User Feedback in Controllable and Explainable Social Recommender Systems: a Linguistic Analysis

Chun-Hua Tsai
Penn State University
University Park, PA, USA
ctsai@psu.edu

Peter Brusilovsky
University of Pittsburgh
Pittsburgh, PA, USA
peterb@pitt.edu

ABSTRACT
Controllable and explainable intelligent user interfaces have been used to provide transparent recommendations. Many researchers have explored interfaces that support user control and provide explanations of the recommendation process and models. To extend the works to real-world decision-making scenarios, we need to understand further the users' mental models of the enhanced system components. In this paper, we make a step in this direction by investigating a free form feedback left by users of social recommender systems to specify the *reasons* of selecting prompted social recommendations. With a user study involving 50 subjects (N=50), we present the linguistic changes in using controllable and explainable interfaces for a social information-seeking task. Based on our findings, we discuss design implications for controllable and explainable recommender systems.

Author Keywords
Linguistic Analysis; Mental Model; Social Recommendation; Explanation; User Control; User Experience

CCS Concepts
•Information systems → Social recommendation; Recommender systems; •Human-centered computing → User interface design;

INTRODUCTION
Recommender systems have been widely adopted to many different real-world applications to facilitate the decision-making process, from daily entertainment purposes [28] to life-threatening situations [21]. With the abundance of data and AI-driven techniques, the recommender systems have become more powerful in providing algorithmically accurate predictions of user preference. However, the user information needs are varied in a different time and situation [5]. The "one-size-fit-all" solution is less useful and also not realistic in real-world situations [52]. Moreover, the recommendation models are usually not transparent or understandable to lay users. It has been shown that if the provided recommendations are opaque or are lack of transparency, the users tend to trust the recommendations less [40].

The lack-of-transparency problems have been addressed in the research of *explainable AI (XAI)* [11], which focuses on explaining and justifying the outcomes of AI-driven recommender models [22, 35, 7, 27, 50]. For example, interactive recommender interfaces helped users to understand how their actions can impact and control the system [17], which contributes to system *inspectability* [22] or *transparency* of the recommendation process [49]. The study of [14] proposed *explanations* to help the users to understand the reasoning process of recommendation models. Providing controllable and explainable interfaces have been studied that positively contribute to the user experience, i.e., trust, understandability, and satisfactions [10, 49].

The user experiments, like online, lab-controlled, or field studies, were commonly adopted to evaluate the proposed intelligent user interfaces. The typical approach is to let the users interact with the system and measure their subjective feedback by survey or interview [24], which was effective in collecting the *explicit feedback* from the users or subjects. However, the *rationale*, for example, *why* does a user select the recommendations, play music [18] or purchase the recommended item, was seldom measured and discussed in the user experiments [31]. However, the human decision process is a complex multi-faceted construct that consists of user psychological states [16, 23, 45]. The *rationale* is important user feedback to uncover *user's mental models* of understanding the way the system works and the experience of using the system [1]. It is also a crucial indicator to determine if the systems and user interfaces are "good" enough for the users [29]. It is an under-explored area of understanding how the users can adopt controllable and explainable recommender interfaces in their decision-making process [4].

In this paper, we aim to understand the user feedback in the controllable and explainable social recommender systems. That is, we would like to answer the research question of "*How do controllable and explainable interfaces affect the user feedback in the social decision process?*" We formulated the user feedback by user-generated text, i.e., the *reasons* of selecting prompted social recommendations. We conducted a lab-controlled user study and then applied linguistic analysis to the collected data. A total of 50 subjects (N=50) were interacting with four different social recommender interfaces in a lab-controlled study. We compared the 25 categories across three linguistic dimensions, include *grammar*, *summary language* and *psychological processes*. The comparison allows us to observe the user feedback changes across different interface components. We found the users may have different decision-

IntRS '20 - Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 26, 2020, Virtual Event

making mental models when the recommender interface is controllable or explainable.

Our works contribute to the literature of recommender systems is three-fold. First, we introduce an empirical dataset. It is our attempt to measure the user feedback changes between controllable and explainable interfaces. Second, we present a linguistic approach to measure the user feedback of social decisions. Third, we discuss the design implications for future recommender research based on our findings.

RELATED WORKS

Controllable recommender systems allow users to interact with the system [17], i.e., the users can rank or re-sort the recommendation based on their preference or information need. The *interaction* was usually powered by user interfaces with visualization techniques [13]. Previous works had made many attempts to enhance the system controllability. For example, 1) the controllable meta-recommendation interface that allowed the user to adjust the recommendation models [43]; 2) the slider-based social recommender interface that allowed users to fusion multiple recommendation sources [2] as well as inspired to support user-driven fusion [35] and exploratory search [8]; 3) a cluster interface enabled users to explore conference papers and talks through an overview map [55]; 4) a two-dimensional scatter-lot for a social recommender system that helped the users to find a suitable social connection in two-dimension visualized space [52].

Explainable recommender systems can achieve different *explanatory goals* by single-style or hybrid explanations [44, 26]. Providing explanations had been studied to improve user satisfaction, user perception, and user experience [48, 33, 26]. Previous works had made many attempts in enhancing the system explainability. For example, 1) rule-based personalized explanations for hybrid recommender systems that using visualization or text [26, 27]; 2) feature-based personalized explanations for the product recommendation based on personal characteristics [28]; 3) algorithmic explanations on visual recommender systems [9]; 4) post-hoc explanation for a social recommender that adopts visualizations [53].

The user-centric evaluation framework was adopted for the controllable or explainable recommender systems in explaining the user experience [24]. The framework contains both explicit (questionnaire, rating, like, etc.) and implicit (click, time, system log, etc.) user feedback. It was common to see the experiment adopted the established assessment tools, e.g., NASA-TLX [12] and ResQue [40], to measure the user subjective feedback, e.g., satisfaction, trust and workload. However, the *post-experiment* survey may not fully reflect the psychological states while interacting with the system. A few implicit user feedback was adopted in previous studies. For example, [25, 52, 50] proposed the behavior interaction factors that measured the variables of *number of clicks*, *use time*, *follow/like*, etc. The behavior factors and variables were very interface-oriented that measured the interaction between the interfaces and users, which may not fully reflect the users' psychological states, e.g., emotion or feeling.

Measuring the users' mental model in recommender systems have been discussed in previous works, e.g., using recommender interfaces to change the user saving energy behavior [23, 45]. [29] argued that explainable AI, such as explainable recommender interface, should further consider the multi-discipline knowledge, not just from the researcher's intuition [29]. The user experience is a complex multifaceted contact that required further understanding of the user's mental model [1]. A more complex user experience (e.g., emotions and cognitive processes) can be measured by users' writing text [37]. The word choice in writing has been shown related to the writer's psychological states, e.g., personality, emotion, and social fluctuations [37]. For controllable and explainable recommender systems, the recent work of [42] argued that user-centered design should consider the user's mental models. [31] explored the mental model in a transparent and controllable recommender system through a qualitative study, the findings uncovered diverse mental models of perceiving the recommender system. However, the user's mental model in social decision-making is still under-explored. In this paper, we plan to explore the *reasons* of selecting the social recommendations, measured by user feedback in a free-form text.

