Controlling Spotify Recommendations: Effects of Personal Characteristics on Music Recommender User Interfaces

Martijn Millecamp
Department of Computer Science, KU Leuven
Leuven, Belgium
martijn.millecamp@cs.kuleuven.be

Nyi Nyi Htun
Department of Computer Science, KU Leuven
Leuven, Belgium
nyinyi.htun@cs.kuleuven.be

Yucheng Jin
Department of Computer Science, KU Leuven
Leuven, Belgium
yucheng.jin@cs.kuleuven.be

Katrien Verbert
Department of Computer Science, KU Leuven
Leuven, Belgium
katrien.verbert@cs.kuleuven.be

ABSTRACT

The “black box” nature of today’s recommender systems raises a number of challenges for users, including a lack of trust when recommendations fail and limited user control. Providing more user control is interesting to enable end-users to help steer the recommendation process with additional input and feedback. However, different users may have different preferences with regard to such control. To the best of our knowledge, no research has investigated the effect of personal characteristics on visual control techniques in the music recommendation domain. In this paper, we present results of a user study on the web using two different visualisation techniques (a radar chart and sliders) that allowed users to control Spotify recommendations. A within-subject design with *Latin Square* counterbalancing measures was used for the study. Results indicate that the radar chart helped the participants discover a significantly higher number of new songs compared to the sliders. We also found that users’ experience with Spotify had an influence on their interaction with different musical attributes. The participants who used Spotify more than 21 hours per week interacted with the attributes significantly more with the radar chart compared to the sliders. Individual musical sophistication also had a significant impact on their interaction with the visual techniques. The participants with high musical sophistication interacted significantly more with the radar chart in comparison to the sliders. Based on the feedback from our participants, we provide design suggestions to further improve user control in music recommendation.

CCS CONCEPTS

• **Information systems → Personalization**; **Personalization**; **Recommender systems**; • **Human-centered computing** → *Visualization design and evaluation methods*;

KEYWORDS

music recommendations, interactive recommender systems, visualisation, personal characteristics

ACM Reference Format:
Martijn Millecamp, Nyi Nyi Htun, Yucheng Jin, and Katrien Verbert. 2018. Controlling Spotify Recommendations: Effects of Personal Characteristics on Music Recommender User Interfaces. In *Proceedings of ACM UMAP conference (UMAP’18).* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Many services that are available on the World Wide Web today utilise recommender algorithms to suggest personalised contents to their users. Amazon.com, Spotify, YouTube and Netflix are well-known examples of such services. With the help of recommender algorithms and abundant data, many of these services can provide users with highly relevant items leading to improved user satisfaction and content discovery.

All recommender systems, at their basic form, rely on user behaviours or preference by recording activities of users who interact with a service or system, or by simply asking users for their preference. While much of previous research has focused on developing and evaluating recommender algorithms, recommender systems still face a number of challenges. One such a challenge is known as the lack of “transparency” [35], meaning that existing recommender systems do not provide users with any insight into the logic of recommendations. To many users, this “black box” raises the issue of not understanding why they are receiving certain recommendations, which in turn could lead to trust issues [22].

In an attempt to combine searching and browsing functionalities, Freyne et al. [16] developed a system based on a community-based web search engine, I-SPY [36], that augments each search result with previous search behaviours of a community. Different icons were attached to each search result to visualise interactions of the community with each result item (e.g. result-query relevancy based on the community, result popularity, related query, etc.). Although this system provides a visualisation of the choices that previous users have made, it did not allow users to interact with the recommendation process itself. Having a control over the recommendation process itself may also help address the “transparency” issue, as

UMAP’18, July 2018, Singapore

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

previous research has shown that the relation between satisfaction and user control is affected by the knowledge level of users [28] and their interests [23].

O'Donovan et al. [31] also argued that many recommender systems are opaque to users and that users need more control in order to tailor recommendations to their current moods and influences. To address this issue, O'Donovan et al. [31] designed PeerChooser to provide users with a visual explanation of the recommendation process and the opportunity to manipulate input weightings to steer the recommendations. SmallWorld, designed by Gretarsson et al. [20], was similar to PeerChooser in that it allowed users to manipulate the recommendation process with additional visual techniques. However, these authors [20, 31] focused on movie and social media domains respectively. Jin et al. [25] investigated the effect of user control on cognitive load and acceptance of recommendations in the music domain and found that a higher level of control produced the best recommendations, while requiring the highest cognitive load. To the best of our knowledge, no research has been conducted to understand the effect of personal characteristics on the perception of visual control techniques in the music domain. Therefore, in this paper, we attempt to address the following research questions:

RQ1: In what way do personal characteristics influence *perception* of the visual control techniques in music recommendation?

RQ2: In what way do personal characteristics influence *interaction* with the visual control techniques in music recommendation?

RQ3: Which visual control technique(s) is/are better suited for users to manipulate music recommendation?

RQ4: How can we design an interface to allow better user control for music recommendation?

The remainder of the paper is organised as follows: in Section 2, we discuss background work on interactive recommender systems, as well as existing work on different personal characteristics that may influence the utility of such systems. In Section 3, we describe our experimental design, including the interface, study procedure, and evaluation metrics. In Section 4, we present the results of the study. In Section 5, findings and their implications are discussed based on the evaluation results. Finally, in Section 6, we conclude the paper by highlighting our contributions and by discussing possible future work.

2 BACKGROUND

2.1 Interactive Recommender Systems

Interactive recommender systems have been researched to some extent by the research community of the past two decades. For example, PeerChooser [31] and SmallWorlds [20] are two approaches that focus on interaction with collaborative filtering recommender engines to users. Both systems allow users to interact with representations of relations between items and other users to support transparency and user control. PeerChooser uses a graph-based visualisation to represent these relationships. SmallWorlds allows users to explore the relationships between recommended items and similar friends in multiple layers of similarity.

