BACHELOR

Predicting Factors for Exploring Music out of Someone's Comfort Zone and the Experience of Music Exploration

Martens, Jolijn G.M.J.

Award date:
2021

Awarding institution:
Tilburg University
Jheronimus Academy of Data Science

Bachelor End Project
Joint Bachelor Data Science

Predicting Factors for Exploring Music out of Someone's Comfort Zone and the Experience of Music Exploration

January 15, 2021

Jolijn Gert Marieke Janine Martens, 1287982
j.g.m.j.martens@student.tue.nl
Eindhoven University of Technology
Department of Mathematics and Computer Science
5612 AZ Eindhoven, The Netherlands

Supervisor: M.C. Willemsen
2nd Supervisor: Y. Liang
Department of Human Technology Interaction

Eindhoven, January 15, 2021

Abstract

Recommender systems provide support when coping with the problem of information overload. Most recommender systems assume that someone's preference will remain the same over years, which does not have to be the case. This study focuses on novel music recommendation and more specifically, what factors make people explore music that is out of their comfort zone. Further, the experience of music exploration is analyzed. An experiment was conducted in which people were given the choice between two concerts: one that laid within their comfort zone and one that was out of their comfort zone. Based on the data that was collected, a logistic regression model was built to predict whether users were open to music exploration. In addition, a multilevel model was created to predict how users experience music exploration. It can be concluded that people are susceptible to nudging to explore music out of their comfort zone. In addition, people tend to have a more positive experience with exploring music when they are aware of the fact that they are exploring music. Finally, people tend to enjoy music more when it is more familiar to them.

1. Introduction

Over the past decade, we have witnessed a rapid growth of digital technologies. More information has become available and there are endless options to get access to information relatively easily. As a consequence, there is a need to help people cope with the problem of information overload. Recommender systems have become a popular technique to prune large information spaces so that users are directed toward those items that best meet their needs and preferences. In their daily lives, people are using recommender systems more often than they are probably aware of. Famous examples of recommender systems are Netflix (which movie will this specific user also like?), Amazon (which product fits a user's preference?) and Tinder (which person is likely to be a match?). This research will focus on music recommendation, and more specifically, novel music recommendation.

Imagine you are using the Spotify recommender system, but you decided you want to explore music that you have not listened to ever in the past, then the recommender system should try to recommend music to you that is out of your comfort zone. In this study, it is investigated what factors make people explore music that is out of their comfort zone. In addition, it is investigated how this music exploration is experienced. This leads to the following research question (RQ):

"*What influences people to explore music out of their comfort zone and how is this music exploration experienced?*"

This research question will be answered by conducting an experiment in which people are given the choice between two concerts: one that lies within their comfort zone and one that lies out of their comfort zone. This study is split into two parts. The first part investigates different factors that could influence whether people are likely to explore music out of their comfort zone, where the main focus will be laid on the effect of making a recommendation list the default option. The main purpose of the default option is to nudge people to explore music that is out of their comfort zone. So, it is researched whether preselecting the stream that is outside their comfort zone increases the probability of choosing to explore music. Further, is investigated whether people are susceptible to nudging is dependent on their Musical Sophistication index for Active Engagement (MSAE score), their age, the difference in recommendation score between streams and gender. Users with a higher MSAE score have more experience with music and therefore a more stable music preference (Müllensiefen et al., 2014). Hence, it could be the case that people with a higher MSAE score might be less susceptible to nudging to explore music as they are

in a more informed position to make the decision to explore music, despite the fact that they are being nudged or not.

The second part of this study investigates how music exploration is experienced. It is investigated whether the participants' rating for how much they enjoyed the performance is dependent on factors like the rank of the track, their MSAE score, whether or not they chose to explore music out of their comfort zone, gender and their familiarity with a song. The main purpose is to investigate whether people tend to enjoy exploring music more when they are aware of the fact that they are exploring music. If this awareness of music exploration matters to users, then this could be implemented in music recommender systems to serve the needs of user even better.

The remainder of this paper is structured as follows. Firstly, the relevant literature will be discussed and the hypotheses will be stated. Secondly, the design of the experiment is explained. The third section presents the proposed methodology regarding the data analysis. Then, the results will be discussed and evaluated. Afterwards, limitations of this research will be discussed and suggestions for future research will be presented. Finally, a conclusion will be drawn.

1.1. Recommender systems

Collaborative filtering is a widely used method to design a recommender system. Collaborative filtering can exist in many forms, but the most common one is user-based collaborative filtering. The general principle can be explained in two steps. First, the system looks for users who share the same rating patterns with the active user, which is the user whom the prediction is for. Secondly, the ratings from those users who share the same rating pattern are used to calculate the prediction for the active user. Another form of collaborative filtering that is used frequently is item-based collaborative filtering. As the name suggests, this approach proceeds in an item-centric manner. In item-based collaborative filtering, an item-item matrix is built which determines the relationships between pairs of items. Secondly, the taste of the user in question is inferred by examining the item-item matrix and matching that user's data. Despite the wide use of collaborative filtering, it has some limitations, like user bias (some users give higher ratings than others) and item bias (tendency for some items to receive higher ratings than others). Within collaborative filtering, these issues are already solved by modelling by the use of matrix factorization. As described by Koren et al. (2009), matrix factorization in its basic form characterizes both items and users by vectors of factors inferred from item rating patterns. A recommendation is generated when there is a high correspondence between an item and the user factors. Unfortunately, collaborative filtering has more limitations, like non-association and cold start problems (Li, Myaeng, & Kim, 2007). Non-association means that the relationship between two similar items cannot be known explicitely if they have never been wanted by the same user. The cold start problem is the challenge to recommend items to users that have no history. Furthermore, traditional collaborative filtering (collaborative filtering without considering user preference change) assumes that someone's taste will stay the same over the years, which does not have to be the case. The importance of modeling temporal dynamics is key for designing proper (music) recommender systems. Ko-