APPROACH

Our goal in this paper is to apply linguistic analysis to free-form user feedback of social recommender systems, which specify the reasons for selecting prompted social recommendations. The free-form user feedback was a couple of sentences that collected through a text input box every-time the user made the selection. We used the user-generated text as a window to the user's mental model that the user build while interacting with the system.

Linguistic Analysis

The word choice in writing has been shown related to the writer's psychological states, e.g., personality, emotion, and social fluctuations [37]. Linguistic Inquiry and Word Count (LIWC) is a linguistic analysis tool that categorized words in psychological meaningful categories [46], which helps to detect the sentiment and psychological process using users' writing text. Table 1 presented the dimensions that adopted in this paper, including *Summary Language Variables*, *Grammar Variables* and *Psychological Processes Variables* dimensions. The *Grammar Variables* dimension included seven categories to present the common linguistic variables: the percentage of "word counts", "words > 6 letters", "verbs", "adjective", "comparisons", "interrogatives", "numbers" and "quantifiers". These categories depict the basic structure of the writing as well as indicators that we can compare the difference between articles or sentences. The "words > 6 letters" (big words) category indicates the percentage of the words with more than six letters. A high score on this category means using words is more complicated and usually less emotional [39].

The *Summary Language Variables* dimension included four categories: 1) "Analytic": this category captures the degree of writing text that "*suggest formal, logical, and hierarchical thinking patterns*" [38]. A high score on "Analytic" category means the writing text is more narrative based on personal

Table 1: Summary of Linguistic Inquiry and Word Count (LIWC) linguistic dimensions and example vocabulary.

Category	**Abbrev**	**Examples**	**Category**	**Abbrev**	**Examples**
Grammar Variables			*Psychological Processes Variables*		
Word Counts	*WC*	-	Affective processes	*affect*	happy, cried
Words > 6 letters	sixltr	advance, combination	Positive emotion	posemo	love, nice, sweet
Common verbs	*verb*	eat, come, carry	Negative emotion	*negemo*	hurt, ugly, nasty
Common adjectives	*adj*	free, happy, long	Social processes	*social*	mate, talk, they
Comparisons	*compare*	greater, best, after	Female references	*female*	girl, her, mom
Interrogatives	*interrog*	how, when, what	Male references	*male*	boy, his, dad
Numbers	*number*	second, thousand	Perceptual Processes	*percept*	look, heard, feeling
Quantifiers	quant	few, many, much	See	see	view, saw, seen
Summary Language Variables			Past focus	*focuspast*	ago, did, talked
Analytical thinking	*analytic*	-	Present focus	*focuspresent*	today, is, now
Clout	*clout*	-	Future focus	*focusfuture*	may, will, soon
Authentic	*authentic*	-	Motion	*motion*	arrive, car, go
Emotional tone	*tone*	-	Time	*time*	end, until, season

experience [36]. 2) “Clout”: this category captures the degree of social status, which indicates “*the confidence, or leadership that people display through their writing*” [38]. A high score on “Clout” implies the writer is in a higher social status or the role in control [20]. 3) “Authenticity”: this category captures the degree of revealing in a honest way that are more “*personal, humble, and vulnerable*” [38]. A high score on “Authenticity” implies the writing is reflecting the real thought of the writer [30]. 4) “Emotional tone”: this category captures the degree of emotions. A high score (more than 50) on “Emotional tone” supports the writing is carrying positive emotion [6].

The *Psychological Processes Variables* is the most notable linguistic dimension provided by the LIWC program, which provides more than 50 different categories. In this paper, we filtered a total of 12 categories in five groups that are most relevant to this analysis. 1) “Affective processes”: it refers to the affective experience of feeling, emotion, or mood. A higher “Affective processes” score means the writings contain more emotion words. In particular, the “Positive emotion” and “Negative emotion” categories were chosen to distinguish the effects of positive and negative sentiment terms. 2) “Social process”: refers to the personal social experience of interact, adjust, and establish relationships. A higher “Social process” score means the writing contains more social terms. We are particularly interested in the gender difference in our analysis. The categories of “Female references” and “Male reference” are also included. 3) “Perceptual processes”: it refers to the human perceptual experience with the environment, e.g., see, hear, or feel. A higher “Perceptual processes” score means there are more terms related to perceptions. We have highlighted the “See” category in this analysis. 4) “Time orientations”: it refers to how the time perception, e.g., “Past focus”, “Present focus” and “Future focus”, i.e., the duration of the events. 5) “Relativity”: “Motion” and “Time” categories were included in this analysis, to capture the percentage of motion and time related terms in the submitted statements.

Experimental Platform

In this paper, we explored the user feedback model of using an experimental platform: *Relevance Tuner+*, a controllable and explainable social recommender user interface [53] of the *Conference Navigator* (CN). CN is a conference support system [3], which has been used at more than 30+ research conferences. The *Relevance Tuner+* was adapted to provide social recommendations to the conference attendees. Figure 1 presents the interface of the *Relevance Tuner+*. The *relevance sliders* (section A) provide user controllability to *tune* (re-rank) the order of social recommendations based on the user assigned weightings. The users can *tune* the relevance sliders based on their information needs or personal preference. Section D provided the scholars’ profile data; the users can inspect the publication list by clicking the name link. The colored stackable bar visualization (section B) shows how the total relevance score is calculated. The explainability was enabled by the *explanation icon* (section C).

Each *relevance sliders* (section A) controls the importance (weighting) of one of the four recommendation models of the hybrid recommender engine. 1) *Publication Similarity:* this similarity was determined by the degree of cosine similarity between two scholars’ publications; 2) *Topic Similarity:* this similarity was determined by matching research interests using topic modeling (LDA) approach [56]; 3) *Co-Authorship Similarity:* this similarity approximated the co-authorship network distance between the user and recommended scholars; 4) *Interest Similarity:* this similarity was determined by the number of co-bookmarked conference papers and co-connected authors in the conference support social system Conference Navigator (CN3).

