which yields the mediation formula [35]:

$$\mathrm{E}\{Y(x, M(x^*))\} = \sum_m \mathrm{E}(Y(x, m))\Pr(m|x^*, C), \tag{1}$$

where C is a set of confounding variables. Since we are interested in the expected improvement over actually observed values for the mediator, $\mathrm{E}\{Y(x, M(x))\}$, we need to calculate the unit effect **UE** of M on Y given X:

$$\mathbf{UE} = \mathrm{E}\{Y(x, M(x^*))\} - \mathrm{E}\{Y(x, M(x))\} \tag{2}$$

We calculated a mediation model with the outcome set to *recommendation quality* in order to emphasize the importance of explanations to support the main goal of RS, i.e. generating good recommendations. Therefore we set $Y = \textit{recommendation quality}, M = \textit{explanations quality}, X = \textit{condition}$. Based on the results of the structural model and further investigations, no confounding variables could be identified. The resulting unit effect is:

$$\mathbf{UE}(\text{recommendation quality}) = 4.11 - 3.63 = 0.48 \tag{3}$$

Altering *condition* from *personal* to *impersonal* while maintaining *explanation quality* therefore increases the expected assessment of *recommendation quality* from 3.63 to 4.11 which corresponds to an improvement of 13 %.

Qualitative Analysis of Explanations

In order to get a better understanding of how participants composed explanations, we provide some examples (see Table 2). Examining those examples more closely shows that explanations vary from very sophisticated statements (e.g. p294) to shallow comments (e.g. p273). Some also use similarities (e.g. p435) or express uncertainty (e.g. p369). Others address general quality of the recommended movie (e.g. p414) or try to be convincing and flattering (e.g. p427). Overall, generated explanations used a similarity relation to the rated movies of the recommendation receiver in 37 %.

Participant	Sample Explanations
p269	"*Based on the rated movies of the buddy I don't know what he likes or dislikes. Hence i chose an entertaining over the top action movie, that is diverting for the short time of the movie.*"
p270	"*I chose the film because I saw it myself and was excited about it. My buddy and I seem to have a similar taste. Besides I wanted to pick a movie in the genre of fantasy/science fiction, based on the rated movies of the buddy.*"
p273	"*Fantasy movie, action*"
p294	"*A classic and atmospheric story, where a noble-minded hero fights an epic battle against the evil (as in most of my buddy's highly rated movies).*"
p307	"*Once is a low budget movie, that has a lot to offer musically. My buddy seems to like films that are emotional and do have melancholic soundtracks. Therefore i chose this nonfamous movie.*"
p369	"*I find it difficult to find a matching movie since the genres of your rated movies are quite different. In addition I do not know most of them. I recommend Disturbia as it mixes action and thriller elements and hope that matches your taste. :)*"
p414	"*A thrilling movie with a tangled plot of hunter and hunted, awesome cast and a whole lot of action.*"
p421	"*I think you don't like romantic comedy or extreme horror movies. As a result I picked this movie. It contains action not too much and a good story that concludes with the movie. Have fun!*"
p427	"*Memonto is very thrilling to watch and contains a whole bunch of light bulb moments. I think this movie is very sophisticated and nothing for bores—thus the perfect movie for guys who like profound stories, like you ;)*"
p435	"*Since my Buddy rated Forrest Gump highly, I guess he/she will like this touching movie with Tom Hanks as well.*"
p445	"*Because it's a good movie*"

Table 2: Some of the explanations created by participants in our experiment (carefully translated to English).

Taking into consideration the language style, 16 % of explanations addressed the buddy directly, 44 % used the third person and 38 % were formulated in a neutral manner. Smileys or other kinds of emoticons were only used scarcely (in 10 % of the explanations). 18 % of the explanations expressed a high certainty regarding the recommended movie, whereas in 4 % of cases it was explicitly stated that the participants were not sure about the recommendation. On average participants used $M = 23.23$ ($SD = 10, 71$) words for their explanations[6].

Explorative Inspection of Temporal Effects

Although the structural model revealed no significant differences for the point in time, we were still interested in explorative investigation. Overall, the reported values in Table 1 are homogeneous within conditions. This is underlined by statistical comparisons: When comparing results between points of measurements, there were no significant differences—neither in assessed quality of recommendations and explanations nor regarding trust in the source of recommendation. Only for the impersonal condition, statistical significant differences are found. Concretely, with ($t(43) = 1.989, p = .053$) *trusting beliefs* were higher at the first point of measurement. This is also true for its subconstruct *competence* ($t(43) = 1, 973$, $p = .055$). Although values for *benevolence* and *integrity* seem to decrease slightly over time, this was not significant. Similar observations can be found regarding *trusting intentions*. Within the impersonal condition, we also found a marginal significant difference here ($t(43) = 1, 984$, $p = .054$). Again, the values at the first point of measurement were slightly higher. This also holds for the subconstruct *follow advice* ($t(43) = 2.126$, $p = .039$). Values of *willingness to depend* were not significant, but seem to slightly decrease, whereas the intention to *give information* nearly remains stable over time.