ren (2009) has implemented the temporal dynamics successfully and concluded that the inclusion of temporal dynamics proved very useful in improving quality of predictions. Besides the fact that your taste is constantly changing as the years pass by, users might want to explore new content that can be out of their comfort zone. In fact, studies have shown that there is a natural drive in humans seeking for novelty and change (McAlister & Pessemier, 1982). It is also indicated that there is a spontaneous devaluation in user preference on music listening meaning that users would sometimes become bored with their current preference and willing to seek for novel content (Kapoor, Kumar, Terveen, Konstan, & Schrater, 2015; Kapoor, Srivastava, Srivastava, & Schrater, 2013).

So, most recommender systems are based on music that a person had listened to and liked in the past, but what if someone wanted to explore new music? There is only few research being conducted on novel music recommendation. For example, Taramigkou et al. (2013) presented an approach and related experimental application that allows users to escape their filter bubble and explore the preferences of others types of music. Further, Zhang et al. (2012) introduced the Auralist recommendation framework, which is a system that explicitly balances the conflicting goals of accuracy, diversity, novelty and serendipity.

This research builds upon the work of Liang & Willemsen (2019), which conducted an experiment to let people explore a new genre, and investigated which recommendation method would be most helpful for users to exploring this new genre. Three different recommendation methods were tested: the non-personalized method, the personalized method and the hybrid method. The non-personalized method recommended the most representative tracks of the genre to be explored. The personalized method recommended songs from the new genre that best fitted the users' current music preferences. The hybrid method made a tradeoff between the non-personalized and the personalized method. This study concluded that the mixed method could be a good way to balance the perceived accuracy and representatives to increase the perceived helpfulness for new genre exploration. In order to predict which songs a user is going to like most, Liang & Willemsen (2019) built multiple Gaussian Mixture Models (GMMs), which will be explained more in detail in Section 1.1.2.

In addition, this research will briefly touch upon the subject of the importance of visualizations to convey a story. Visualizations aim to reveal to users those regions of their recommendations space that are unknown to them, i.e, their blind spots (Tintarev et al., 2018). By the use of understandable visualizations, users get a better understanding of their profile, which helps them to identify their consumption blind spots (Kumar & Tintarev, 2018). This research will build upon the idea of visualizing the recommended music as proposed by Meeuwisse (2019).

1.1.2. Gaussian Mixture Models

In order to predict which song a user is going to like most, this research used multiple one-dimensional Gaussian Mixture Models as already used in Liang & Willemsen (2019). A GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities (Reynolds, 2009). GMMs are initialized

using K-means clustering and the parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm (Bishop, 2006). One important characteristic of GMMs is that the model is able to generate an output of a probability of how much a data point is associated with a specific cluster. This is therefore also the main difference with hard clustering methods like K-means clustering, as K-means clustering will associate each data point to one and only one cluster (Carrasco, 2020).

For the scope of this research, the focus is on GMMs in music recommendation. Like in the research of Liang & Willemsen (2019), the assumption needs to be drawn that top tracks of the user's Spotify profile represents the user's musical preference well. If this assumption holds true, the GMM is trained on the audio features from the user's top tracks that are extracted from the user's Spotify profile, and results in a representation of the user's preference as a probability density function. As six different audio features are taken into account (accousticness, danceability, energy, liveness, valence and speechiness), multiple one-dimensional GMMs are needed. Each GMM is initialized by K-means and trained with the EM algorithm. The ranking of the track is obtained in each feature dimension. The multiple GMMs together provide a personalized recommendation score per track, together with the position of track i in the ranked list of all tracks. For this research, the same assumption is made as stated above. Furthermore, this research used an adjusted version of the recommendation algorithm, as this research obtained a personalized recommendation score per session, instead of per track, by taking the average of all tracks in the session.

1.2. The effect of nudging

Leonard (2008) defined a nudge as the following:

> A nudge, as we will use the term, is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting fruit at eye level counts as a nudge. Banning junk food does not. (p.6)

Thaler (2015) described three principles that guide the use of nudges. Firstly, a nudge should be transparent and never misleading. Secondly, it should be as easy as possible to opt out of the nudge, preferably with as little as one mouse click. Thirdly, there should be a good reason to believe that the behavior that is encouraged will improve the welfare of those being nudged. There is a fine line between nudging and manipulation. Manipulation is defined as a form of influence that is neither coercion nor rational persuasion (Noggle, 2018). The main difference between nudging and manipulation is that manipulation may not be in the best interest of the individual, whereas nudging should be. Nudges are currently implemented in many fields in society: governance, business and healthcare. A famous example is the printing of a housefly on the inside of urinals to improve bathroom cleanliness (Bikker, 2020).