The *explanation icon* (section C) opens an explanation window so the users can inspect the rationale behind each recommendation model. The examples of explaining the four recommendation models are presented in Figure 2. The publication similarity was explained by a *Two-Way Bar Chart*, the text-level similarity between the publication of the user and the attendee. The topic similarity was explained by *Topical Radar*, showing the research topics in a radar chart and the topical words of each research topic in the table. The co-authorship similarity was explained by a *strength graph*, which shown the co-authorship network in a path graph. The CN3 interest

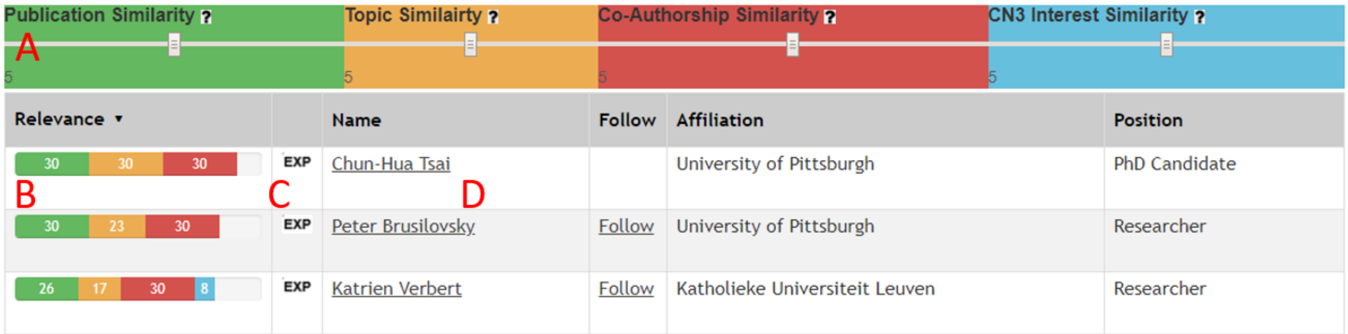

Figure 1: The *FULL* design of Relevance Tuner+: (A) relevance sliders; (B) stackable score bar; (C) explanation icon; (D) user profiles. The sections were controlled in different condition: controllable only (*CONT*) was disabled the section C; explainable only (*EXPL*) was disabled the section A; the baseline interface was disabled both section A and C.

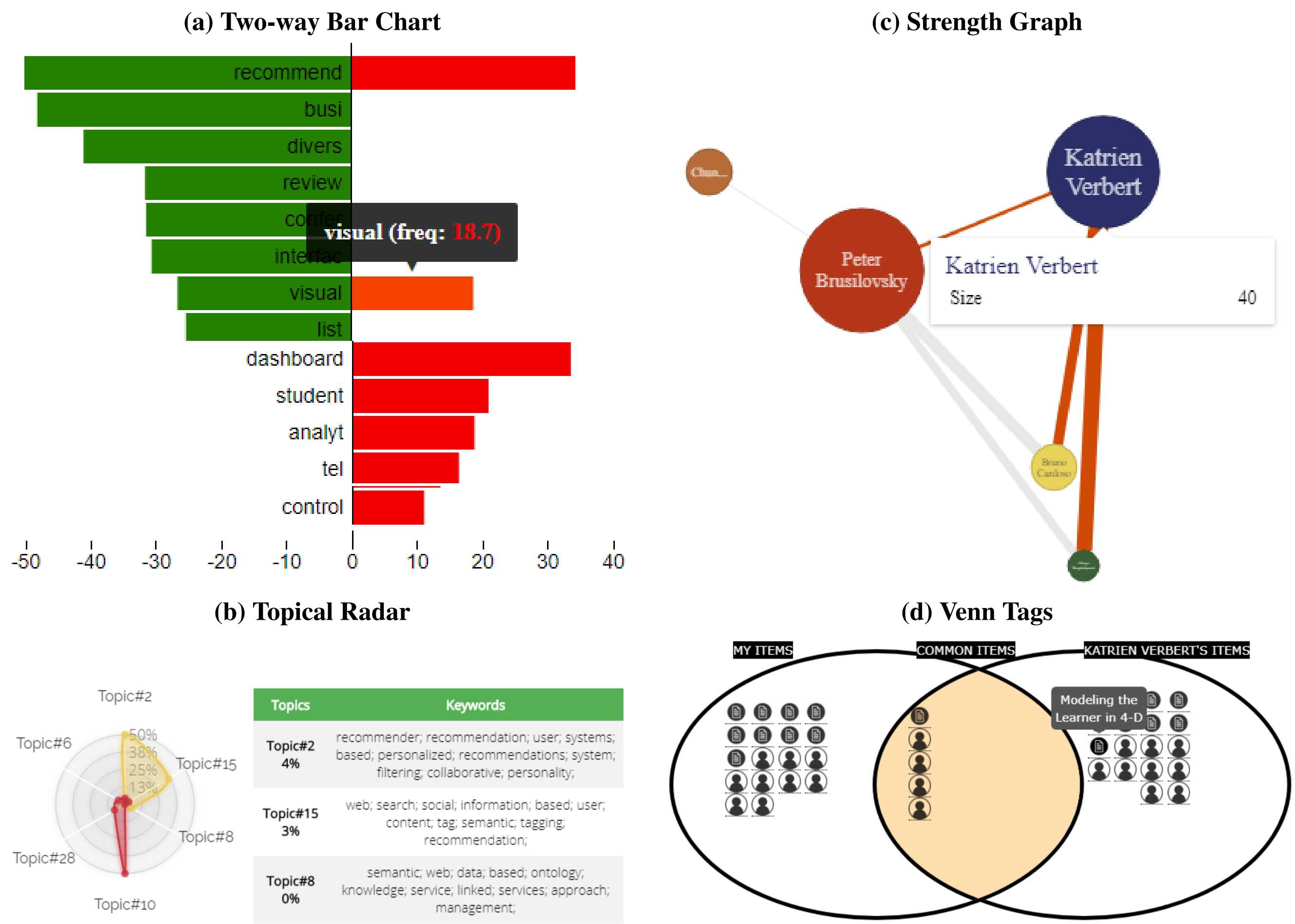

Figure 2: The visualisations used to explain the four recommendation models: (a) *Two-Way Bar Chart* for explaining Publication Similarity. (b) *Topical Radar* for explaining Topic Similarity, (c) *Strength Graph* for explaining Co-Authorship Similarity, and (d) *Venn Tags* for explaining CN3 Interest Similarity in the explanation page. The users can access these interfaces by clicking the *Explanation Icon* (Figure 1, Section C).

similarity was explained by *Venn Tags* interface, to present the bookmarked item in the Venn diagram. The effectiveness of the explanation interfaces was evaluated by previous studies. The detailed designs and evaluations of each component can be found in [54, 53].