In addition, a number of visualisations have been developed to interact with hybrid recommender systems. TasteWeights [4] is a system that allows users to control the impact of different algorithms as well as different input data sources on the recommendation results, eliciting preference data and relevance feedback from users at runtime in order to adapt recommendations. This idea can be traced back to the work of Schafer et al. [34] on meta-recommendation systems, where users are provided with personalised control over the generation of recommendations by altering the importance of specific factors on a scale from 1 to 5. Similarly, SetFusion [32] is a more recent example that allows users to fine-tune weights of a hybrid recommender system, using a Venn diagram [38] to visualise relationships between recommendations. MoodPlay [2] is a hybrid music recommender system that integrates different techniques in an interactive interface supporting explanation and control of affective data. The system allows the user to explore a music collection through latent affective dimensions, thereby improving acceptance and understanding of recommendations. MyMovieMixer [30] is an interactive movie recommender that integrates different recommender techniques with interactive faceted filtering methods, called "blended recommending". The approach allows users to interact with a set of filter facets representing criteria that can serve as input for different recommendation methods, including collaborative and content-based filtering.

Our previous work [11] is focused on various factors that affect acceptance of recommendations, such as user satisfaction, trust and sense of control. Specifically, based on the analysis of research on interactive recommender systems, we derived a framework proposing five important attributes for trust-aware and interactive recommender systems, namely: transparency, controllability, justification, diversity and context. We also investigated how information visualisation can improve user understanding of the rationale behind recommendations in order to increase their perceived relevance and meaning and to support exploration and user involvement in the recommendation process. To this end, we performed a study using TalkExplorer [40], an interactive visualisation tool developed for attendees of academic conferences based on a Cluster Map [18]. We combined different user-generated data sources in the study, but rather than automatically merging these data as it is done in hybrid recommender systems, end-users were allowed to select which users or tags should be considered. In addition, users could select different recommendation techniques that are represented as agents, similar to Ekstrand et al.'s [15] idea of enabling users to switch between recommenders. While the results of user studies indicated an increase in recommendation effectiveness when using the visualisation as opposed to a ranked list representation, we found that non-technical users did not receive the same benefits from the interface as technical users as they only interacted in a limit way with the interface [39]. In this paper, we are researching the use of two different representations to interact with recommender systems: a radar chart and sliders. The overall objective is to gain insight into the utility of more advanced versus simpler visualisations for interacting with recommender systems and the interplay of these representations with different personal characteristics.

2.2 Personal Characteristics

The influence of personal characteristics on the performance of users has been researched elaborately. In this existing body of research, the influence of a variety of personal characteristics has been investigated. Because of this variety, we will use the classification of Aykin et al. [3] to list the different personal characteristics.

2.2.1 Level of experience. One of the most common researched personal characteristic is expertise or experience [1, 3, 9, 13, 14, 24, 37, 44]. Depending on the domain of the research, this experience is measured in different ways. For example, in a user interface domain, experience can be seen as the experience with computers [44] or the visualisation expertise [9, 13].

2.2.2 Demographic characteristics. Demographic characteristics have also been researched extensively [3, 8, 10, 14, 17, 29].
Some research only takes into account basic demographic characteristics such as age, sex and gender [3, 14, 29]. Other research goes deeper and investigates personal interests, goals, background, country, education level, marriage status, (sector of) job, income and first language [8, 10, 17, 44].

2.2.3 Personality traits. In previous research, it is been shown that personality traits can have an impact on the performance and preference of a user [3]. Aykin et al. [3] list seven different personality traits: Jungian personality type, Field dependence/independence, Locus of control, Imagery, Spatial ability, Type A/B personality and Ambiguity tolerance. Brusilovsky [8] researched more in depth cognitive and learning styles. Other research investigated colour characteristics such as colour perception, colour memory, colour ranking [14] and psychograpic or psychological characteristics (e.g. sensitivity, disabilities, emotion, etc.) [29].

2.2.4 Others. There are a number of other, user related, characteristics that are influencing perception and performance of the user [3]. One popular category of personal characteristics not mentioned above are cognitive skills [1, 8, 9, 13, 14, 37]. Especially working memory is a popular metric that is commonly measured, except for the research of [14]. Working memory can be split up in visual and verbal working memory, but because the number of text in both interfaces is the same, we only measured the visual working memory in our research.

In our research, we measured *level of experience* as musical sophistication, the number of hours participants listen to Spotify, and tech-savviness, *demographic characteristics* as age and gender, and *visual working memory.* Musical sophistication was measured by 10 seven point Likert scale questions based on the Goldsmiths Musical Sophistication Index (Gold-MSI)[1]. Visual working memory was measured by a block-tapping test based on Corsi block-tapping [27]. As for tech-savviness, participants were asked to rate themselves between *confident*, *not confident* and *somewhere in-between* when it come to trying new technology.

3 EXPERIMENTAL DESIGN

The following sub-sections present a detailed description of the experimental design deployed in the study.

3.1 Participants

Participants were recruited from personal contacts, Reddit, research groups and university contacts for the study. A total of 80 people participated, of which 40 were removed as they did not finish the study. Of the remaining 40 participants, 10 were female and 30 were male. Twenty-three participants belonged to the age group of 15-24, 15 to the group of 25-34 and 2 to the group of 35-44. We also asked the participants to report their confidence with trying new technology. Twenty-nine out of 40 participants reported that they were confident with trying new technology, while one participant reported to feel not confident and 10 participants reported that they were somewhere in-between. Regarding Spotify usage, 8 out of 40 participants reported that they used Spotify between 1 and 5 hours per week, 10 between 6 and 10 hours per week, 11 between 11 and 15 hours per week, 2 between 16 and 20 hours per week and 9 more than 21 hours per week. For visual working memory, the participants were divided into two group at the 50th percentile. Both low and high visual working memory groups had 20 participants each. Similarly for music sophistication, the participants were divided into two groups at the 50th percentile. The high music sophistication group at 18 participants and the low group had 22 participants.