As a consequence of the rapid growth of digital technologies, the concept of Digital Nudging was introduced. Digital Nudging is

the use of user-interface design elements to guide people's behavior in digital choice environments (Weinmann et al., 2016). More and more frequently, choices are being made in a digital environment. The design of a digital environment always (either deliberately or accidentally) influences people's choices and nudges them to make a particular decision. In fact, it is impossible to present choices in a neutral way. The key of implementing a nudge is that it should help the person being nudged to make better choices, so the interest of the designer of the digital environment is not important and should not be taken into account.

Jesse & Jannach (2020) examined the relationship between Digitial Nudging and recommender systems, in which 87 nudging mechanisms were identified and categorized in a novel taxomony. Figure 1 shows the different categories of the proposed taxonomy. Four categories are identified: Decision Information, Decision Structure, Decision Assistance and Decision Affection. Firstly, the category Decision Information is based on changing the information that is shown to the decision maker without changing the options themselves. Examples that can be thought of are how information is phrased on the interface, and how to increase the salience of information. The second category, Decision Structure, focused on the arrangement of options. Examples are setting defaults and changing the order of options. The third category is Decision Assistance, which encompasses mechanisms that support decision makers in accomplishing their goals. An example of Decision Assistance is implementing reminders to nudge people in the right direction. The fourth category is Decision Affection, which focuses on the emotional and social implications of the change. An example is that you show the choices that other users made, and investigate whether this influenced the decision of the user in question.

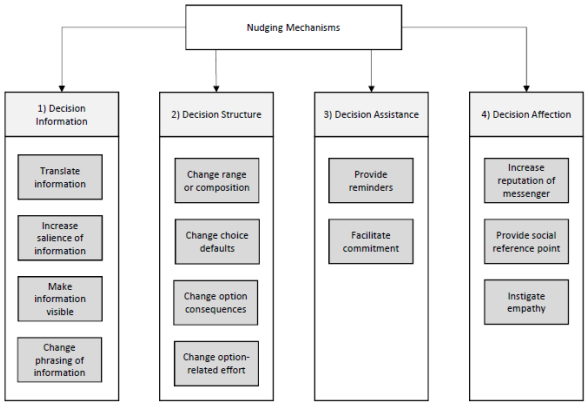

Figure 1: The Taxomony of Nudging Mechanisms (Jesse & Jannach, 2020).

This research will investigate the effect of nudging users to explore music. As far as novel music recommendation is concerned, the nudge should be subtle, as people are exploring their preferences (what kind of music would you might like in the future?). A more direct nudge, like described in the third category of the Taxomony of Nudging Mechanisms (Decision Assistance), is not necessary as the choice to explore music has no consequences for other users or for society in general. Examples of choices that do influence other users and society in general are issues like nudging people for organ donation or the issue to reduce CO2 emissions. By setting a default option, people draw inferences from that default, as they think it is the recommended option (McKenzie et al., 2006). Further, setting defaults is expected to work because of the reduced effort while sticking with the default (Dinner et al., 2011; Thaler & Sunstein, 2009).

Multiple studies have investigated

the effect of nudging mechanisms of setting defaults (Bauer & Schedl, 2017; Bothos et al., 2015, 2016; Lee et al., 2011). This research will investigate this effect in music recommendation, as users will be nudged by preselecting a playlist for them. Therefore, the current research will mostly fall into the second category: Decision Structure.

Furthermore, it is also investigated whether the MSAE score, gender, age and the difference in recommendation score between the streams influence whether users are susceptible to nudging. If the susceptibility of nudging to explore music is based on these factors, recommender systems can be personalized accordingly. It is expected that people with a higher MSAE score are less susceptible to nudging, as the assumption is made that people with a higher MSAE score have a more stable preference, as they have a higher musical expertise (Müllensiefen et al., 2014). Therefore, it is expected that people with a higher MSAE score will be in a more informed position to make the decision to explore music, despite the fact that they are being nudged or not. Further, it could be the case that people are more susceptible to nudging when the difference in preference for the different streams is small, as people are less willing to make a big change at once, but rather want to take small steps (Wendel, 2020).

Limited research is available on nudging and novel music recommendation, and the susceptibility to nudging in this field, as, in existing research, music is most often used as a tool to investigate how music serves as a nudge for healthier choices (Biswas et al., 2016) or for more ethical and peaceful business behavior (Fort, 2018). Therefore, this study can be considered as a first exploratory view of the approach of setting defaults to novel music recommendation.

1.3. Hypotheses

In order to answer the proposed research question, the following hypotheses are formulated:

- **H1:** Making a recommendation list the default will increase the probability of selecting that list.
  - **H1.1:** Whether people are susceptible to nudging is dependent on their MSAE score, age, the difference in recommendation score between streams and gender.
- **H2:** Ratings for music performance will depend on someone's rank of a track, MSAE score, whether they chose to explore music out of their comfort zone, gender and their familiarity with a song.

In order to investigate hypothesis H1, a logistic regression model is built, which investigated the effect of making the recommendation list the default list on exploring music out of someone's comfort zone. Further, the variables MSAE score, age, the difference in recommendation score and gender are included as control variables. By including the interaction effects between those variables and the effect of making a recommendation list the default, hypothesis H1.1 can be answered: whether people are susceptible to nudging is dependent on their MSAE score, age, the difference in recommendation score, and gender.

