Controlling Spotify Recommendations: Replication Study

Joseph Grendzinski
grend016@umn.edu

Abstract

This paper lays out the replication of a study previously conducted. This study was focused on the "black box" nature of today's recommendation systems in which users may show a lack of trust due to a lack of transparency within each system. The study set out to provide users with more control during the recommendation process in an attempt to help them steer their results to better fit their taste. The study also attempts to decipher which personal characteristics have the greatest impact on one's perception of visual techniques. Participants are asked to create a playlist using two different interfaces that allow the adjustment of certain musical attributes to help steer recommendations. It is found that personal characteristics do not have a significant impact on the use of each interface.The replication of the study reflected this. It was also found in both studies that participants preferred the radar chart for adjusting each musical attribute (as opposed to the slider chart).

Introduction

As technology has advanced in that last decade or so, a lot of services have moved partially or completely online. However, without somebody behind the counter to assist you, it can be hard for some consumers to find exactly what they're looking for. To address this problem, many of the services available online utilise some sort of recommendation system to suggest more personalized content to their users. Well-known companies such as Amazon, Spotify, YouTube, and Netflix make use of such systems to offer their users much more relevant items.

Recommendation systems rely on information that is received from the user. This can be achieved by simply asking users for their preferences or by analyzing their interactions with a service or a system. This information can be extremely useful if properly taken advantage of. Still, there are a couple of issues regarding recommendation systems. One of which is the lack of "transparency" such systems provide. While recommendations are constantly being provided to users, it is rare to include a proper explanation of *how* the conclusion was reached. Because of this, the process of recommendation has become a "black box," leading to mistrust between the company and its users.

One study sought to give users more control over recommendations to better understand the effect of personal characteristics on the perception of visual control techniques in the music domain. The study attempts to answer the following four research questions:

- In what way do personal characteristics influence perception of the visual control techniques in music recommendation? (RQ1)

- In what way do personal characteristics influence interaction with the visual control techniques in music recommendation? (RQ2)
- Which visual control technique(s) is/are better suited for users to manipulate music recommendations? (RQ3)
- How can we design an interface to allow better user control for music recommendation? (RQ4)

In an attempt to replicate this study, I recreated both interfaces used in order to conduct the experiment on a different set of participants. The replication of this study serves to both:

- Support/oppose the results
- Critique how difficult/easy it was to successfully carry out the experiment with the details provided in the writeup

Background

The original study measured *level of experience* as musical sophistication, the number of hours each participant listens to Spotify (or any music service), and tech-savviness; *demographic characteristics* as age, 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). Visual working memory was measured by a block-tapping test based on Corsi block-tapping. Tech-savviness was simply the score that participants rated themselves when it comes to trying new technology (on a scale of *confident*/*not confident*/*somewhere in-between*).

Experimental Design

Participants

Both the original study and the replication involved participants recruited from personal contacts. The original study had a total of 40 participants complete the inquiry. The replication was conducted with only 20 participants. The personal characteristics of the two groups are shown in Figure 1. For both visual working memory and musical sophistication, the two groups were split at the 50th percentile into “high” and “low” MS or VWM.

Implementation

The authors of the original study found that certain visualization techniques can be used as a basis to improve recommendations or even allow the user to explore the recommendation space. The techniques chosen to pursue were sliders and a radar chart. Both with the ability to drag and drop to adjust different values. Each interface was designed with a three column layout. The first column allows users to select up to five specific artists to be used for generating recommendations. The second column contains the visualization technique currently in use. The visualization lets users adjust the weights of certain musical attributes that can be used as parameters to tweak recommendations. The third column contains the generated recommendations that each user can either like, dislike, or preview.