Experimental Procedure

The controlled lab user study was conducted in May/June 2019 at the campus of the University of Pittsburgh with a group of 50 graduate students. The study followed a within-subject design for testing four social recommender interfaces using *Relevance Tuner+*: *baseline (BASE), controllable (CONT), explainable (EXPL) and controllable+explainable (FULL)* conditions. Section B and D was enabled in all conditions, but different rules applied to section A and C. The *FULL* interface (shown in Figure 1) had both section A and C enabled. Section A or C was enabled in *CONT* and *EXPL* interfaces, respectively (i.e., has only one section enabled in the interface). Both sections A and C were disabled in the *BASE* interface. To minimize the learning effect and bias, we followed a Latin square design to balance the conditions appeared to each participant.

In the study, the subjects were told to act as a researcher who is attending the conference. The experiment subjects were asked to select suitable candidates to meet at an academic conference based on the scenario shown below (the same scenario was used in each interface), based on their best judgment. Participants were given one training session and one information search tasks for each interface. In the training session, we urged the user study participants to follow a few steps, so they have a chance to familiarize the system. The subjects were then to be asked to complete an information search task by a scenario of finding advisors or mentors for their graduate school admission.

Scenario of *Finding Advisor/Mentor*:

1. If you plan to pursue a doctoral degree after your current degree program, it is an excellent opportunity to find your prospective advisor or mentor at the conference. For this task, you will select scholars to follow as potential advisors/mentors. Please follow *four scholars* whose work is more relevant to your research interest(s). The ideal candidates will be scholars who the system identified as 'more' connected to your chosen SCI professors, so they can provide you a strong recommendation letter;
2. Please "follow" four scholars whose work in more relevant to your research interest(s). The ideal candidates will be scholars who the system identified as more connected to your chosen SCI professors (so they can provide you a strong recommendation letter, etc.
3. You are also expected to justify your selections (for example, to the Ph.D. admission committee), so it is important to pay attention to why do you make the selection.

It is a typical similarity-based scenario that prompts the participants to select scholars (in an academic conference) who have similar research interests as well as have a close connection with the subject's social (co-authorship) network. In the design of *Relevance Tuner+*, the users can tune the *publication similarity* and *co-authorship similarity* sliders for re-ordering the recommendation to better-fit their information needs. For each selected scholar, the subjects were asked to *justify* their selections using a free-form text. The user input text was collected by a pop-up box when the user clicked on "follow" button (shown in Figure 1, section D). It provides an insight to observe the implicit user feedback of social decision-making through writing text, which implied the users' psychological states while interacting with a user-controllable and explainable social recommender interface [37].

Table 2: Metadata of the UMAP conference data

	UMAP 2015	UMAP 2016	UMAP 2017	UMAP 2018
Number of Papers	143	129	168	108
Number of Authors	231	305	345	289
Number of Attendees	116	115	151	131
Number of Bookmarks	664	660	714	342
Assigned Interface	BASE	CONT	EXPL	FULL

To control the data sparsity, the participants were asked to fill-up a pre-study questionnaire for user preference elicitation. The questionnaire required the participants to pick 10 (out of 100) research keywords (As the participant's research topics, e.g., data mining, HCI, etc.), 5 professor (from 18 professor at the University of Pittsburgh whose works related to UMAP conferences, as the participant's co-author) and 12 preferred UMAP papers that were bookmarked in the CN system (as the publication text and interest), so we can generate personalized social recommendations using the four recommendation models. The conference data (papers, authors and conference attendees, bookmarks, etc.) was gathered from the UMAP conference proceeding from the year 2015 to 2018, the meta-data can be found in Table 2. The conference data can be accessed through the Conference Navigator (CN) system[1].

Data Description

A total of 50 participants (N=50) were recruited for the lab-controlled user study. There were 28 males and 22 females whose ages ranged from 22 to 44 (M=28.82, SD=4.83). A total of 22 masters' students joined the study, including 21 IS and 1 MST majors. There were 28 doctoral students, including 18 IS, 3 LIS, 2 CS, and 5 ISP majors. All doctoral students had at least one publication and one conference attending experience, but no Masters' students had any publication or any conference attendance. Subjects took between 52 and 192 minutes (M=106.05, SD=28.80) to complete the study. Each subject was asked to select at least four scholars in each interface, based on the Scenario of *Finding Advisor/Mentor*.

The dataset contains 829 user choice statements (included 29 extra). The word count of each statement ranges from 1 to 51 words (M=10.23, SD=7.44). One example statement of the participant who using *BASE* interface is shown below.

> "*We have the highest publication similari[ti]es. His research is on cognitive science which is also my interest.*" (Subject 3; Female Master Student).

The subjects were expected (but not required) to use all interface components to help them make the selections. In the

[1] http://halley.exp.sis.pitt.edu/cn3/portalindex.php

statement above, the decision was made by the highest publication similarity score, which the subjects can explore from the colored stackable bar (Figure 1, section B). However, the research interest information can only be determined by inspecting the publication list (Figure 1, section D). Another statement submitted by the participant who of using the *CONT* interface, the subject's response indicates the usage of both countable components, shown below.

> "*Second highest relevance when co-auth[o]rship similarity is set to 10*" (Subject 36; Male Master Student)

The participant adopted the slider component (Figure 1, section A) and made his decision based on the ranking of relevance score. A further example from the participant who was using the *EXPL* interface may show the adaption of the explanation component.

> "*His work in social media text analysis and recomm[e]ndation system attracts me. And he has a closed con[n]ection with my advisors.*" (Subject 22; Male PhD Student)

The co-authorship information ("*close connection with advisor*") was accessed from the *explanation* of the co-authorship similarity, i.e., to inspect the visualization (Figure 2c) after clicking the explanation icon (Figure 1, section C). The user-generated data provides us an opportunity to reveal the inner psychological states changes while using different controllable or explainable interface components.

To compute the linguistic changes between interfaces, we first group the submitted statements by the participants and interfaces. We combined multiple statements into a text string, so each interface has an equal number of combined statements (a total of 50). The statements were analyzed for measuring the change of users' psychological states of using the controllable and explainable social recommender interfaces. We ran the normality test and found our data were not normally distributed. Hence, we applied a nonparametric paired Wilcoxon signed-rank test [34] to assess the population mean rank difference. The pairwise comparison of 4 recommender interfaces raises the need to control for the Type I errors (i.e., false positives). We applied the false discovery rate procedure by the Benjamini and Hochberg (BH) method to adjust the P values in our analysis [15].