In order to investigate H2, a multilevel model is built, with as dependent variable the performance rating of a song given by a participant. This performance rating

measured how much a participant enjoyed the song. The variables rank, MSAE score, whether or not they chose to explore music out of their comfort zone, gender and their familiarity with a song, together with their interaction effects, are the independent variables. More detail on the data analysis techniques that are used is provided in Section 3.2.

Furthermore, the importance of visualizations to understand the music recommendation algorithm will be investigated. The following usability question will be answered: Does this visualization help understanding the user's own music taste and the music to be explored? It is hypothesized that visualizations do help the user's to understand their own music taste and the music to be explored. It is investigated if this is the case, together with whether this effect is mediated by variables like age and MSAE score.

2. Experiment Design

The conducted experiment was part of the Den Bosch Data Week[1], and more specifically, took place on October 27, 2020 from 20:00 to 21:00. In the 'Culture Night Music x Data', participants were able to listen to multiple short concerts (15-20 minutes) consisting of 4-6 songs (depending on the duration of the song) of artists performing in his or her own style. The following artists were performing: a harpist, a jazz musician, a singer-songwriter and a pop musician. Participants could choose between two streams each consisting of two artists. So, each participant listened to 8-12 songs in total. In order to help the participants in choosing which stream to listen to, a recommendation algorithm recommended sessions based on their Spotify account. Simultaneously to watching the performances of the artists, paintings were shown and rated, which is outside the scope of this research.

Afterwards, people that were not able to participate on the given date of the experiment had the possibility to participate at a later time. In this way, more data could be collected in order to carry out the data analysis.

This section will be structured as follows. First, the registration process will be explained. Second, procedure of the virtual concert will be explained.

2.1. Registration process

The registration process took place via an interface, which is an adjusted version of the interface used experiments conducted in previous Den Bosch Data Weeks by Liang & Willemsen (2019). In order to generate recommendations, participants needed to sign up with their Spotify account. Spotify was chosen as our experiment platform, as Spotify has become one of the largest music streaming platforms with millions of active users in the Netherlands. The Spotify Web API[2] is used to get access to user's top tracks and audio features in order to generate content-based recommendations.

After the Spotify login, participants were asked to complete the questionnaire about their musical sophistication. The Musical Sophistication Index is an instrument to assess self-reported musical skills and behaviours on multiple dimensions Müllensiefen et al. (2014). This research collected data on the user's the Musical Sophistication Index for Active Engagement (MSAE