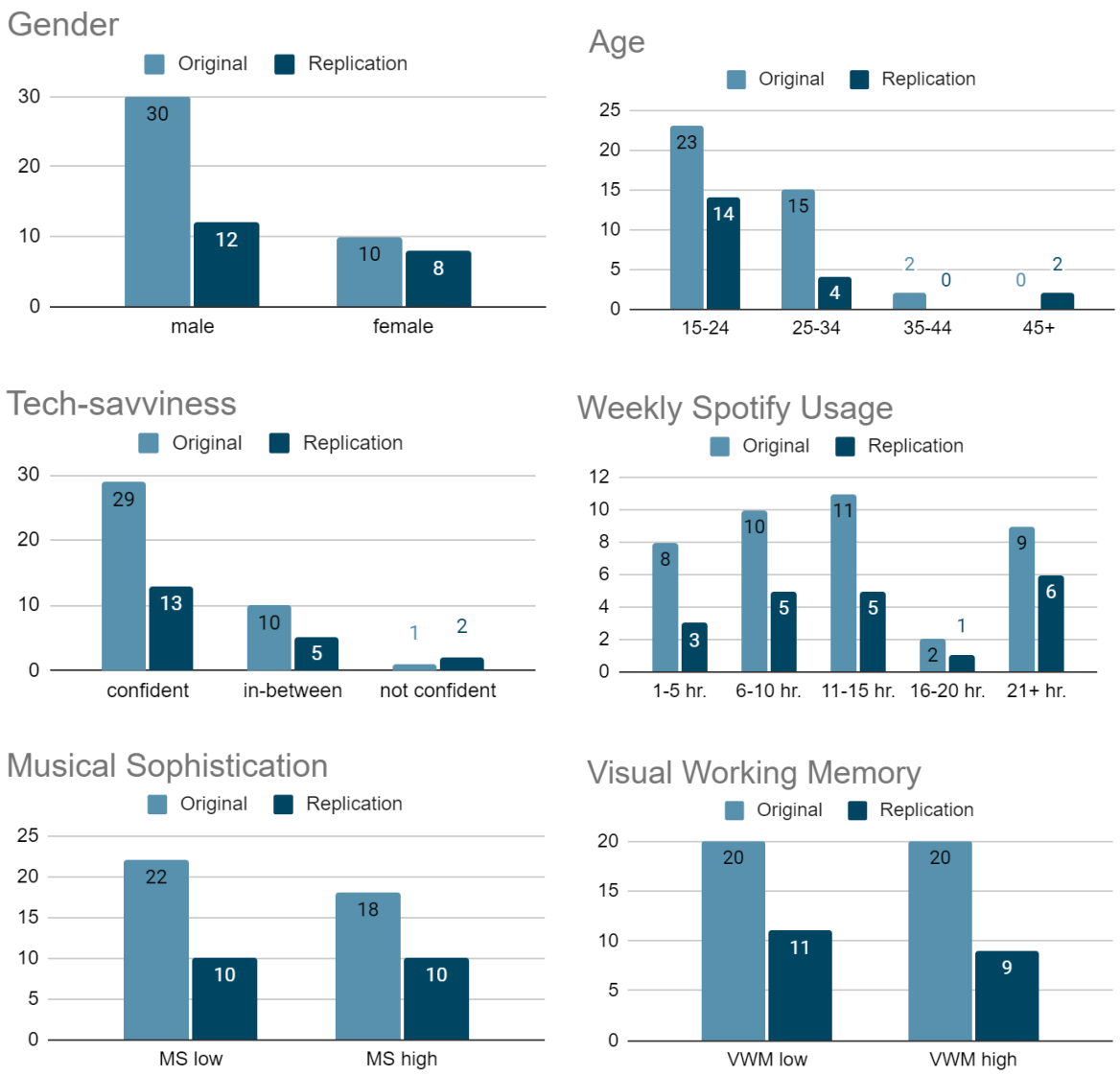

Figure 1: Personal characteristics of participants

The actual recommendation is achieved through the use of the Spotify API. The API allows you to generate recommendations based on up to 5 artists. Along with the artists, the API also lets you modify the weights of 14 different musical attributes. The original study explains that different song genres can be represented by only three categories: arousal (intensity and energy of the music), valence (spectrum of emotions), and depth (intellect and sophistication). Because of this, the authors chose these five musical attributes of the fourteen available: energy, danceability, valence, instrumentalness, and acousticness.