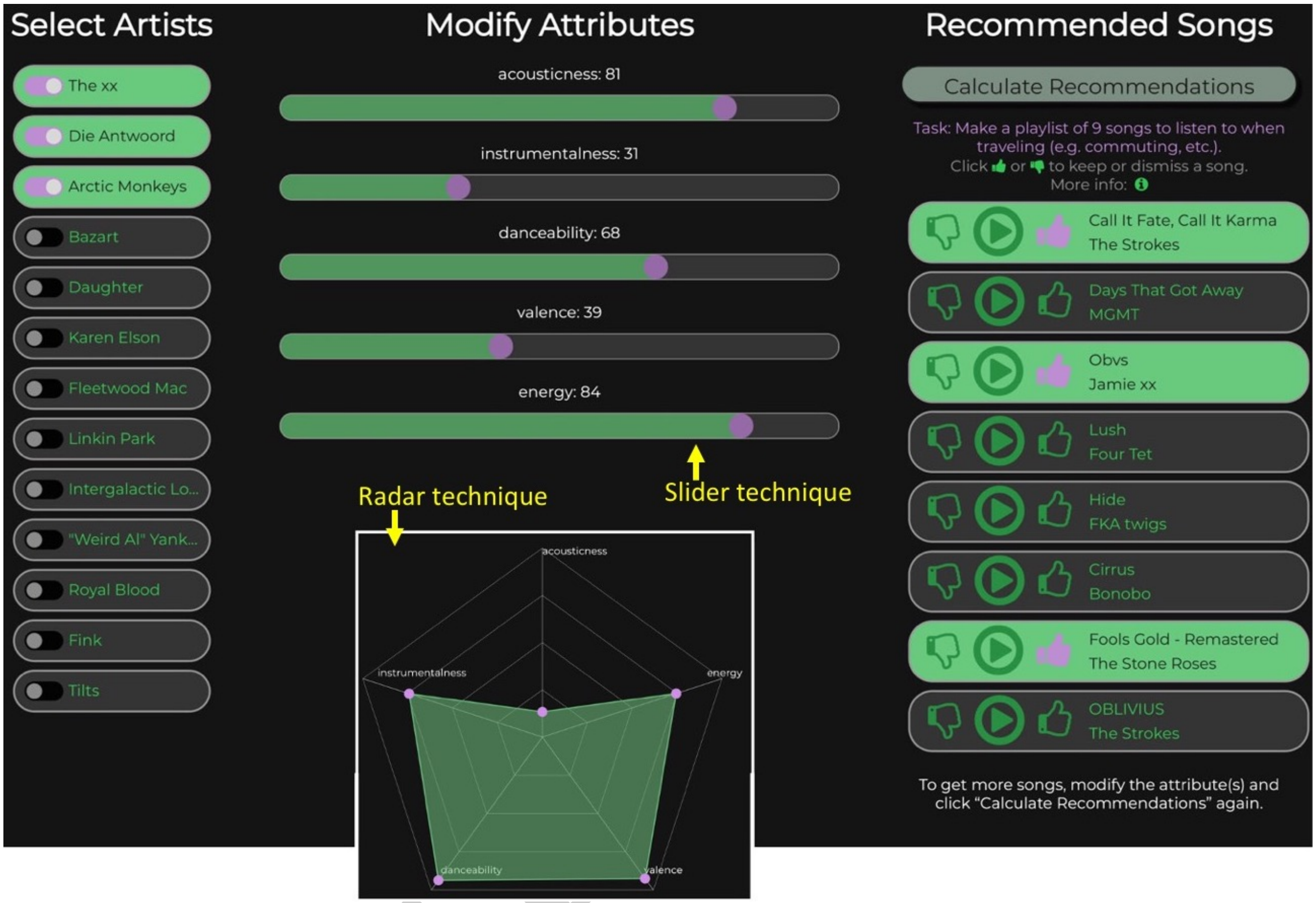
3.2 Implementation

In our earlier work, we analysed existing interactive recommender systems in detail [11] and found different visualisation techniques that have been used to support user control as a basis to improve recommendations or to explore the recommendation space. Among these techniques, we found that sliders and graphs with draggable and droppable elements are the most popular. Sliders are often just visualised below each other [4, 32]. Other draggable elements are often visualised in a circle [12, 26, 41–43] or draggable nodes in a graph [7, 20, 31]. Because of the common use of sliders and draggable and droppable elements, we also adopted these elements, but in two different modalities: sliders and a radar chart, both with draggable and droppable elements. As shown in Figure 1, the two visual control techniques were implemented into two separate interfaces. Both interfaces were designed using a 3 column format similar to previous music recommender systems [4, 5, 25, 32]. The column on the left side enables users to select artists from the list of top artists they listen to. These artists are used as input for generating recommendations. The visualisation in the second column represents different parameters that can be used to adjust recommendations. Users can for instance increase the weight of parameters such as danceability and energy. The third column represents the generated recommendations. As explained above, two different visualisations were implemented to enable users to adjust parameters weights: sliders and a more advances, and potentially more appealing, radar chart.

The Spotify API[2] allows to generate recommendations based on up to 5 favourite artists. In addition, the API also allows modification of 14 musical attributes in order to describe one's preferred musics. We selected 5 out of 14 available musical attributes based on [19], where the authors found that different song genres can be represented by only three categories: arousal, valence and depth. According to the authors, *arousal* represents intensity and energy in music, *valence* the spectrum of emotions in music, and *depth* the

[1] https://www.gold.ac.uk/music-mind-brain/gold-msi/

[2] https://developer.spotify.com/web-api/get-recommendations



Task: Make a playlist of 9 songs to listen to when traveling (e.g. commuting, etc.).
Click 👍 or 👎 to keep or dismiss a song.
More info: ℹ

To get more songs, modify the attribute(s) and click "Calculate Recommendations" again.