[1] https://www.denbosch.nl/nl/denboschdataweek

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

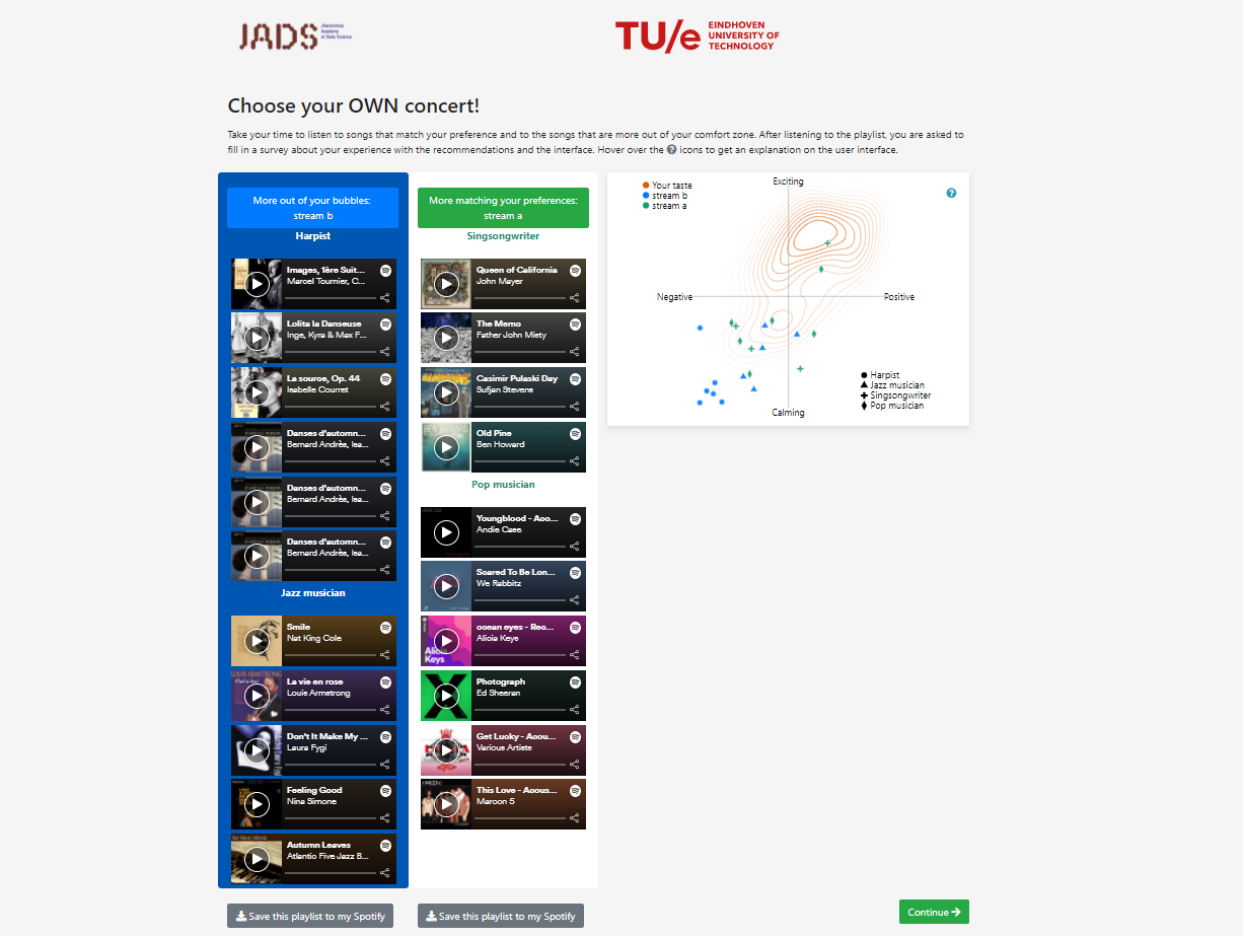

Figure 2: Main interface of the application that covers the registration process

score) and Musical Sophistication Index for Emotion (MSE score). For the scope of this research, only the MSAE score is taken into consideration. This questionnaire was already used in prior research of Liang & Willemsen (2019) and remained unchanged for the purpose of this research.

The main screen of the interface is shown in Figure 2. In this interface, participants were able to explore the music that was going to be performed during the concert. At the left, two streams (a and b) are shown. For each participant, each stream is labeled either with "more out of your comfort zone" or "more matching your preference". This labelling is based on the comparison of the recommendation scores from both streams. Recall that the recommendation scores per stream were calculated by averaging the recommendation scores of all tracks of the corresponding stream. The recommendation scores per track were calculated by multiple one-dimensional GMMs, one for each audio feature (accousticness, danceability, energy, liveness, valence, speechiness). The stream that had the lowest recommendation score was furthest from the user's current preference, so was labeled as more out of their comfort zone. The stream with the highest recommendation score was labeled as the

stream that was more matching their preference.

In order to investigate the first hypothesis (H1), we randomized which stream was preselected by default. This means that some users had the stream that was more out of their bubble as the default preselected stream, whereas other users had the stream that was more in their comfort zone as the default preselected stream. In the interface, the default choice is presented as the lefthand stream and is highlighted to indicate that this stream is preselected.

Furthermore, participants were able to listen the music and to save the playlist to their Spotify account. Each stream is composed of two artists. Stream a is composed of the singer-songwriter and the pop musician and stream b is composed of the harpist and the jazz musician. These artists are set together based on content-based features of the songs they were going to perform at the concert. The lefthand stream represents the stream that is chosen by default. If the participant wanted to switch to the other stream, they could click the right stream and a box will appear to clarify that that stream is chosen.

At the right, the visualization of the user's music preference and songs that were going to be performed is shown. The visualization is an adjusted version of the visualization made by Meeuwisse (2019). In the visualization, the user can see how the songs that match their preference (green dots) together with the songs that are more out of their bubble (blue dots) relate to their music taste (orange outline). The graph shows two features of the song: energy (the excitement of the song) and valence (the positiveness of the song) extracted from the tracks using the Spotify API[2]. The songs are labeled with four symbols representing the different artists: singer-songwriter, harpist, pop musician and jazz musician.

In the bottom right corner, users can, after selecting the stream they want to listen, click continue, which will redirect them to the next screen.

In the next screen, participants are asked to fill in a short questionnaire about their experience with the recommendations and the interface. The questionnaire is generated using the Likert Scale Robinson (2014). An overview of the questions that were asked can be found in Table 1.

In the next screen, participants are asked to rate a number of paintings, the outcomes of which were not considered a part of the current research. Finally, the participants received a personalized link to the concert. A full overview of the interface can be found using this link[3].

2.2. Concert

In Figure 3, the interface of the concert stream is shown. In the top left, the stream is shown. At the bottom, some artist information is given. At the top right, the participant's personal information (email, chosen stream) is shown. At the bottom right, the name of the song, the name of the painting and questions regarding the song and the painting are given. Only the first two questions ('How familiar is this song to you?' and 'Rate how much you liked the song.') are relevant for this research. The questions are described in Table 1. The idea is that participants listen to the song, and rate the song afterwards. The rating of the songs will be used to investigate the second hypothesis