Each song is presented as its title and artist. The study makes use of like and dislike buttons in order to either dismiss or keep a song in the list. The replication uses an “X” button for dislikes and a song toggle for likes. Also, the Spotify API gives access to 30 second previews of each song. The “play” button in each interface launches this preview in the user’s browser. The

main difference between the two interfaces displayed in Figure 2 is the “Search for artist” button present in the replication. This button was added in order to avoid complications with retrieving one’s Spotify information. The original study would automatically generate the artists in column one by retrieving each participant's top Spotify artists while the replication would require the artists to be put in manually. Although this can be tedious for participants, it ultimately gives them more control over the system and allows those that do not use Spotify to participate in the study. This also gives the participants an opportunity to diversify their results by utilizing a greater variety of artists.

Study Procedure

The original study was conducted online while the replication was conducted in-person (interface had to be run locally). Both involved participants recruited from personal contacts. The original study began by prompting participants to authorize their Spotify account. By implementing the “Search for artist” button, this step was avoided in the replication. Next, each participant was required to answer demographic questions to collect their personal characteristics (age/gender/music sophistication/visual working memory/tech-savviness/Spotify usage). In addition, participants were also asked to rate their familiarity with recommender systems and their attitude towards them.

Once the demographic questions were answered, each participant was given instructions on how to launch/use the interface to create a playlist consisting of 9 songs to listen to when travelling or to listen to during personal maintenance (each paired with one of the two visualization techniques). The original authors chose these prompts since travelling and personal maintenance are the two most common activities associated with listening to music. After the current user had 9 songs liked, they were presented with 13 questions designed to rate the system’s effectiveness.

Once the questions were completed, the participants were to repeat the task with a different visualization technique and prompt. Again, the same 13 questions were presented for the second interface. Following the completion of both playlists, each participant was presented a set of exit questions. These questions were designed to help better improve the visualization techniques/effectiveness of the system. To avoid fatigue, the order in which the interfaces were presented was rotated using a *Latin Square* counterbalancing measure.