RESULTS

The Grammar Variables

We reported the linguistic analysis results of the grammar variables in Table 3. We observed the participants submitted less words in the *EXPL* interface. The Wilcoxon signed-rank test supported the word counts of *EXPL* interface was significantly less than the *CONT* ($V = 835.5, P\_Adj = 0.10$) [2] and *FULL* ($V = 431.5, P\_Adj = 0.10$) interfaces. The percentage of using more than 6 letters words (*sixltr*, the "big words") was lower in the *CONT* interface, compared to the *BASE* with the most big words, e.g., "similarity", "authorship", "interest", and "publication"[3]. The Wilcoxon signed-rank test indicated the number of big words in the *CONT* interface was significantly lower than the *BASE* ($V = 807, P\_Adj = 0.06$) and *EXPL* ($V = 388, P\_Adj = 0.07$) interfaces. We did not find significance in the categories of *verb*, *adj*, and *compare*, however, the usage of adjectives and comparison words were higher in the *EXPL* and *FULL* interfaces. The adjective words (*adj*) in our data included: "high", "similar", "good", and "interesting". The comparison words (*compare*) in our data included: "similar", "highest", "more", and "very".

It was interesting to see there were more interrogative words in the *CONT* interface that outperformed the other three interfaces, especially the *EXPL* interface which has the least interrogative words. The interrogative words (*interorg*) in our data included: "which", "when", "how", and "what". The Wilcoxon signed-rank test indicated the percentage of interrogative words in the *BASE* ($V = 58, P\_Adj = 0.008$) and *CONT* ($V = 132, P\_Adj = 0.006$) interface were both significantly higher than the *EXPL* interfaces. We did not find significance in the category of *number*, but we observed the difference in the *quantifier* category. The common *quantifier* words included: "very", "many", "few", and "much". The Wilcoxon signed-rank test indicated the percentage of *quantifier* words in the *BASE* interface was significantly lower than the *FULL* ($V = 124, P\_Adj = 0.031$) interface.

The Summary Language Variables

We reported the linguistic analysis results of the summary language variables in Table 4. We observed the score of *Analytic* dropped in the *CONT* interface versus the other three interfaces. The score of *Analytic* variable of *CONT* interface was lower than the *FULL* interface, but we did not find significance after the BH adjustment. The analysis result implied the participants used less logical and hierarchical word structure in their statements when interacting with a controllable interface. A low *Analytic* score (=9.72) statement example, using *CONT* interface, was shown as below.

> "*His topic is social community; this topic is social community; his topic is about city and urban area; her topic is social network.*" (Subject 13; Male PhD student)

As we can see, *Subject 13* only pointed out the topic relevance and used less logical and hierarchical words to explain or justify in his statement. We can compare it to a high *Analytic* score (=99) statement example, using *FULL* interface, was shown as below.

> "*He has the highest score with the highest CN3 interest similarity score. Good network shown in the Co-authorship similarity. The most relevant publications to my interest among all. The second-highest CN3 interest similarity.*" (Subject 8; Male PhD student)

We can see the *Subject 8* adopted the interface components and provide a thoughtful reason to justify his decision. The sentences adopted the information from the *relevance score*,

[2] "P_Adj" (Adjusted P-Value): adjusting the false discovery rate using the Benjamini and Hochberg (BH) method.

[3] The sample words was ordered by the term frequency in our dataset, applied to all examples below.

Table 3: Linguistic Analysis of the Grammar Variables

Category	Interface				Significance Test					
	BASE	CONT	EXPL	FULL	BASE &	BASE &	BASE &	CONT &	CONT &	EXPL &
	M(SD)	M(SD)	M(SD)	M(SD)	CONT	EXPL	FULL	EXPL	FULL	FULL
WC	43.52 (26.14)	44.62 (23.14)	38.48 (20.11)	45.62 (31.38)		-		-		-
sixltr	35.80 (13.86)	32.95 (15.56)	35.36 (14.16)	34.34 (14.95)	-			-		
verb	10.17 (6.03)	10.98 (5.95)	10.43 (6.31)	9.27 (5.87)						
adj	14.84 (13.74)	13.89 (11.35)	15.40 (12.11)	15.15 (11.17)						
compare	7.92 (7.61)	7.94 (7.11)	9.30 (7.66)	8.94 (6.79)						
interrog	0.44 (1.00)	0.99 (1.92)	0.11 (0.46)	0.58 (1.70)		-		**		
number	0.76 (1.65)	0.99 (2.27)	0.75 (2.31)	0.86 (1.69)						
quant	1.35 (2.23)	1.94 (2.82)	1.81 (2.87)	2.96 (3.75)			*			

*Significance Level: (**) p<0.01; (*) p<0.05; (-) p<0.1

Table 4: Linguistic Analysis of Summary Language Variables

Category	Interface				Significance Test					
	BASE	CONT	EXPL	FULL	BASE &	BASE &	BASE &	CONT &	CONT &	EXPL &
	M(SD)	M(SD)	M(SD)	M(SD)	CONT	EXPL	FULL	EXPL	FULL	FULL
Analytic	70.44 (27.89)	65.13 25.20)	71.26 (26.95)	71.69 (25.38)						
Clout	61.90 (67.04)	67.04 (23.49)	60.13 (26.98)	64.67 (24.70)						
Authentic	26.20 (29.22)	27.15 (30.23)	24.07 (27.56)	23.98 (26.02)						
Tone	79.91 (28.95)	78.42 (29.86)	79.24 (30.52)	85.48 (26.17)						

ranking, *explanation of co-authorship similarity* and *publication similarity*. It is interesting to mention the score was even lower (although not significantly) than the *BASE* interface, which implied the controllable interface might facilitate a quick but less thoughtful decision process. A high *Analytic* score (=99) text example, using *BASE* interface, was shown as below.

> "*because of similar interest publications like facial recognition following; because of interest in recommender system and connections following; be[ca]use of publications in domain of visualiz[a]tion and computer vision; because of interest in data mining*" (Subject 34; Male Master Student)

The statement from *Subject 34* showed the participant could leverage different components for making the decision. In the *BASE* interface, the participant had no supports on controllable and explainable components. The subject needed to inspect the candidate's publication list and made the selection based on the shared research interests. It is doable, but it was not surprising to find out the subject needs extra effort (more time and clicks) in searching and organizing this information.

We observed the score of *Clout* in the *CONT* interface was the highest versus the other three interfaces. The *Clout* variable score of *CONT* interface was higher than the *EXPL* interface, but we did not find significance after the BH adjustment. The result implied the controllable interface grants the "power" to the users that letting them feel in control, which may be less support when only the explanations were provided. A high *Clout* score (=99) text example, using *CONT* interface, was shown as below.

> "*She has a highest score of adding four scores up. She is the most similar scholars with me. We have high publication similarity and co authorship similar[i]ty. We can share opinions of our publications. We have a high topic similarity. We care about same topics, maybe we can col[l]ab[o]rate in the future. We have a high CN3 interest similarity and co authorship similarity. We have similar interest.*" (Subject 45; Female Master student)

We can find the confidence and assurance in the submitted text, from the statement of *Subject 45*. The subject used more *we have* and *we can* in the statement, which indicates an equal social status to the scholars she has chosen, even when she was a Master student with limited research experience.