[3]https://dbdw-culture-night.herokuapp.com/

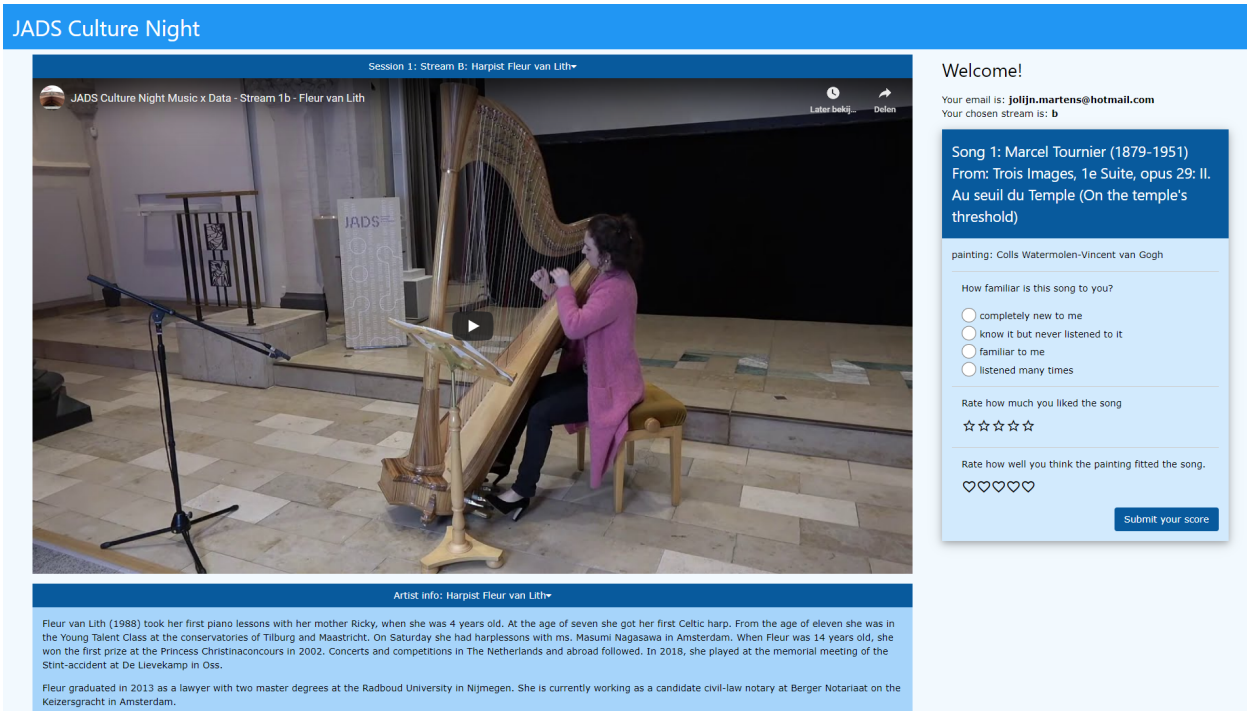

Figure 3: Main interface of the application that shows the concert

(H2). After the session is completed, participants are asked to answer two questions about the session as a whole. Again, for an overview of the questions, see Table 1. The questions on the session level fall outside the scope of this study but could be interesting for future work.

After the first session, participants are asked the following question: "Do you want to stay at this stream or switch to the other stream?". In both cases, the experiment will remain the same as in the first session, only a different stream is presented this time. The main goals of giving participants the possibility to switch between streams in the middle of the concert are firstly, to analyze their behavior after exploring music out of their comfort zone. Do participants tend to switch back to music they are used to after exploring music? Secondly, to analyze whether participants want to start exploring music after listening to music that is familiar to them.

3. Methods

In this section, a brief overview of the collected data is given. Further, the preprocessing and data analysis techniques are explained.

The sample consisted of 58 participants, of which 47 participants participated in the live experiment. The other 11 participants took part in the not-live version. There were 19 participants that only took part in the registration process and did not watch and rate the songs performed in the concert. We tried to maintain as much data as possible by using data from the 58 participants in the first part of our data analysis. Only in the part where the ratings needed to be considered, we limited our scope to the data of 39 participants. Note that there

Stage	Statements / Questions
Registration	The stream that should match with my music preference actually matched my music preference.
	The stream that was more out of my bubble challenged me to explore a new music taste.
	I like the stream that was out of my bubble more than I expected beforehand.
	The visualization helped me in understanding my own music taste.
	The visualization helped me in understanding the song characteristics of the different sessions
	The visualization helped me in choosing which session to go
Concert (song level)	How familiar is this song to you?
	Rate how much you liked the song (5 star scale)
Concert (session level)	I enjoyed the session a lot.
	The session helped me explore a new genre / music style

Table 1: Overview of the questions asked during the experiment

were also participants that were present at the concert, but dropped out halfway. Those participants were still included in the data analysis, as otherwise too few data points remained. However, only the songs that they have rated were included.

3.1. Preprocessing

In order to be able to answer the main research question, some preprocessing has been done. An overview of the dataset that is used to investigate hypothesis H1 and hypothesis H1.1 can be found in Appendix A. The variables that were added to the original dataset are *user_id*, *age*, *male*, *default_outbubble*, *recommended_stream*, *chosen_out_of_bubble_stream*, and *rec_score_difference*. We define the *default_outbubble* group as the group of users that had the more out of their comfort zone as the default playlist in the registration process. The *default_inbubble* group is defined as the group that had the playlist that fitted their preference most as default playlist in the registration process. The participants were randomly divided into the *default_outbubble* group and the *default_inbubble* group, meaning that the participants were not able to choose which stream was selected by default.

The recommended stream is calculated by taking the maximum of all recommendation scores from a particular user an selecting the stream that belongs to that maximum. The recommendation scores are calculated by making use of the recommendation algorithm as provided in the study of Liang & Willemsen (2019). However, some adjustments needed to be made, as the purpose of this research was to recommend a whole playlist instead of individual songs. The main principle stayed the same, but now the recommendation scores are averaged for each playlist. The variable *chosen_out_of_bubble_stream* is created by comparing the recommended stream with the chosen stream. If the recommended stream corresponds to the chosen stream, participants did not choose to explore music out of their comfort zone. On the other hand, if the recommended stream did not correspond to the chosen stream, participants did choose to explore music out of their comfort zone.