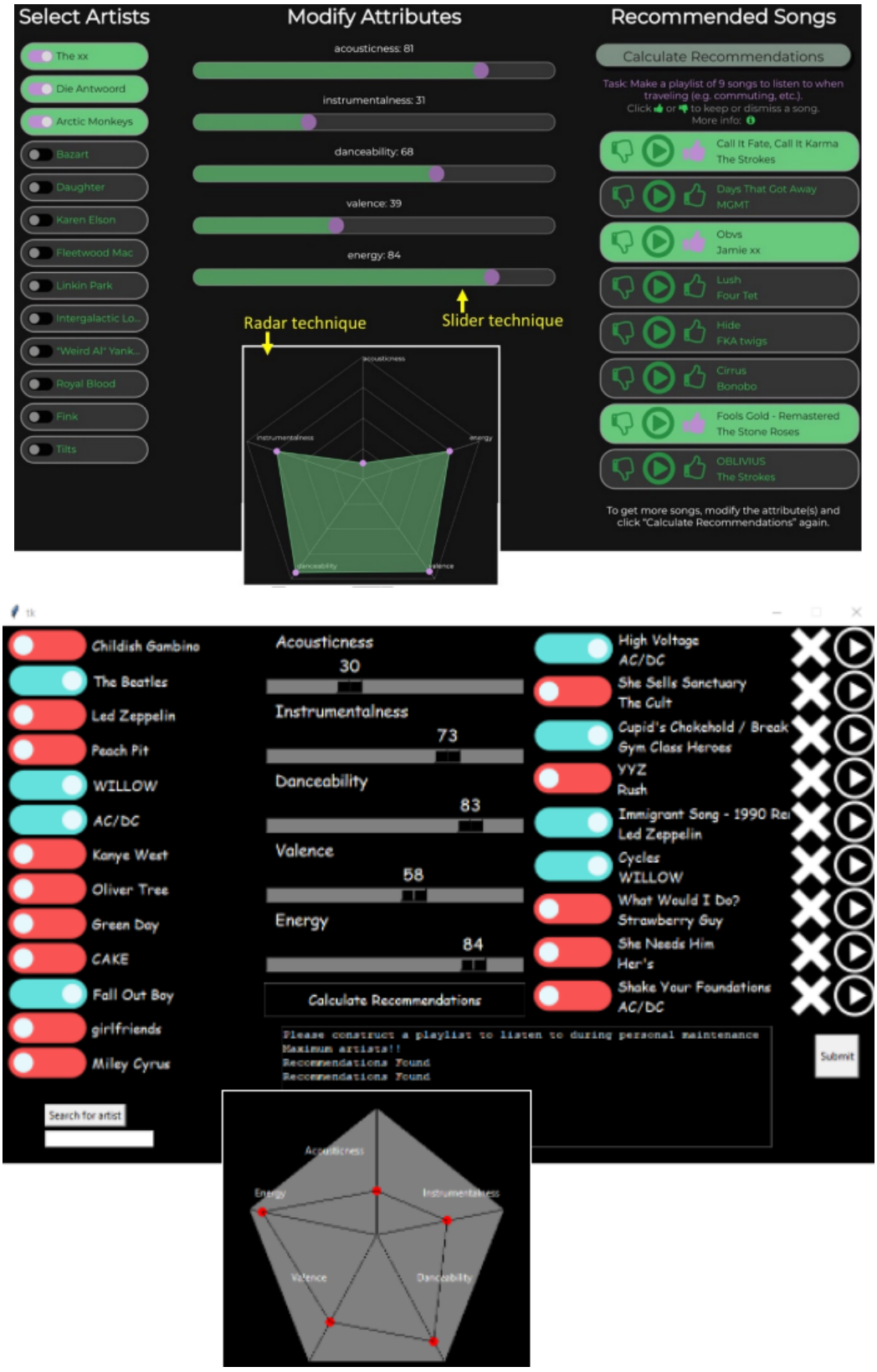

Figure 2: Interface variations used in each study (original = top, replication = bottom)

Evaluation Metrics

System Evaluation Questions (7-point Likert scale; 1 = strongly disagree, 7 = strongly agree):

- 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)

Interaction Log:

- The number of times the musical attributes were changed (attribute)
- The number of times any given musical attribute was changed (nb + attribute name)
- 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)

  Note: the original study receives these values directly from the interface while the replication requires each participant to manually input them. After each playlist is finished, users must select the "submit" button to receive the log values.

Exit Questions

In this section, participants were first presented with all 14 of the available musical attributes, along with their definitions, through the Spotify API. They were then asked to rate each attribute on a 7-point Likert scale. Afterwards, participants were asked to suggest any other visualization techniques they think may be useful to improve the recommendation process.

Results

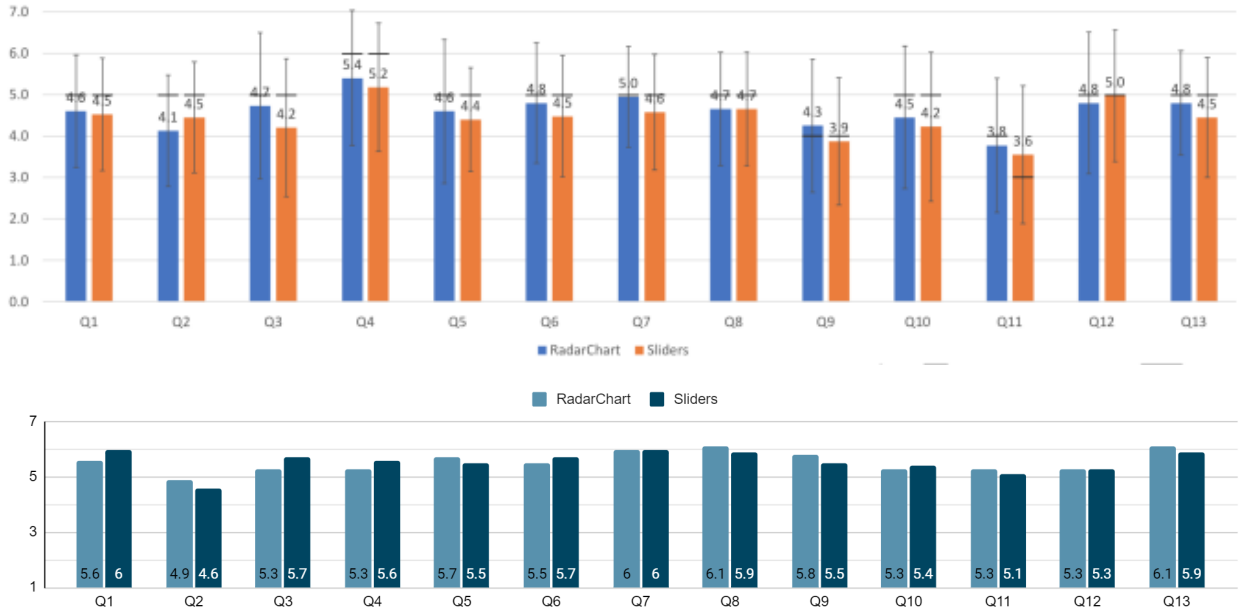

Figure 3: Results of evaluation questions (original = top, replication = bottom)