Figure 1: Interface variations used in the study.

intellect and sophistication in music. In line with these three categories, the 5 music attributes we selected were: energy, danceability, valence, instrumentalness and acousticness.

As in the Spotify application, we presented each recommended song by its title and artist. Album art and album name were not displayed in order to have a clean and manageable layout. The Spotify API provides a way to play a preview of up to 30 seconds for each recommended song (complete songs are inaccessible). We attached this feature with a play button in our interfaces which allowed users to listen to a preview of the recommended songs. Similar to the Spotify radio feature, we used "Thumb up" and "Thumb down" buttons to allow users to like or dislike the recommended songs. Disliking a song dismisses it from the list whereas liking a song keeps it in the list.

3.3 Study Procedure

The study was conducted online with participants recruited from personal contacts, Reddit research groups and university contacts. As soon as the study URL was loaded, participants were presented with detailed information about the study and a consent form. After they agreed to participate in the study, a new page was presented where they had to authorise their Spotify account to be accessible by the study. Following the authorisation, a new page with demographic questions was presented to collect individual participants' personal characteristics which include age, gender, music sophistication, visual working memory, tech-savviness, Spotify usage, familiarity with recommender systems and attitude towards recommender systems.

Once the demographic questions were completed, the participants were shown detailed instructions and a task requesting them to make a playlist of 9 songs to listen to when travelling (e.g. for commuting). This task was chosen for the study because travelling and personal maintenance are the two most common activities associated with listening to music [21]. Next, the participants completed the task by selecting their favourite artists, manipulating the musical attributes and selecting 9 songs from the recommendations. The participants were then presented with a number of evaluation questions which included a set of questions from the ResQue user-centric recommender systems evaluation framework [33] and open-ended questions (see Section 3.4.1 for details).

Next, the participants were displayed a different task and instructions which requested them to make a playlist of 9 songs to listen to during their personal maintenance. Similar to the steps for the first task, the participants completed the second task and answered another set of evaluation questions. The participants were then asked to complete a set of open-ended exit questions (see Section 4.2 for details).

To counteract any fatigue effect during the study, the order in which the interfaces were presented was rotated using a *Latin Square* counterbalancing measure.

3.4 Evaluation Metrics

3.4.1 Recommender System Evaluation Questions. To evaluate the interfaces, a set of modified ResQue questions [33] and open-ended questions were used. A total of 13 ResQue questions were selected and made minor modifications to fit our evaluation requirements. These questions are as follow:

- The songs recommended to me are of various kinds (Q1).
- The songs recommended to me are similar to each other (Q2).
- This recommender system helped me discover new songs (Q3).
- I haven't heard of some songs in the list before (Q4).
- This recommender system helped me find ideal songs (Q5).
- Using this recommender system to find what I like was easy (Q6).
- This recommender system gave me good suggestions (Q7).
- Overall, I am satisfied with this recommender system (Q8).
- I am convinced of the songs recommended to me (Q9).
- This recommender system made me more confident about my selection/decision (Q10).
- I will use this recommender system again (Q11).
- I will tell my friends about this recommender system (Q12).
- I will keep the songs recommended so that I can listen again (Q13).

The questions were in the form of 7-point Likert scales and the answers ranged from 1 (strongly disagree) to 7 (strongly agree). Following the ResQue questions, a number of open-ended questions were also administered to capture feedback from the participants about the most and the least useful parts of each interface.

3.4.2 Interaction Log. Both interfaces recorded a log of the participants interactions with different components. Specifically, the log captured:

- The number of times the musical attributes were changed (*attribute*).
- The number of times any given musical attribute was changed (*nb + attribute name*, e.g. nbEnergy for the energy attribute)
- The number of times the "Calculate Recommendations" button was clicked (*calculate*).
- The number of times the dislike button was clicked (*disliked*).
- The number of times the like button was clicked. (*liked*)
- The total number of clicks on all components of the interface throughout a session. (*interactions*)

This log was then used to understand the impact of participants' personal characteristics on their interactions with the interfaces.

3.4.3 Exit Questions. In the exit questions, the participants were first presented with all of the 14 musical attributes, as well as their definitions, that are supported by the Spotify API. They were then asked to rate each of these musical attributes on a scale of 1 (least likely) to 7 (most likely) to indicate how likely they will use it in order to control music recommendations. This allowed us to understand if there may be other potentially useful attributes than those we used in our current interfaces. Finally, the participants were asked to suggest any other visual techniques that they think may be helpful for them to better control the recommendation process.

Table 1: Result of Wilcoxon signed rank tests showing significant differences between the two interfaces without personal characteristics. M = mean, SD = standard deviation

	Z	p	Radar Chart			Slider		
			M	Median	SD	M	Median	SD
Q3	-2.623	0.009	4.72	5	1.8	4.2	5	1.7
liked	-2.073	0.038	11.2	9.5	4.6	9.975	9	3.5
nbEnergy	-2.032	0.042	5.475	4	4.5	4.275	2	5.8

4 RESULTS

A normality test was initially performed on data using the Shapiro-Wilk test. To compare the differences between the two interfaces, we used t-tests for normally distributed data and Wilcoxon signed rank tests for non-normally distributed data. Independent variables were the interfaces whereas dependent variables were questionnaire scores and log data. Details of the statistical analysis results are presented in Section 4.1. Responses for open-ended questions were analysed using Thematic Analysis [6] and the results are presented in Section 4.2.