Summary of Findings

The main focus of the statistical analysis presented was the investigation of causal paths along a structural model (Figure 2) that lead from the effects of our experimental *condition* to trust-related constructs. Our results suggest that the higher-quality explanations provided by participants had an overall positive effect on their buddies' *trusting beliefs* and *trusting intentions*, despite the lower *recommendation quality*. We discuss the implications of our findings in the next section.

5 DISCUSSION

Close inspection of the relations discussed in the previous section hint at a pivotal role of *explanation quality*. *Recommendation quality*, *social presence* and the trusting beliefs *competence*, *benevolence* and *integrity* were all significantly and directly affected by the quality of explanations.

Recommendation and Explanation Quality

By distinguishing between direct and indirect influences, we were able to detect systematic effects that would otherwise have been obstructed. Concretely, no differences could be found descriptively between the two conditions for perceived *recommendation quality* (see Section 4). However, by taking into account the mediating function of explanations, a negative direct effect became evident. That is, if we look at the chosen items in isolation and control for any influence explanations might have, human recommendations were less likely to conform to the receiver's preferences. This finding is in line with previous research[25]: Humans tend to listen

[6] Explanations were restricted to a maximum of 250 characters.

to their "gut feelings" and rely on vague emphatic estimations, whereas the RS, due to its statistical nature, has access to a vast factual basis from which to derive its decision.

However, the parameter weights on the indirect path [*condition → explanation quality → recommendation quality*] cause the total effect to become insignificant. Our deductions are twofold: First, a good explanation can, at least to a certain degree, make up for a poor recommendation. Second, humans compose explanations of significantly superior quality[7].

We originally solicited *movie quality* besides *recommendation quality* but discarded it. While in some cases there surely is a difference (imagine recommending a movie the user already knows and likes: even though the item is liked, the recommendation would not be considered very helpful), it seems that participants in our experiment could not draw a mental line between these concepts. When replacing *recommendation quality* with *movie rating* inside the SEM, effects stay identical. This should not be the case had the movie been rated solely on watching experience. Especially *explanation quality* would not have influenced subjective *movie quality*.

People dislike the explanations generated by the RS because they are, in essence, a verbalization of the similarity relation between a previously rated movie and the recommendation. Without any further context given, they appear arbitrary to users. In our experiment the RS did not disclose its decision criteria, thus making it difficult for participants to understand the foundation for the similarity estimation. Moreover, the system did not explain why a particular rated item was chosen as the basis for an explanation and not any other.

On the other hand, humans conveyed their explanations in an argumentative manner that resembles very closely the process of how people exchange information in reality. Overall, they gave more nuanced explanations by justifying their choice and contextualizing the recommendation with respect to a plurality of dimensions. Interestingly, they often also used similarity to rated movies, revealing that this style is per se suitable for explanation purpose. Yet, humans often combined several explanation styles, e.g. by summarizing content commonalities in rated movies and bring them into similarity-context with the recommendation (e.g. see p307 of Table 2). We thus believe that combinations of different explanation styles lead to explanations with a higher perceived value, which is backed up by prior findings [5].

RS in general should be equipped with more sophisticated means of explaining their decisions. Counterfactual inference give us concrete hints about the effect size we can expect: While maintaining the same algorithmic accuracy and only by adopting to a human-like rather than a similarity-based explanation style, the quality of recommendations would be improved, on average, by around 0.5 points on the rating scale. Expected improvements over RS that do not provide any explanations at all—which are still very common—would be even greater.