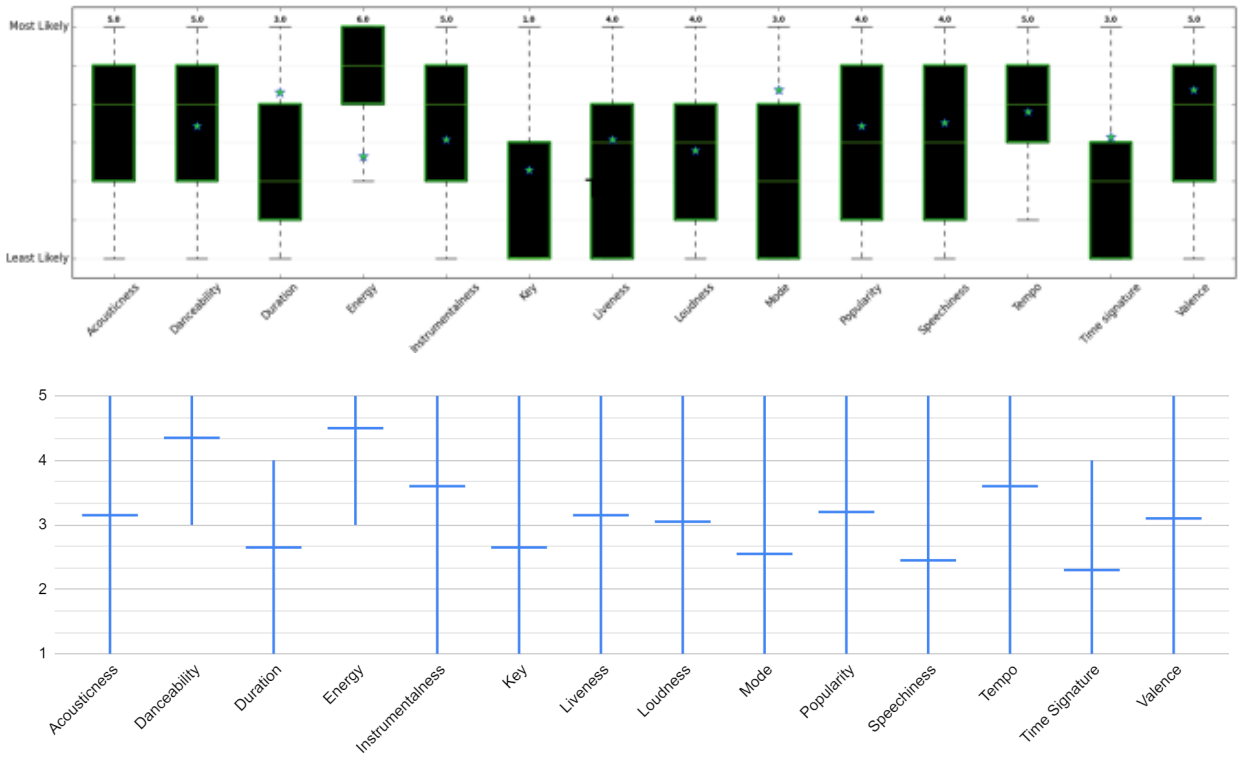

Figure 4: Attitude towards 14 musical attributes (original = top, replication = bottom)
Top: green star represents mean while wicks indicate low and high values
Bottom: horizontal bar represents mean while wicks indicate low and high values

User Perception and Log Analysis

Comparisons without Personal Characteristics

First, results of each study were examined without taking into consideration the personal characteristics of each participant in order to better understand the overall differences in the two interfaces. Results of the original study suggest that the two different interfaces show the most significant difference in their perceived discovery of new songs. The original chart from Figure 3 displays this difference in Q3. The results from the replication also show a large gap between the scores of Q3. However, the replication shows a flipped score for each interface. Rather than one interface greatly outperforming the other, this implies that Q3 may be subject to more uncertainty.