In order to investigate hypothesis H1.1, the interaction effect of the variable *chosen_out_of_bubble_stream* with the variables MSAE score, age, recommendation score difference is investigated. This indicates whether the susceptibility to nudging is dependent on the variables MSAE score, age, gender and recommendation score difference.

The recommendation score difference is calculated by taking the difference in *rec_scores* between the user's out of their comfort zone's and in their comfort zone's playlist. So the higher the value of *rec_score_difference* for a particular user, the more the streams differ in how far they are from their preferences.

In order to investigate our second hypothesis, the data had to be transformed by having one row containing information about one specific rating, instead of information about one specific user. An overview of the dataset used to investigate H2 can be found in Appendix B. The variables contain information about either the user in general (email, user id, recommendation id, MSAE score, gender, whether they chose to explore music in the registration process, whether the user switched in the middle of the concert), about the song in general (artist, place in the concert), or about the interaction of the user with the song (performance rating, familiarity rating, rank). The performance rating of a song is defined as the answer from a participant to the statement: "Rate how much you liked the song". The answers are in a 5 star scale. The performance rating column will be the dependent variable of our multilevel model.

The independent variables of the multilevel model are: the rank of a song, MSAE score, *chosen_out_of_bubble_stream*, gender and the familiarity rating. The rank of a song is defined as a number between 1-20 where 1 indicating the song with the highest recommendation score and 20 indicating the song with the lowest recommendation score. The familiarity rating of a song is defined as the answer from a participant to the question: "How familiar is this song to you". The answers are divided into four categories: (1) the song is completely new to me, (2) I know the song but never listened to it, (3) the song is familiar to me, (4) I listened to the song many times.

As participants got the possibility to switch halfway through the concert, the variable *chosen_out_of_bubble_stream* has been switched accordingly for the users that switched.

There were some mismatches between the songs in the registration process and the songs in the concert. The third song of the singer-songwriter in the concert did not match the song in the registration process, which made it impossible to match the rank with the rating. Therefore, it was decided to remove this song from the multilevel model. Note the performed song was properly rated, therefore it is included in Figure 4 and Figure 5. Secondly, the third song of the harpist consisted of three different songs that each contained a different rank. Therefore, for each user, the mean rank for this song is calculated. Thirdly, it was decided to exclude the last song of the pop musician from both the exploratory analysis as from the multilevel model, as the last song was not performed due to technical difficulties.

In order to investigate the usability question of whether the visualization in the registration process helped the participants to understand their own music taste and the music to be explored, the answers from statement 4-6 in Table 1 are investigated. Statement 4 is defined as "the visualization helped me in understanding my own music taste". Statement 5 is defined as "the visualization helped me in understanding the song characteristics of the different sessions". Statement 6 is defined as "the vi-

sualization helped me in choosing which session to go to". In the next section, the distribution of the ratings of these statements are visualized and the effect of age and MSAE score on these ratings are analyzed.

Now, some descriptive statistics about the 58 participants are presented. The data consisted of 24 males and 34 females, ranging from 18 years to 64 years old (mean = 30.28, sd = 12.40). The default option to preselect the out of your bubble stream held true for 27 participants (46.6%) and held false for the other 31 participants (53.5%). From all 58 participants, 67.2% had chosen stream a (the stream of the singer-songwriter and the pop musician) during the registration process. The other 32.8% had chosen stream b (the stream of the harpist and the jazz musician). Further, 29.3% of the participants had chosen the stream that was out of their bubble and thus decided to explore music. The other 70.7% chose the stream that was most inside their comfort zone. The Musical Sophistication index for Active Engagement (MSAE) (range 1-7) has a mean of 3.39 (sd = 1.80). Some basic statistics about the ratings: the data consists of 269 ratings of 39 users. The mean performance rating is 3.5 out of 5 (sd = 1.30).

3.2. Data analysis techniques

The data analysis techniques that are used are Pearson's correlation coefficients, chi-square tests, logistic regression analysis and multilevel modelling. In this section, the data analysis techniques will be explained one by one.

The Pearson's correlation coefficient is the test statistic that measures the statistical relationship between two variables X and Y. The Pearson's Correlation Coefficient is defined as

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

The chi-square test is a statistical hypothesis test that is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories. In our data analysis, a chi-square test is performed on, among other things, the relative differences of exploring music out of your comfort zone between the *default_inbubble* group and *default_outbubble* group.

Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable. Like all regression analyses, the logistic regression is a predictive analysis, as it is used to describe the data and to explain the relationship between one (or more) dependent variable and one or more independent variables. The result of the logistic regression model is the impact of each variable on the odds ratio of the observed event of interest. Mathematically, in our experiment, the logistic regression estimate is defined as

$$log\left(\frac{p(y=1)}{1-p(y=1)}\right) =$$

$$\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 * \beta_3 * (x_1 * x_2)$$

in which $y = 1$ indicates that the user chose the stream that was out of their comfort zone, x_1 corresponds to whether the out of their comfort zone stream that was selected by default, x_2 corresponds to the user's gender and x_1 * x_2 corresponds to the interaction term between the independent variables.