Trusting Beliefs and Trusting Intentions

Beyond improving the quality of recommendations, our experiment shows that good explanations can also increase the trust in their authors as expressed by the significant effects on all subconstructs of *trusting beliefs*. It is safe to assume that individuals who can articulate profoundly how they chose a movie as recommendation, e.g. by contextualizing their choice, will be considered competent advisors. Moreover, integrating direct speech and other subtleties of human language into the explanation text may trigger associations of benevolence and integrity. These competences, which humans learn naturally through socialization, are typically not reflected in RS explanations. Lower values in *explanation quality* and therefore trustworthiness are possible consequences. This assumption is underlined, although not with statistical significance, by the fact that we observed diminishing trust over time in the RS that was not traceable for humans. After a high initial trust, which is not uncommon when establishing new relations [31], users were supposedly disappointed by the explanatory capabilities of the system. The resulting asymmetry of information and unfulfilled expectations probably led to the observed decrease in trust. As a consequence of these factors, we suggest developing systems for incorporating explanatory components in a manner that resembles more closely the way in which humans exchange information.

Apart from that, we found some interesting relations regarding the subconstructs of trusting beliefs that we shortly want to discuss: First, there was a direct (negative) effect from *condition* on *competence*. That is, a priori RS are perceived as more competent, likely because of a dispositional attitude people seem to have. Second, although the total effect from *condition* on *benevolence* indicates that humans are assessed as being more benevolent than a machine, it is still surprising that we could not identify a predisposition—expressed by a significant direct effect—in favor of humans that transcends the indirect influences. Third, *recommendation quality* seems to have no effect on *integrity*. This can easily be explained against the background that assessing someone as upright is rarely connected with the perception of how good they are at a particular task.

The prediction of *trusting intentions*—and thus trust-related behavior—on the basis of the degree of trust into a source of recommendations is, in contrast to prior research [e.g.

[7] Please note that the R^2 value in explanation quality is rather low at 13%. We account this to the fact that our binary exposure variable obviously cannot explain variance that occurs within *conditions* but only between.

30], only possible via *competence*. We assume that *benevolence* and *integrity* were not conveyed sufficiently in our experimental setup. Moreover, the movie domain may not create the necessity of such traits in order to follow an advice. Thus, we believe that with more information about the recommendation source available, further communication possibilities, and an item domain in which such traits are more important (e.g. real estate business), benevolence and integrity would become more influential. Interestingly, the effect on *trusting intentions* outgoing from *social presence* does not get completely mediated through *trusting beliefs*. *Social presence* directly influences the tendencies to *follow advice* and to *give information*. Certain undetected social cues seem to have been present during interaction, distinct from the recommendation source, that facilitate such social behavior.

Social presence itself is partially affected by *explanation quality*. We can observe at least a small effect that indicates that better explanations increase social awareness. The major portion of explained variance, however, originates in the differences between the two *conditions*. The knowledge about whether the interlocutor is human or not significantly influences one's perception of being in a social situation. This effect seems also to occur through the restricted information channel determined by the recommendation platform which corresponds to prior research [6, 26]. One important factor for the observed perceived *social presence* is probably also the conceptualization of the user study as a reciprocal act between humans. Users in the human condition were not only receivers of recommendations but also producers. This will likely lead to elevated feelings of social exchange and probably also to situational sympathies.

Finally, there are some limitations to our study to consider. We are aware of possible biases in our sample since a requirement for taking part in our study was to have a streaming account. For this reason, we believe that participants were somewhat technically skilled and probably less picky about recommendations and their explanations. Additionally, only a single item was recommended for each point in time, which is not a typical RS situation. This decision was made because we wanted to isolate the situation from as many stimuli as possible. If more than one recommendation would have been shown, other aspects, such as diversity, may have influenced perceived quality of recommendations. However, since this was the case in both conditions, possible biases (e.g. on perceived *recommendation quality*) can be neglected.

6 CONCLUSIONS AND FUTURE WORK

We have provided a detailed analysis of the causal effects that determine the outcome of trust in personal vs. impersonal recommendation sources. We laid particular focus on exposing systematic effects that can be causally ascribed to the fundamental differences between personally composed and automatically generated explanations. Structural equation modeling offered us the tools to uncover subtle cause-effect relationships. By tracing back indirect influences over elongated paths we could identify the relative impact of *recommendation quality*, *social presence*, and especially *explanation quality* on trust. Thereby, our structural model provides an indication of general mechanisms relevant for generating good recommendations that could not have been derived with correlative studies. Counterfactuals helped us answer questions about hypothetical situations in which RS are able to generate human-like explanations. Unit effect values indicate that being capable of doing so will likely turn out to have a significant impact on the perceived quality of recommendations. The impersonal nature of automated RS can, at least to some degree, be overcome by approaching an explanation style that humans tend to employ in everyday interaction.