We did not find a significant difference in the score of *Authentic*. In general, the scores are lower than 50, which represents the degree of the writing, reflecting the subjects' real thought was small. The issue was due to the *Authentic* category was trained by the text, which reflects more "*personal, humble, and vulnerable*", which may not apply to the experimental dataset. In the lab-controlled user study, the subjects were asked to provide feedback based on the selecting decision rather than the personal life experiences. It is not surprising that the scores are below the average. An example of high *Authentic* score (=92), using *CONT* interface, was provided below.

> "*related work in viz and interactive systems interested in visual decision making no publications listed, but I am familiar with [Scholar J] interests health related interests in UMAP.*" (Subject 2; Male PhD student)

The text of *Subject 2* mentioned the personal research interests and personal connection with a scholar listed in the social recommendation list. It is evident that the text revealing the real thoughts.

We did not find a significant difference in the score of *Emotion Tone*, but we observed all submitted text were more align with positive emotions. The average score was 79 to 85, which was higher the cut-off line 50, in between the positive and negative emotions texts. The *FULL* interface had the highest score that showed high positive feedback in the submissions. An example of high *Emotion Tone* score (=99), using *FULL* interface, was provided below.

> "*Big figure in recommendation domain. Have many good papers Using deep learning models on the recommendation. many good papers Very interesting work in video engagement prediction Interesting MOOC research, can be combined with concept/knowledge learning.*" (Subject 25; Male PhD student)

Many positive words were mentioned in the statement submitted by *Subject 25*, for example, "good paper" and "interesting work", which indicates a more positive emotion tone.

The Psychological Processes Variables

We reported the linguistic analysis results of the psychological processes variables in Table 5. We did not find a significant difference in the categories of *affect*, *posemo* and *negemo*. However, we found the *FULL* interface has more affective and positive emotion words, i.e., feeling and emotions words, the common words in our dataset included: "strong", "interest", "interesting", and "nice". We observed very few negative emotion words in our experimental dataset, and only a few subjects used the word "low" in their statement, which usually indicated the lower bar in the relevance slider. The result is not surprising since we asked the participants to select a set of academic advisors, which is not commonplace for receiving negative user feedback.

We observed the *CONT* interface has more social words (*social*) than the other interfaces. The common social word in our dataset included: social, related, they, their, etc. The Wilcoxon signed-rank test indicated the percentage of social word in *CONT* interface was significantly higher than the *BASE* ($V = 225.5, P\_Adj = 0.05$) and *EXPL* ($V = 559, P\_Adj = 0.05$) interfaces. We did not find a significant difference in the *male* category, however, we observed the *female* category was dynamic across interfaces. Surprisingly, we found the *female* related words (e.g., "she" and "her") was used more in non-explainable interfaces, i.e., the *BASE* and *CONT* interfaces. The Wilcoxon signed-rank test indicated the percentage of female word in *BASE* interface was significantly more than *FULL* ($V = 106, P = 0.019$) interfaces. The same pattern repeated in *CONT* interface that outperformed both *EXPL* ($V = 120, P = 0.019$) and *FULL* ($V = 134.5, P = 0.019$) interfaces. The *female* words were mentioned less in the statements when the extra explanations were given. Interestingly, the usage *female* words was maintained across the interfaces. The result implied when no extra explanations were provided, which the recommendations were not transparency, the gender information may be more appearing.

In the categories of *percept* and *see*, we found the *BASE* interface was outperformed the other three interfaces, although no significant difference was found in the pairwise testing. The common *percept* and *see* words in our data included: "look", "see", and "eye". The score of *percept* words in *BASE* interface was higher than the *CONT* and *EXPL* interfaces. The same pattern repeated in *see* category, the Wilcoxon signed-rank test indicated the percentage of *see* words in *BASE* interface was higher than *EXPL* ($V = 113, P\_Adj = 0.006$) and *FULL* ($V = 111, P\_Adj = 0.024$) interfaces. The result supports the limited baseline interface, the subjects tend to use more perceptual words in their statements, e.g., *"his topics (papers) look like interesting"*.

In the categories of *focuspast*, *focuspresent* and *focusfuture*, we found the "present tense" was adopted the most in the submitted text. The score of *focuspresent* words in *CONT* interface was higher than the *BASE* interface. The result implied the interactive (controllable) user interface gave the users real-time response, which may let them had more sense of current time. The interaction also make the participants had more *motion* words in their statements, e.g., "mine", "tune", "walk", and "run". The *CONT* interface has more *motion* related words than the *EXPL* and *FULL* interfaces. We also observed the users mentioned more *time* words in *CONT* interface, compare to the *BASE* interface, It is interesting to observe the *motion* words were only used more when the controllability, but not explainability was enabled. The effect did not persist with the extra explanation involved.

DISCUSSION

Controllable Recommender Interfaces

The goal of a user-controllable recommender interface is to put the user in control, so the users can influence or further filter

Table 5: Linguistic Analysis of Psychological Processes Variables

	Interface				Significance Test					
Category	BASE	CONT	EXPL	FULL	BASE &	BASE &	BASE &	CONT &	CONT &	EXPL &
	M(SD)	M(SD)	M(SD)	M(SD)	CONT	EXPL	FULL	EXPL	FULL	FULL
affect	8.02 (8.56)	6.88 (6.06)	8.15 (9.26)	8.53 (7.83)						
posemo	7.98 (8.59)	6.76 (6.04)	8.03 (9.30)	8.24 (7.85)						
negemo	0.02 (0.16)	0.11 (0.46)	0.10 (0.48)	0.09 (0.44)						
social	7.48 (6.85)	9.57 (9.00)	7.08 (6.48)	8.92 (9.60)	*			*		
female	1.04 (2.22)	1.11 (2.25)	0.48 (1.39)	0.25 (0.77)		-	**	*	**	
male	1.57 (2.99)	1.97 (3.14)	1.69 (2.63)	1.55 (2.45)						
percept	1.20 (2.51)	0.48 (1.17)	0.46 (1.49)	0.51 (1.87)						
see	1.11 (2.52)	0.34 (0.89)	0.30 (1.01)	0.08 (0.42)		-	*			
focuspast	0.67 (1.77)	1.10 (2.39)	1.12 (2.61)	0.64 (1.70)						
focuspresent	8.63 (6.11)	10.12 (6.78)	8.84 (5.86)	8.54 (5.16)						
focusfuture	0.44 (1.32)	0.37 (0.99)	0.24 (0.89)	0.68 (1.49)						
motion	0.44 (1.50)	0.52 (1.14)	0.12 (0.50)	0.13 (0.66)				-	-	
time	0.32 (0.88)	1.00 (1.97)	0.76 (1.94)	0.35 (0.98)						

the recommendations based on their preference or difference user needs [17]. The design is aim to expedite the decision process. Based on the linguistic analysis, we found the users felt more in charge while using the controllable interface, which was supported by the high score on the *clout* category. We also observed the score of *analytic thinking* was lower. The results indicated when the controllable interface was provided, the user feedback shown more on *filtering* the recommendations instead of *determining* the good recommendations. We also found the controllable interface was used less big words and analytic words in their statements than the non-controllable interface (either baseline or the explainable interface). The interaction was showing less emotional, which may also be implied it is more rational and precise.