4.1 User Perception and Log Analysis

4.1.1 Comparisons without Personal Characteristics. To understand an overall difference between the two interfaces, comparisons were firstly performed between the interfaces without taking personal characteristics into account. Results showed that the two interfaces yielded significantly different outcomes in terms of interactions and perceived discovery of new songs. As shown in Table 1, the radar chart interface had a significantly higher score than the slider interface (Z = -2.623, p = 0.009) in terms of perceived discovery of new songs (Q3). Looking at Figure 2, it was found that the scores between all of the 13 questions were similar. Participants rated the highest for Q4 (I haven't heard of some songs in the list before). Having the ability to control musical attributes may have allowed the participants to discover the songs which may otherwise have been ignored by either themselves or the system.

In addition, the participants used the like button significantly higher in the radar chart interface (Z = -2.073, p = 0.038). Since the participants were required to choose exactly 9 songs in each interface, it appears that they refined their list as more favourite songs appeared through the session which led to using the like button more frequently in the radar chart interface. Interestingly, it was also found that the energy attribute was more frequently used in the radar chart interface (Z = -2.032 p = 0.042). Although not proven, we suspect that this could be due to the way the attributes

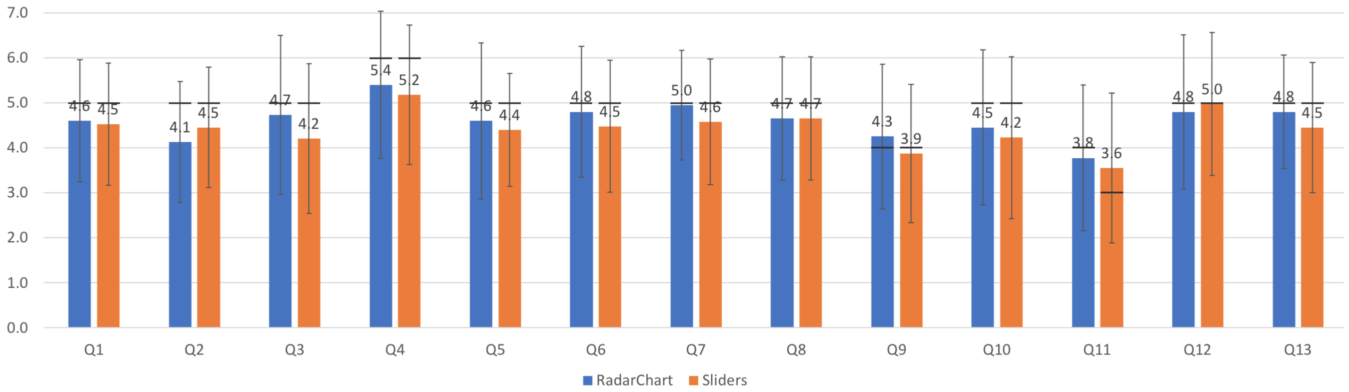


Figure 2: Results of the ResQue questions between the two interfaces without personal characteristics. Values above the columns = mean

are positioned in the two interfaces. Further studies should be performed to investigate this.

4.1.2 Comparisons by Personal Characteristics. To analyse the effect of personal characteristics, the participants were divided based on each category of their personal characteristics (e.g. male, female, high music sophistication, low music sophistication, etc.). According to the analysis results, certain characteristics such as age, gender, familiarity with recommender systems, and attitude towards recommender systems had no significant impact on usage and perception between the two interfaces. In the followings, results of other characteristics that had significant impact on usage and perception between the two interfaces are presented.

Tech-savviness: Tech-savviness was divided into three categories by individuals who are 1) *confident*, 2) *not confident*, and 3) *somewhere in-between*. Only the participants in the category of *somewhere in-between* indicated that they found more ideal songs (Q5) with the radar chart interface than with the sliders interface (Z = -2.032, p = 0.042, see Table 3). For the participants in the *confident* and *not confident* categories, however, neither of the interfaces was significantly better at finding ideal songs than the other.

Spotify Usage: Interestingly, the participants who use Spotify more than 21 hours a week interacted with the musical attributes significantly higher within the radar chart interface (Z = -2.08, p = 0.038, see Table 3). In addition to this, for the participants in the same category, we found that the instrumentalness attribute was used significantly more with the radar chart interface (Z = -2.27, p = 0.023, see Table 3). It appears that experience with Spotify may play a role when it comes to interaction with the musical attributes for certain visual techniques.

Musical Sophistication (MS): Based on their musical sophistication, the participants were divided into high and low MS categories. Interestingly, those with a high MS had significantly higher overall interactions with the radar chart interface (Z = -2.2, p = 0.028, see Table 3). In addition for the radar chart interface, the same group of participants had significantly higher interactions with the acousticness attribute (t(17) = 2.114, p = 0.05, see Table 2) and the "Calculate Recommendations" buttons (Z = -2.078, p = 0.038, see Table 3). On the contrary, the participants with a low MS had significantly higher interactions within the sliders interface with the acousticness attribute (t(21) = -2.46, p = 0.015, see Table 2) and the "Calculate Recommendations" buttons (Z = -2.138, p = 0.033, see Table 3). This suggests that an individual's musical sophistication may also greatly impact on interaction with the musical attributes for certain visual techniques.

Table 2: Result of t-tests showing significant differences between the two interfaces based on personal characteristics. M = mean, SD = standard deviation

Personal Characteristics	Category	Metrics	t	df	p	Radar Chart		Slider	
						M	SD	M	SD
MS	low	nbAcousticness	-2.46	21	0.015	3.86	3.47	5.81	8.89
	high	nbAcousticness	2.114	17	0.05	4.83	4.72	2.89	2.40
VWM	low	nbAcousticness	-2.238	19	0.037	3.25	2.98	4.9	5.74

Visual Working Memory (VWM): To test the visual working memory of individual participants, we used a test based on the Corsi block-tapping test [27]. Just as musical sophistication, we divided the participants into low and high VWM categories. Again, we found that the participants with a high VWM had significantly higher interactions with the danceability attribute within the radar chart interface (Z = -2.71, p = 0.007, see Table 3). Meanwhile, the participants with a low VWM had significantly higher interactions with the acousticness attribute within the sliders interface (t(19) = -2.238, p = 0.037, see Table 2).