On the basis of these results, we conclude that the positive impact of adequate explanations is considerably underestimated and receives too little attention in research and—even more decisively—in industry. If we look at contemporary explanations on online platforms, they are, if anything, a subordinate component, be it in Netflix, Spotify or YouTube. We argue for a more prominent role of explanations in RS—especially due to the mediating effects of explanation quality: While automated RS seem to generate recommendations of superior quality, this benefit is countered by the quality of human explanations to the degree of complete equalization. In other words, the tremendous accuracy of recommending algorithms, emerging from decades of research in that area, remains next to meritless, when RS fail to convey rationales behind their recommendations.

Finally, considering the trend of incorporating more and more natural language into human–computer interaction (e.g. personal voice agents such as *Siri* or Amazon's *Echo*), in future work we will aim at analyzing human-generated explanations in more detail to derive insights into features used and their impact on trust. We also plan to utilize more sophisticated explanations in our experimental setting and intend to take conversational explanation patterns into consideration, enabling RS to answer on specific questions about recommendations.

REFERENCES

[1] James Bennett and Stan Lanning. 2007. The netflix prize. In *Proceedings of KDD Cup and Workshop (KDDCup '07)*, Vol. 2007. San Jose, CA, 3–6.

[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 (IUI '17)*. ACM, New York, NY, 287–300. https://doi.org/10.1145/3025171.3025209

[3] M. Bilgic and Raymond J. Mooney. 2005. Explaining recommendations: satisfaction vs. promotion. In *Proceedings of Beyond Personalization Workshop, IUI*.

[4] André Calero Valdez, Martina Ziefle, and Katrien Verbert. 2016. HCI for Recommender Systems: The Past, the Present and the Future. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, 123–126. https://doi.org/10.1145/2959100.2959158

[5] 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)*. ACM, New York, NY, 175–182. https://doi.org/10.1145/2959100.2959153

[6] Jaewon Choi, Hong Joo Lee, and Yong Cheol Kim. 2009. The influence of social presence on evaluating personalized recommender systems. In *Pacific Asia Conference on Information Systems (PACIS)*.

[7] Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum, Hillsdale, NJ.

[8] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van 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 (2008), 455–496. https://doi.org/10.1007/s11257-008-9051-3

[9] Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. 2018. Explaining Recommendations by Means of User Reviews. In *Proceedings of the 1st Workshop on Explainable Smart Systems (ExSS '18)*. http://ceur-ws.org/Vol-2068/exss8.pdf

[10] A. Felfernig and B. Gula. 2006. An Empirical Study on Consumer Behavior in the Interaction with Knowledge-based Recommender Applications. In *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE '06)*. 159–169. https://doi.org/10.1109/CEC-EEE.2006.14

[11] David A Freedman. 2006. On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician* 60, 4 (2006), 299–302. https://doi.org/10.1198/000313006X152207

[12] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382. https://doi.org/10.1016/j.ijhcs.2013.12.007

[13] David Gefen. 1997. *Building Users' Trust in Freeware Providers and the Effects of This Trust on Users' Perceptions of Usefulness, Ease of Use and Intended Use of Freeware.* Ph.D. Dissertation. Atlanta, GA.

[14] Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Transactions on Management Information Systems* 6, 4 (2015), 13:1–13:19. https://doi.org/10.1145/2843948

[15] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 265–308.

[16] Ido Guy. 2015. Social Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 511–543.

[17] 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)*. ACM, New York, NY, 241–250.

[18] Miguel Angel Hernán. 2004. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health* 58, 4 (2004), 265–271. https://doi.org/10.1136/jech.2002.006361

[19] Anthony Jameson, Martijn C. Willemsen, Alexander Felfernig, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Li Chen. 2015. Human Decision Making and Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 611–648.

[20] Michael Jugovac and Dietmar Jannach. 2017. Interacting with Recommenders–Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 10:1–10:46. https://doi.org/10.1145/3001837

[21] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (Oct. 2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[22] Sherrie Y. X. Komiak and Izak Benbasat. 2006. The Effects of Personalizaion and Familiarity on Trust and Adoption of Recommendation Agents. *MIS Quarterly* 30, 4 (2006), 941–960. https://doi.org/10.2307/25148760

[23] Joseph A. Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 101–123. https://doi.org/10.1007/s11257-011-9112-x

[24] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37. https://doi.org/10.1109/MC.2009.263

[25] Vinod Krishnan, Pradeep Kumar Narayanashetty, Mukesh Nathan, Richard T. Davies, and Joseph A. Konstan. 2008. Who Predicts Better?: Results from an Online Study Comparing Humans and an Online Recommender System. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08)*. ACM, New York, NY, 211–218.