We analyzed the linguistic changes between controllable and non-controllable interfaces. We observed the users mentioned more *interrogative* words while making the decision. The users tend to "fine-tune" the relevance sliders to a point and pick the recommendations on the top. The user feedback went through a *conditional decision process* that the users tend to select top-recommended items *when* one relevance slider was tuned (high or low). The controllable interface can be seen as a "contrasting explanation" that users are tuning "counterfactual cases" in comparing the recommendations [29]. The tuning process provides evidence to gain the confidence of selecting the recommendation. The users tend to use more *social* words and focus on the *present* time as well as mentioned more motion words, which indicated the interaction between the users and the interactive interface component.

The interaction was also a cognitive load that prevents the users from inspecting the profile of the further recommend scholar (e.g., the publications and the scholars' profiles). The users tend to set a clear goal to efficiently *search* or *compare* the scholars who fitted certain criteria. Based on the user feedback, we found the rationale of the "*why do they want to select the scholars*" is less salient. The decision process may be less thoughtful due to the low score on *analytic* category. Based on the sense of *motion* and *female* aspects, the gender information seems more appearing when the recommender interface was not explainable. In addition, when both controllable and explainable components were enabled, the controllable component was adopted more by the users.

Explainable Recommender Interfaces

The goal of an explainable recommender interface is to make the recommendation reasoning process or outcome transparent [49], i.e., let the user understand the rationale behind the system [41]. The design aims to gain system transparency that lets the users know the reasoning process of the recommendation models. Based on the linguistic analysis, we found, in the *EXPL* interface, the users were able to inspect the recommendation models through different visual explanations. We observed the users did use more complicated words in their text, which leads to the decision with analytic thinking. The users were able to make thoughtful decisions based on the information provided. The users tend to use more certain words that are less hesitation as well as show more positive emotion during the decision-making process.

Due to the users cannot directly interact with the interface, the users used less *motion* words in their text. We observed the users mentioned less *interrogative* words while making the decision, it is not conditional decision-making that causes by user-driven filtering. It is surprising to observe that, when explanations were granted, the usage of female words was reduced. One possible reason is when limited information was provided; the gender characteristic was more appealing so that the users may make their decision based on this. Providing explainable recommendations may be a solution to a social recommender when gender equality is a major concern.

Providing extra explanations also brings a significant cognitive load that we observed the users significantly reduced the word counts in their statements. The detailed information may *distract* the users and make them less confident in the decision-making process. It may be another *information overloading* to the users, which may violate the original intention of introducing a recommender system into the decision-making process [57]. The explainable interface is like a double-sided blade that can help the users to make thoughtful decisions but required extra cognitive loads to process the explanations, which may not be useful in many low-stake situations. For example, recommendation books, movies, and items for purchases. Instead, a useful recommender system may "shield" the distracting information so the users can make the decision faster. We argue the extra explanations are useful only when the decision is high-stake, and the users really need it. However, it is not surprising the users perceived the providing explanations with *bias* [29]. The finding suggests to further explore the possible solution to identify *when* and *how* to provide these explanations to the users, so the providing explanation can be customized based on the users' mental model.

Design Implications

We believe both controllable and explainable interfaces contribute to different *level of transparency*. The controllable and explainable interfaces focus more on the *fusion* [32, 35, 51] and *rationale* [14, 49, 53] of hybrid recommendation models, respectively. Controllable user interfaces are known to improve recommendation efficiency, i.e., to help the users to make decisions faster [47]. However, a user-controlled interface for the hybrid fusion of recommender sources cannot assure that the users understand the underlying rationale of the recommendation algorithm [17]. In the case when the recommendation mechanism is too complicated for non-professional users to explain, some considerable transparency could be achieved by explainable user interfaces.

We found it is important to further consider the psychological states of designing the countable and explainable interfaces, e.g., the personal characteristics [19]. We argue the users may need a different level of transparency in their information decision process. Foe example, the Master students tend to *accept* the recommendations from the system, due to they do not have related research experience (e.g., publishing papers). However, the doctoral students tend to *confirm* the recommendation before *accept* them, e.g., the users will confirm the co-authorship similarity explanations to see if they can find the known scholars in the graph. The psychological states can be a lens to understand how we can design the interfaces for users with different expectations and information needs.

Our findings shed light on further consider the level of *fidelity* in the controllable and explainable interfaces. For example, the current controllable interface was letting the user tune the recommendation weighing, which unified in all recommendation models to reduce the cognitive loading. However, a high-level fidelity interface can provide a different level of detail, that is, to let the users tune the parameters of the algorithm for fine-tune the recommendation models. The experimental explainable interfaces are more like mid-level fidelity that we hide many algorithmic details of the recommendation models. For example, in the explanation of topic modeling, we only showed a few topics and topical words, which was not full disclosure of all the details. It is crucial to further consider the cognitive loading of the user and customize the design for users with different backgrounds or expertise.

We also find it is rewarding to further consider gender awareness in designing the recommender interfaces. Based on our study, when more information (explanation) was provided, gender awareness was gone. It shed light on fining an effective design in prompting or hiding the gender information and provide more empirical evidence that influences the decision-making process. For example, to further control different explanations in a social recommender and identify how the users are influenced by the appearance. We believe the explainable interface can be a promising solution to the prospect of gender bias that brings by AI-driven applications.

CONCLUSION

In this paper, we analyzed user free-form feedback in a controllable and explainable social recommend system collected through a controlled study (N=50). Our main focus were the user-submitted statements specifying the reason for selecting a recommendation (a scholar to meet at an academic conference) based on a scenario of finding academic advisors. Based on the user-generated text, we conducted a pioneering attempt to understanding the psychological states of using a controllable and explainable social recommender system. We applied a linguistic analysis to the collected data and discussed the results in three linguistic dimensions: *Grammar*, *Summary*, and *Psychological Process* variables, with a total of 25 categories were analyzed. Based on the pairwise testing and analysis, we discussed the users' psychological state changes in using the controllable and explainable social recommender components. Our works provided empirical evidence on how these components affect the user's social decision process. We then discussed the design implications based on our findings.