The results of the original study also display Q4 as the highest overall rated question. It's no surprise that having more control over the recommendation process allowed participants to discover songs that may have otherwise been ignored by themselves or the system. Q4 did not stand out as much during the replication of the study. In the replication, the questions Q7, Q8, and Q13 are rated the highest overall. All three of these questions involve the user's overall satisfaction with the system. Q13 in particular reinforces the idea of Q4 that participants did in fact discover songs that they did not already have in their library.

Comparisons by Personal Characteristics

To analyze the effect of personal characteristics, the participants were divided based on each category of their personal characteristics. This was done during the replication as well. The characteristics that proved to have a significant impact on the usage and perception between the two interfaces are as follows.

Tech-savviness: Results from the original study state that those who rated themselves "in-between" for tech-savviness indicated that they found more ideal songs with the radar chart over the slider chart (Q5). Those in the other two categories did not seem to show a preference for interface when it came to finding ideal songs. The results from the replication do not show a significant difference for Q5 when considering the same group of participants. Rather, results from the replication indicate that those who rated themselves "confident," interacted with both interfaces significantly more than those who rated themselves "in-between" or "not confident."

Spotify Usage: The original study explains that those who use Spotify more than 21 hours a week interacted with the musical attributes significantly higher within the radar chart. In addition to this, those in the same category would tend to use the instrumentalness attribute significantly more with the radar chart. The results from the replication also relay this with both the slider and radar chart. Those with greater Spotify usage on average would report more interactions with each attribute, not just instrumentalness. Similar to those that are more tech-savvy, those with greater Spotify usage ultimately interact more with the attributes.

Musical Sophistication (MS): As explained earlier, based on their MS, participants were divided into high and low MS groups. The original study reports much higher overall interactions with the radar chart interface. The replication did not reflect these results. MS did not seem to have a great effect on overall interactions. MS did seem to have a direct correlation

with the use of the preview button. Those with a high MS interacted with the preview button significantly more than those with a low MS.

Visual Working Memory (VWM): To test the visual working memory of each participant, each study made use of the Corsi block-tapping test. The original study does not specify how this test is conducted. For the replication, participants were to complete a Corsi block-tapping of span 3 (in which the participants must remember the order and placement of images flashing in a 4x4 grid, a span of 3 would flash 3 images). If the current participant were to correctly pass at a span of 3, they would then increase the span by one until they fail. The original study shows that those with a high VWM had significantly more interactions with the danceability attribute within the radar chart interface. Those with a VWM had significantly higher interactions with the acousticness attribute within the slider chart interface. During the replication of the study, those categorized with a high VWM had significantly more interactions with the slider chart than the radar chart. None of the attributes stood out as being adjusted much more than the others for this group of participants. This could mean that those with a lower VWM may prefer the radar chart over sliders while those with a higher VWM do not need the extra visualization.

Design Feedback

The suggestions received for improving the system were divided into three main themes by the authors of the original study. Each is presented in detail in the following sections.

Track-Attribute Visualization: Participants of the original study expressed a desire for a way to better understand how each attribute is attributed to the recommended songs. Users can benefit greatly from this as it may allow them to better understand why each song is being recommended. Furthermore, visualizing songs with their track-attributes can help users better understand their music taste and information about their favorite songs. Such responses were not reflected in the replication; however this is most likely due to the smaller sample size lacking diversity in their suggestions.

Relevance feedback: A handful of participants expressed some concern about the musical attributes. For those that are unfamiliar with the attribute terms or just unsure how to utilize them, it may be easier to simply find songs that are similar to each other. In this case, the recommended songs would display their relationship with the input song in the musical attribute spectrum. This concern was present in the replication study as well. Participants explained that a glossary of some kind would help in their understanding and usage of the musical attributes presented.

Usability: The majority of participants displayed a positive attitude towards having the ability to steer recommendations. In both the original and its replication, nearly all participants preferred the radar chart over the sliders. Those that preferred the radar chart explained that it provides a greater overview of the current settings. One concern that was much more prevalent during the replication was the desire for more color. Seven out of the twenty participants mentioned that the interface could be more visually appealing.