4.2 Design Feedback

A thematic analysis [6] of the responses for the open-ended questions resulted in three main themes: track-attribute visualisation, relevance feedback and usability. We present these themes in detail in the following sub-sections.

4.2.1 Track-Attribute Visualisation. Four participants reported that a visualisation of the relationship between recommended tracks and selected attributes may be helpful for better understanding of the recommended tracks. For example, one participant explained that "one way to gain intuitive understanding would be to have an option to scroll through a 2d space of different features such as valence vs mode, and be able to see some of my most played tracks

Table 3: Results of Wilcoxon signed rank tests showing significant differences between the two interfaces based on personal characteristics. M = mean, SD = standard deviation

Personal Characteristics	Category	Metrics	Z	p	Radar Chart			Sliders		
					M	Median	SD	M	Median	SD
Tech-savviness	somewhere in-between	Q5	-2.032	0.042	5.5	6	0.97	4.4	5	1.34
Spotify usage	21+	attributes	-2.08	0.038	11.6	11	6.44	6.89	7	2.67
	21+	nbInstrumentalness	-2.27	0.023	4.2	4	1.86	2	2	1.41
MS	Low	calculate	-2.138	0.033	6.18	4	4.75	8.91	6	9.00
	High	calculate	-2.078	0.038	7.33	6	6.2	4.33	3	3.86
	High	interactions	-2.2	0.028	104.8	101.5	62.77	74.55	69	37.63
VWM	High	nbDanceability	-2.71	0.007	5.75	4	5.70	2.7	2	2.08

there." (P3). Another participant explained that "It would be great to visualise the timeline of the playlist. For example if I was creating a 2 hour long playlist it would be great to see the length of each track and the BPM associated with each song so you can better create a playlist for an extended period of time for a workout session, party or during work." (P14). We believe that users can benefit greatly from having a visual representation of the recommended songs. By looking at the relationship between the recommended songs and the attributes, they may better understand why particular songs are being recommended. At the same time, this visualisation technique itself should allow users to explore further and refine the list. In addition, we believe that by showing the songs from their playlists on this track-attribute visualisation, users can better understand their music taste and information about their favourite songs.

4.2.2 Relevance feedback. While musical attributes are a great way to describe one's preference, some might find them difficult to utilise. Therefore, one participant suggested to "Let me 'dial' [the songs I liked] to influence the future song recommendations. It requires less intimacy with the musical traits and seems less subjective? I mean, people dance to different things." (P13). We believe that when their desired musical attributes are unknown, it may be easier for users to express their preference as "show me more songs like this". The recommended songs in this case should display their relationship with the input song in the musical attribute spectrum so that users can easily understand why particular songs are being recommended.

4.2.3 Usability. The majority of the participants expressed positiveness towards having the ability to steer the recommendations. Three participants reported that they preferred the sliders and 10 reported that they preferred the radar chart. Those who preferred the radar chart also mentioned that it offers a better overview of the current settings. Therefore, while the ability to steer recommendations was seen as a good aspect, many participants preferred the radar chart.

Finally, five of the participants reported that they could not easily understand which musical attribute had an impact on a particular song in the recommended list. Therefore, one participant explained, "...it was difficult to say which aspect I should change to get better recommendations." (P4). This is similar to the findings presented in Section 4.2.1, confirming that visualisation of track-attribute relationship will be greatly beneficial for users.

5 DISCUSSION

In the context of our first research question (RQ1): "In what way do personal characteristics influence perception of the visual control techniques in music recommendation?", we found that techsaviness of an individual had an influence on the outcomes between the two visual control techniques. The participants who expressed themselves as somewhere in-between on the tech-savviness scale also indicated that they found more ideal songs with the radar chart interface than with the sliders interface. However, those who expressed themselves as either confident or not confident at trying new technology had no different responses between the two interfaces. Overall, personal characteristics do not seem to play much role on perception of the visual control techniques in music recommendation. In the future, it may be interesting to explore the outcomes of other visual control techniques.

In the context of our second research question (RQ2): "In what way do personal characteristics influence interaction with the visual control techniques in music recommendation?", we found that Spotify usage, music sophistication and visual working memory had an influence on the outcomes between the two visual control techniques. The participants with a high Spotify usage interacted with the musical attributes significantly more with the radar chart interface. In addition, the participants with a high musical sophistication had significantly higher overall interactions with the radar chart interface than with the sliders. Similarly, the participants with a high visual working memory had significantly higher interactions with the danceability attribute within the radar chart interface. Overall, the radar chart encouraged more interactions with the interface itself and musical attributes for those who had a high Spotify usage, musical sophistication and visual working memory.

In the context of our third research question (RQ3): "Which visual control technique(s) is/are better suited for users to manipulate music recommendation?", we found that the radar chart interface scored higher at discovering new songs. In addition, 10 of the participants reported that they preferred the radar chart as it is better at displaying an overview of the attribute settings. Meanwhile, only three participants reported that they preferred the sliders. Different