[26] Johannes Kunkel, Tim Donkers, Catalin-Mihai Barbu, and Jürgen Ziegler. 2018. Trust-Related Effects of Expertise and Similarity Cues in Human-Generated Recommendations. In *Companion Proceedings of the 23rd International on Intelligent User Interfaces: 2nd Workshop on Theory-Informed User Modeling for Tailoring and Personalizing Interfaces (HUMANIZE)*. http://ceur-ws.org/Vol-2068/humanize5.pdf

[27] Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. 2017. A 3D Item Space Visualization for Presenting and Manipulating User Preferences in Collaborative Filtering. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion (IUI '17)*. ACM, New York, NY, 3–15. https://doi.org/10.1145/3025171.3025189

[28] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[29] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2018. Impact of Item Consumption on Assessment of Recommendations in User Studies. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, New York, NY, 49âĂŞ53. https://doi.org/10.1145/3240323.3240375

[30] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13, 3 (2002), 334–359. https://doi.org/10.1287/isre.13.3.334.81

[31] D. Harrison McKnight, Larry L. Cummings, and Norman L. Chervany. 1998. Initial Trust Formation in New Organizational Relationships. *Academy of Management Review* 23, 3 (1998), 473–490. https://doi.org/10.5465/amr.1998.926622

[32] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. ACM, New York, NY, 1097–1101. https://doi.org/10.1145/1125451.1125659

[33] Bengt Muthén. 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* 49, 1 (1984), 115–132. https://doi.org/10.1007/BF02294210

[34] Umberto Panniello, Michele Gorgoglione, and Alexander Tuzhilin. 2015. Research Note—In CARSs We Trust: How Context-Aware Recommendations Affect Customers' Trust and Other Business Performance Measures of Recommender Systems. *Information Systems Research* 27, 1 (2015), 182–196. https://doi.org/10.1287/isre.2015.0610

[35] Judea Pearl. 2012. The Mediation Formula: A Guide to the Assessment of Causal Pathways in Nonlinear Models. *Causality: Statistical Perspectives and Applications* (2012), 151–179. https://doi.org/10.1002/9781119945710.ch12

[36] Pearl Pu and Li Chen. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20, 6 (2007), 542–556. https://doi.org/10.1016/j.knosys.2007.04.004

[37] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 317–355. https://doi.org/10.1007/s11257-011-9115-7

[38] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender Systems: Introduction and Challenges. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 1–34.

[39] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. http://www.jstatsoft.org/v48/i02/

[40] Donald B Rubin. 1990. Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference* 25, 3 (1990), 279–292. https://doi.org/10.1016/0378-3758(90)90077-8

[41] Sylvain Senecal and Jacques Nantel. 2004. The influence of online product recommendations on consumers' online choices. *Journal of Retailing* 80, 2 (2004), 159–169. https://doi.org/10.1016/j.jretai.2004.04.001

[42] Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. ACM, New York, NY, 830–831. https://doi.org/10.1145/506443.506619

[43] B. Smith and G. Linden. 2017. Two Decades of Recommender Systems at Amazon.com. *Internet Computing, IEEE* 21, 3 (2017), 12–18. https://doi.org/10.1109/MIC.2017.72

[44] Donnavieve Smith, Satya Menon, and K. Sivakumar. 2005. Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of Interactive Marketing* 19, 3 (2005), 15–37. https://doi.org/10.1002/dir.20041

[45] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4 (2012), 399–439. https://doi.org/10.1007/s11257-011-9117-5

[46] Nava Tintarev and Judith Masthoff. 2015. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 353–382.

[47] Patricia Victor, Martine de Cock, and Chris Cornelis. 2011. Trust and Recommendations. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer US, Boston, MA, 645–675. https://doi.org/10.1007/978-0-387-85820-3_20

[48] J Christopher Westland. 2010. Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications* 9, 6 (2010), 476–487. https://doi.org/10.1016/j.elerap.2010.07.003

[49] Sherrie Xiao and Izak Benbasat. 2003. The Formation of Trust and Distrust in Recommendation Agents in Repeated Interactions: A Process-tracing Analysis. In *Proceedings of the 5th International Conference on Electronic Commerce (ICEC '03)*. ACM, New York, NY, 287–293. https://doi.org/10.1145/948005.948043

[50] Jingjing Zhang and Shawn P. Curley. 2018. Exploring Explanation Effects on Consumers' Trust in Online Recommender Agents. *International Journal of Human–Computer Interaction* 34, 5 (2018), 421–432. https://doi.org/10.1080/10447318.2017.1357904