Discussion

When considering the first research question (RQ1: In what way do personal characteristics influence perception of the visual control techniques in music recommendations?), it was found that personal characteristics in general do not play a significant role in the perception of visual control techniques. Although the original study found some correlation between tech-savviness and the preference of interface, this correlation is not very strong and was not reflected in the replication. No strong conclusions could be drawn about personal characteristics and their effect on perception of visual control techniques.

As for the second research question (RQ2: In what way do personal characteristics influence interaction with the visual control techniques in music recommendation?), it was found that Spotify usage, musical sophistication, and visual working memory had some influence on interactions. In both studies, it was obvious that Spotify usage had a strong correlation with the total number of interactions on each interface. Those that use Spotify more than 21 hours per week showed significantly more interactions than those below. As for musical sophistication, the replication suggests that those with a higher MS reported much more interactions with the preview button. Finally, visual working memory seemed to have an effect on the interactions recorded between both interfaces. Those with a higher VWM showed far more interactions with the slider chart rather than the radar chart. As explained before, those with a higher VWM may not need the extra visualization that is provided through the radar chart.

In the context of the study's third research question (RQ3: Which visual control techniques are better suited for users to manipulate music recommendations?), the radar chart proved much more effective in both cases. 10 of the participants from the original study reported that they preferred the radar chart over the slider while only 3 reported a preference for the sliders. As for the replication, 12 reported a preference for the radar chart while only 1 showed a preference for the sliders.

In the context of the fourth research question (RQ4: How can we design an interface to allow better user control for music recommendations?), it was found that providing a visualization of the relationship between recommended tracks and musical attributes could be greatly beneficial to users. This in turn can help users better understand why they are receiving specific recommendations and even help them better understand how their music taste is presented in the recommendation space. Those that participated in the replication mainly posed concerns with the appearance of the interface rather than its functionality.

It is also important to consider the participants' opinions on each of the 14 musical attributes available through the Spotify API. Figure 4 contains the results from both the original study and its replication. It is obvious that the most popular attribute across both studies was energy. The next most popular were danceability, instrumentalness, tempo, popularity, and valence. Therefore, it could be useful to include attributes such as tempo and popularity to help users better explore the recommendation space and find songs more tailored to their taste.

Conclusion

In this paper, a previously conducted experiment was replicated in which two different visual control techniques were evaluated in their effectiveness of steering the music recommendation process. The original interfaces were imitated in order to conduct the study on a different set of participants and compare results. The Spotify API was employed to generate song recommendations using 1-5 seed artists and the weights of five different musical attributes. The visual control techniques allowed each user to manipulate the musical attributes. Evaluation questions were presented to each participant in order to compare the effectiveness of each interface. The results from both the original study and its replication show a significant preference for the radar chart over the slider chart. It was concluded that personal characteristics do not play a significant role on the perception towards different visual control techniques. It was found that the radar chart encouraged more interactions from those who had a high Spotify usage, musical sophistication, and visual working memory. Replication of this study was fairly straightforward once the replication of the two interfaces were recreated. The evaluation metrics were clearly laid out and the experimental process was explained in detail. The replication reinforced the conclusion that personal characteristics do not have a direct correlation to one's perception of visual techniques. In addition to this, the replication also supported the claim that the radar chart was overall more effective than the slider chart. Future work would require a revamping of the replicated interfaces and perhaps include other musical attributes such as popularity or tempo along with the other five.

References

[1] Millecamp, M., Htun, N. N., Jin, Y., & Verbert, K. (2018, July). Controlling Spotify recommendations: effects of personal characteristics on music recommender user Interfaces. In Proceedings of the 26th Conference on user modeling, adaptation and personalization (pp. 101-109).

[2] Corsi block tapping test. MemoryHealthCheck. (2021, December 5). Retrieved January 14, 2022, from https://www.memorylosstest.com/corsi-block-tapping-test/

[3] Welcome to spotipy!. Welcome to Spotipy! - spotipy 2.0 documentation. (n.d.). Retrieved January 14, 2022, from https://spotipy.readthedocs.io/en/2.19.0/