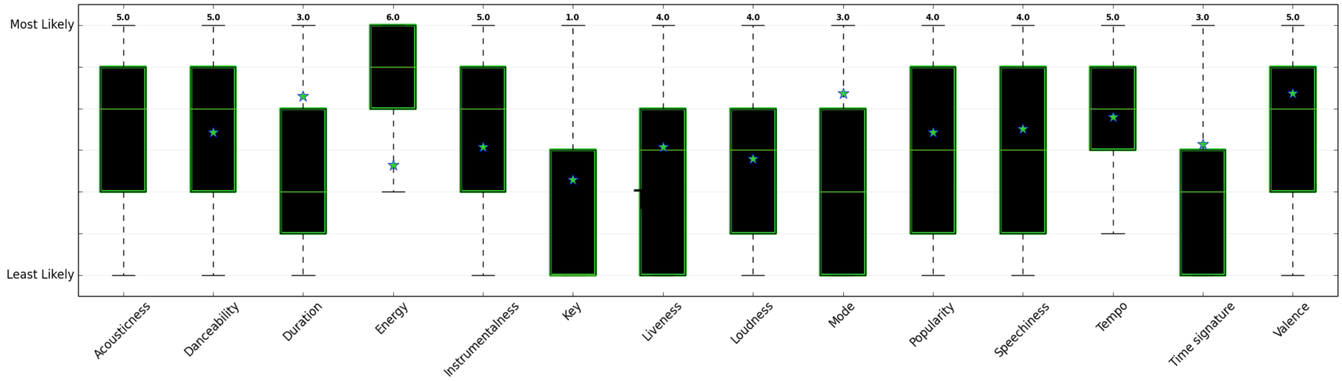


Figure 3: Evaluation of the 14 Spotify musical attributes based on how likely they will be used by participants. The star indicates the mean. The value above and the green line inside the box are representing the median.

personal characteristics did not have considerable impact on the outcomes between the two techniques.

As shown in Figure 3, amongst the 14 available musical attributes of Spotify, the majority of the participants preferred the energy attribute the most, followed by acousticness, danceability, instrumentalness, tempo and valence. The key attribute was the least preferred, suggesting that the key of a track (e.g. C, C♯, D, etc.) will not be used by many users to describe their music preference. Therefore, these 6 attributes: energy, acousticness, danceability, instrumentalness, tempo and valence should be considered when implementing a visual control technique.

In the context of our fourth research question (RQ4): “How can we design an interface to allow better user control for music recommendation?", we found a number of design considerations. The first consideration should be given to visualising the relationship between recommended tracks and musical attributes. Such a visualisation can help users understand why particular songs are being recommended to them. In addition, this visualisation technique itself should allow users to explore further songs and keep refining their list. In addition, to help users better understand their music taste and information about their favourite songs, we believe that the songs from their playlist could also be associated within this visualisation.

The second but equally important consideration is that the interface should not only support controlling musical attributes but also a way for users to express their preference by selecting a song. In the latter, the interface should also be able to visualise the relationship between the recommended songs and the input song using the musical attribute spectrum so that users can easily understand why particular songs are being recommended.

6 CONCLUSION

In this paper, we presented an online evaluation of two different visual control techniques: sliders and a radar chart for steering the music recommendations process. The two techniques were implemented into two separate music recommender interfaces. The Spotify API was employed in order to generate recommendations. The visual control techniques allowed users to manipulate five musical attributes used to produce recommendations by Spotify. A within-subject design with *Latin Square* counterbalancing measures was used in the study. A number of evaluation questions including ResQue [33] and open-ended questions were administered. Results showed that the radar chart helped the participants to discover a significantly higher number of new songs compared to the sliders. In addition, a number of participants reported that they preferred the radar chart over the sliders as it provides an overview of the musical attribute settings. Next, we found that personal characteristics did not play much role on the perception towards the visual control techniques. Interestingly, the radar chart encouraged more interactions with the interface itself and musical attributes for those who had a high Spotify usage, musical sophistication and visual working memory. When implementing a visual control technique for music recommender systems, considerations should be given to these particular music attributes: energy, acousticness, danceability, instrumentalness, tempo and valence. In addition, based on the feedback from our participants, we found that visualisation of the relationship between recommended tracks and musical attribute can be greatly beneficial for users. Also, when their desired musical attributes are unknown, it may be easier for users to express their preference by indicating a song. Recommender systems must be able to take such an input and display recommendations together with a visualisation of the relationship between the input song and the recommended songs using the musical attributes. In summary, our findings presented in this paper provide an important contribution for personalised music recommender systems. Future work should focus on implementing a new generation of music recommender systems that provide users with comprehensive visual techniques and input methods to steer the recommendation process.

7 ACKNOWLEDGEMENTS

Part of this research has been supported by the KU Leuven Research Council (grant agreement C24/16/017).

REFERENCES

[1] Azzah Al-Maskari and Mark Sanderson. 2011. The effect of user characteristics on search effectiveness in information retrieval. *Information Processing &*

Management 47, 5 (2011), 719–729.

[2] Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2016. Moodplay: Interactive Mood-based Music Discovery and Recommendation. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. ACM, 275–279.

[3] Nuray M Aykin and Turgut Aykin. 1991. Individual differences in human-computer interaction. *Computers & industrial engineering* 20, 3 (1991), 373–379.

[4] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In *Proc. RecSys'12*. ACM, 35–42.

[5] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2013. LinkedVis: exploring social and semantic career recommendations. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 107–116.

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[7] Simon Bruns, André Calero Valdez, Christoph Greven, Martina Ziefle, and Ulrik Schroeder. 2015. What should i read next? a personalized visual publication recommender system. In *International Conference on Human Interface and the Management of Information*. Springer, 89–100.

[8] Peter Brusilovsky and Eva Millán. 2007. User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web*. Springer, 3–53.

[9] Giuseppe Carenini, Cristina Conati, Enamul Hoque, Ben Steichen, Dereck Toker, and James Enns. 2014. Highlighting interventions and user differences: informing adaptive information visualization support. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1835–1844.

[10] Zohreh Dehghani Champiri, Seyed Reza Shahamiri, and Siti Salwah Binti Salim. 2015. A systematic review of scholar context-aware recommender systems. *Expert Systems with Applications* 42, 3 (2015), 1743–1758.

[11] He Chen, 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.