LIMITATIONS

We are aware of some limitations in our analysis. First, each user-generated text is relatively short and may not reliably reflect the users' psychological states. We tried to combine multiple statements into one to solve this issue. There is room to improve the data quality by expanding the length of the text-based user feedback. Second, the user-submitted text usually covered just a few linguistic categories, i.e., we will have many zeros in our dataset. The p-value may be influenced by the zeros in our datasets. Third, we noticed some typos in the user-submitted text so some words might be misinterpreted in the linguistic analysis. A further correction will be considered to improve data quality. Fourth, some linguistic aspects were influenced by paper titles. For example, in the UMAP conference, there was much paper contains the word *social*, which will be counted as a social word in the linguistic analysis, which may need further correction in the future studies.

REFERENCES

[1] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental Models of Mere Mortals with Explanations of Reinforcement Learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 2 (2020), 1–37.

[2] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 35–42.

[3] Peter Brusilovsky, Jung Sun Oh, Claudia López, Denis Parra, and Wei Jeng. 2016. Linking information and people in a social system for academic conferences. *New Review of Hypermedia and Multimedia* (2016), 1–31.

[4] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[5] Jilin Chen, Werner Geyer, Casey Dugan, Michael Muller, and Ido Guy. 2009. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 201–210.

[6] Michael A Cohn, Matthias R Mehl, and James W Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science* 15, 10 (2004), 687–693.

[7] Cecilia di Sciascio, Peter Brusilovsky, and Eduardo Veas. 2018. A study on user-controllable social exploratory search. In *23rd International Conference on Intelligent User Interfaces*. ACM, 353–364.

[8] Cecilia di Sciascio, Vedran Sabol, and Eduardo E Veas. 2016. Rank As You Go: User-Driven Exploration of Search Results. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, 118–129.

[9] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. 2019. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 408–416.

[10] 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.

[11] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[12] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[13] Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive recommender systems: a survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56 (2016), 9–27.

[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*. ACM, 241–250.

[15] Mohieddin Jafari and Naser Ansari-Pour. 2019. Why, when and how to adjust your P values? *Cell Journal (Yakhteh)* 20, 4 (2019), 604.

[16] Irving L Janis and Leon Mann. 1977. *Decision making: A psychological analysis of conflict, choice, and commitment.* Free press.

[17] Dietmar Jannach, Sidra Naveed, and Michael Jugovac. 2016. User control in recommender systems: Overview and interaction challenges. In *International Conference on Electronic Commerce and Web Technologies*. Springer, 21–33.

[18] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. 2010. Comparison of implicit and explicit feedback from an online music recommendation service. In *proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems*. 47–51.

[19] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of personal characteristics on music recommender systems with different levels of controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 13–21.

[20] Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology* 33, 2 (2014), 125–143.

[21] Been Kim. 2015. *Interactive and interpretable machine learning models for human machine collaboration*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[22] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and Control in Social Recommenders. In *6th ACM Conference on Recommender System*. 43–50.

[23] Bart Piet Knijnenburg, Martijn Willemsen, and Ron Broeders. 2014. Smart sustainability through system satisfaction: Tailored preference elicitation for energy-saving recommenders. (2014).

[24] Bart P Knijnenburg and Martijn C Willemsen. 2015. Evaluating recommender systems with user experiments. In *Recommender Systems Handbook*. Springer, 309–352.

[25] 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 (2012), 441–504.

[26] 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*. 84–88.

[27] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 379–390.

[28] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To Explain or not to Explain: the Effects of Personal Characteristics when Explaining Music Recommendations. In *Proceedings of the 2019 Conference on Intelligent User Interface*. ACM, 1–12.

[29] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[30] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin* 29, 5 (2003), 665–675.

[31] Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. 2020. Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 183–191.

[32] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1085–1088.

[33] Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. 2012. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery* 24, 3 (2012), 555–583.

[34] Albert Park and Mike Conway. 2017. Longitudinal changes in psychological states in online health community members: understanding the long-term effects of participating in an online depression community. *Journal of medical Internet research* 19, 3 (2017), e71.

[35] Denis Parra and Peter Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies* 78 (2015), 43–67.

[36] James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one* 9, 12 (2014), e115844.

[37] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[38] Inc. Pennebaker Conglomerates. 2020. Interpreting LIWC Outputs. (2020). `https://liwc.wpengine.com/interpreting-liwc-output/`

[39] Ulrike Pfeil, Raj Arjan, and Panayiotis Zaphiris. 2009. Age differences in online social networking–A study of user profiles and the social capital divide among teenagers and older users in MySpace. *Computers in Human Behavior* 25, 3 (2009), 643–654.

[40] Pearl Pu and Li Chen. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20, 6 (2007), 542–556.

[41] Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies* 1, 1 (2019), 33–36.

[42] Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable AI and users' mental models. In *CHI 2019 Workshop: Where is the Human? Bridging the Gap Between AI and HCI*. Association for Computing Machinery, Inc.

[43] J Ben Schafer, Joseph A Konstan, and John Riedl. 2002. Meta-recommendation systems: user-controlled integration of diverse recommendations. In *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, 43–51.

[44] Amit Sharma and Dan Cosley. 2013. Do social explanations work?: studying and modeling the effects of social explanations in recommender systems. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1133–1144.

[45] Alain Starke, Martijn Willemsen, and Chris Snijders. 2017. Effective user interface designs to increase energy-efficient behavior in a Rasch-based energy recommender system. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 65–73.

[46] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[47] Nina Tintarev. 2017. Presenting Diversity Aware Recommendations: Making Challenging News Acceptable. (2017).

[48] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (1 Oct. 2012), 399–439.

[49] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.

[50] Chun-Hua Tsai. 2020. *Controllability and explainability in a hybrid social recommender system*. Ph.D. Dissertation. University of Pittsburgh.

[51] Chun-Hua Tsai and Peter Brusilovsky. 2017. Providing Control and Transparency in a Social Recommender System for Academic Conferences. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. ACM, 313–317.

[52] Chun-Hua Tsai and Peter Brusilovsky. 2018. Beyond the Ranked List: User-Driven Exploration and Diversification of Social Recommendation. In *23rd International Conference on Intelligent User Interfaces*. ACM, 239–250.

[53] Chun-Hua Tsai and Peter Brusilovsky. 2019a. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 391–396.

[54] Chun-Hua Tsai and Peter Brusilovsky. 2019b. Exploring Social Recommendations with Visual Diversity-Promoting Interfaces. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 1 (2019), 1–34.

[55] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 351–362.

[56] Yao Wu and Martin Ester. 2015. FLAME: A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborative Filtering. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. ACM, New York, NY, USA, 199–208.

[57] Bo Xiao and Izak Benbasat. 2007. E-commerce product recommendation agents: use, characteristics, and impact. *MIS quarterly* 31, 1 (2007), 137–209.