[12] Yu Chen and Pearl Pu. 2013. Cofeel: Using emotions to enhance social interaction in group recommender systems. In *Alpine Rendez-Vous (ARV) 2013 Workshop on Tools and Technology for Emotion-Awareness in Computer Mediated Collaboration and Learning*.

[13] Cristina Conati, Giuseppe Carenini, Enamul Hoque, Ben Steichen, and Dereck Toker. 2014. Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. In *Computer Graphics Forum*, Vol. 33. Wiley Online Library, 371–380.

[14] Gitta O Domik and Bernd Gutkauf. 1994. User modeling for adaptive visualization systems. In *Visualization, 1994., Visualization'94, Proceedings., IEEE Conference on*. IEEE, 217–223.

[15] Michael D Ekstrand, Daniel Kluver, F Maxwell Harper, and Joseph A Konstan. 2015. Letting Users Choose Recommender Algorithms: An Experimental Study. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*. ACM, New York, NY, USA, 11–18. https://doi.org/10.1145/2792838.2800195

[16] Jill Freyne, Rosta Farzan, Peter Brusilovsky, Barry Smyth, and Maurice Coyle. 2007. Collecting community wisdom: integrating social search & social navigation. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 52–61.

[17] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. 2007. User profiles for personalized information access. In *The adaptive web*. Springer, 54–89.

[18] Vladimir Geroimenko and Chaomei Chen. 2006. *Visualizing the semantic web: XML-based internet and information visualization*. Springer, London.

[19] David M Greenberg, Michal Kosinski, David J Stillwell, Brian L Monteiro, Daniel J Levitin, and Peter J Rentfrow. 2016. The Song Is You: Preferences for Musical Attribute Dimensions Reflect Personality. *Social Psychological and Personality Science* 7, 6 (2016), 597–605. https://doi.org/10.1177/1948550616641473

[20] Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Christopher Hall, and Tobias Höllerer. 2010. Smallworlds: visualizing social recommendations. In *Computer Graphics Forum*, Vol. 29. Wiley Online Library, 833–842.

[21] Ruth Herbert. 2013. *Everyday music listening: Absorption, dissociation and trancing*. Ashgate Publishing, Ltd.

[22] 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.

[23] Yoshinori Hijikata, Yuki Kai, and Shogo Nishida. 2012. The Relation Between User Intervention and User Satisfaction for Information Recommendation. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC '12)*. ACM, New York, NY, USA, 2002–2007. https://doi.org/10.1145/2245276.2232109

[24] Satoru Inoue, Hisae Aoyama, and Keiichi Nakata. 2011. Cognitive analysis for knowledge modeling in air traffic control work. In *International Conference on Human-Computer Interaction*. Springer, 341–350.

[25] Yucheng Jin, Bruno Cardoso, and Katrien Verbert. 2017. How do different levels of user control affect cognitive load and acceptance of recommendations?. In *Proceedings of the 4th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2017)*. CEUR-WS, 35–42.

[26] Antti Kangasrääsiö, Dorota Glowacka, and Samuel Kaski. 2015. Improving controllability and predictability of interactive recommendation interfaces for exploratory search. In *Proceedings of the 20th international conference on intelligent user interfaces*. ACM, 247–251.

[27] Roy PC Kessels, Martine JE Van Zandvoort, Albert Postma, L Jaap Kappelle, and Edward HF De Haan. 2000. The Corsi block-tapping task: standardization and normative data. *Applied neuropsychology* 7, 4 (2000), 252–258.

[28] Bart P Knijnenburg, Niels J M Reijmer, and Martijn C Willemsen. 2011. Each to His Own: How Different Users Call for Different Interaction Methods in Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, New York, NY, USA, 141–148. https://doi.org/10.1145/2043932.2043960

[29] Zacharias Lekkas, Nikos Tsianos, Panagiotis Germanakos, Constantinos Mourlas, and George Samaras. 2011. The effects of personality type in user-centered appraisal systems. In *International Conference on Human-Computer Interaction*. Springer, 388–396.

[30] Benedikt Loepp, Katja Herrmanny, and Jürgen Ziegler. 2015. Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 975–984.

[31] 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.

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

[33] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proc. RecSys'11*. ACM, 157–164.

[34] James Schaffer, Tobias Höllerer, and John O'Donovan. 2015. Hypothetical Recommendation: A Study of Interactive Profile Manipulation Behavior for Recommender Systems.. In *FLAIRS Conference*. 507–512.

[35] Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*. ACM, 830–831.

[36] Barry Smyth, Evelyn Balfe, Jill Freyne, Peter Briggs, Maurice Coyle, and Oisin Boydell. 2004. Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction* 14, 5 (2004), 383–423.

[37] Dereck Toker, Cristina Conati, Giuseppe Carenini, and Mona Haraty. 2012. Towards adaptive information visualization: on the influence of user characteristics. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 274–285.

[38] John Venn. 1880. I. On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 10, 59 (1880), 1–18.

[39] Katrien Verbert, Denis Parra, and Peter Brusilovsky. 2016. Agents Vs. Users: Visual Recommendation of Research Talks with Multiple Dimension of Relevance. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 2 (2016), 11.

[40] 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.

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

[42] Michail Vlachos and Daniel Svonava. 2012. Graph embeddings for movie visualization and recommendation. In *First International Workshop on Recommendation Technologies for Lifestyle Change (LIFESTYLE 2012)*. 56.

[43] David Wong, Siamak Faridani, Ephrat Bitton, Björn Hartmann, and Ken Goldberg. 2011. The diversity donut: enabling participant control over the diversity of recommended responses. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1471–1476.

[44] Xiangmin Zhang and Mark Chignell. 2001. Assessment of the effects of user characteristics on mental models of information retrieval systems. *Journal of the Association for Information Science and Technology* 52, 6 (2001), 445–459.