Multilevel models are statistical models of parameters that vary at more than one level, which is particularly appropriate for nested data. In our experiment, this multilevel model is a 2-level model as the data contains different ratings on the experience of different songs that are grouped by individual. There exist multiple types of multilevel models: random intercepts models, random slopes models and random intercepts and slopes models. In this study, a random intercept model is built, as we do not want the effect of the predictor to vary between individuals. The multilevel model used in this study is defined as

$$y_{ij} = \beta_0 + \beta_1 * x_{1ij} + \beta_2 * x_{2ij} + \beta_3 * x_{3ij}$$
$$+\beta_4 * x_{4ij} + \beta_5 * (x_{2ij} * x_{4ij}) + u_j + e_{ij}$$
$$u_j \sim N(0, \sigma_u^2)$$
$$e_{ij} \sim N(0, \sigma_e^2)$$

in which y_{ij} is defined as the performance rating of a specific song given by an individual. The independent variables (fixed part) are defined by 5 variables: x_{1ij} the MSAE score, x_{2ij} whether the song is out of the user's comfort zone, x_{3ij} the familiarity rating, x_{4ij} the gender, and $(x_{2ij} * x_{4ij})$ the interaction term between whether the song is out of the user's comfort zone and their gender. The random part of the model is defined by two variances. Firstly, σ_e^2 represents the variance in ratings. Secondly, σ_u^2 represents the variance between individuals.

4. Results

In Section 4.1., hypothesis H1 is investigated by performing chi squared tests and interpreting the logistic regression model as described in Section 3.2.

In Section 4.2., hypothesis H1.1 is investigated by investigating the interaction effects of the logistic regression model.

In Section 4.3., the performance rating is analyzed to be able to investigate hypothesis H2. First of all, some exploratory analysis on the performance and familiarity ratings is performed. Secondly, the multilevel model as described in Section 3.2. is interpreted and visualized.

In Section 4.4., the usability question whether the visualization in the registration process is helpful in understanding their own music taste and the music to be explored is investigated. This is done by visualizing the ratings to the statements that were given in the registration process. Further, it is researched whether there is a mediating effect of MSAE score and age on the helpfulness of the visualization in the registration process.

4.1. The effect of the default playlist (H1)

In order to investigate the effect of the default playlist, the variable *default_outbubble* is investigated. From the *default_outbubble* group, 44.4% chose the stream that was outside their comfort zone. From the *default_inbubble* group, only 16.1% chose the stream that was outside their comfort zone. A chi-square test was performed and resulted in a significant relationship between the default option being the out of your comfort zone stream and the choice to explore music out of your comfort zone, $\chi^2(1, N = 58) = 4.3, p = .038$. This already indicates a significant effect of selecting the out of their comfort zone playlist as default playlist on exploring music out of their comfort zone.

In order to validate this effect and to check whether this effect might by mediated by other variables like age, gender, MSAE

Table 2: Logistic regression model with interaction effect

Model:	Logit	Pseudo R-squared:	0.087
Dependent Variable:	chosen_out_of_bubble_stream	AIC:	72.0708
Date:	2020-12-27 21:55	BIC:	80.3126
No. Observations:	58	Log-Likelihood:	-32.035
Df Model:	3	LL-Null:	-35.085
Df Residuals:	54	LLR p-value:	0.10692
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
const	-1.7346	0.6262	-2.7699	0.0056	-2.9620	-0.5072
default_outbubble	1.7346	0.8233	2.1068	0.0351	0.1209	3.3483
male	0.2305	1.0016	0.2301	0.8180	-1.7326	2.1937
male x default_outbubble	-0.7005	1.2704	-0.5514	0.5814	-3.1905	1.7895

score and the recommendation score difference between the streams, a logistic regression model is built. The complete logistic regression model can be found in Appendix C. In Table 2, only the variables with the strongest effect are included. It should be noted that the p-value of the overall model was insignificant ($p = 0.107$), meaning that the null hypothesis that the model with no independent variables fits the data as well as our model could not be rejected. When investigating the individual regression coefficients, a strong positive and statistically significant coefficient of *default_outbubble* can be found, $\beta_1 = 1.73, t(55) = 2.1, p = .035$. This positive relationship indicates that participants more often chose the out of their comfort zone stream when it was set as the default stream, indicating that the nudge to explore music was successful. Therefore, our results are in line with hypothesis H1.

4.2. The effect of MSAE score, age, the difference in recommendation score and gender on the susceptibility to nudging to explore music (H1.1)

In order to investigate the effect on MSAE score, age and gender on the susceptibility to nudging to let participants explore music out of their comfort zone, the interaction effects of the logistic regression model are investigated. Note that the interaction effects were included one by one in the model, but for simplicity reasons, not all models are included separately. The complete logistic regression model is shown in Appendix C. Based on the complete logistic regression model, none of the interaction effects (1) *MSAE * default_outbubble* (2) *age * default_outbubble* (3) *rec_score_dif * default_outbubble*, (4) *gender * default_outbubble* were significant. Based on this result, we reject hypothesis H1.1. Therefore, whether people are susceptible to nudging to explore music out of their comfort zone is *independent* of their MSAE score, age, the difference in recommendation score between streams and their gender.

4.3. Predicting user's performance ratings (H2)

During the concert, users were asked to answer the questions as stated in Table 1. This section used the ratings to the questions on a song level. For the performance rating, the statement of interest that was given to the participants was: "I enjoyed the session a lot", which the user's could answer by giving a rating between 0 and 5. For the familiarity rating, the question that was asked to the participants was: "How familiar was this song to you?". The performance rating is considered the independent variable of the multilevel model that was built, whereas the familiarity rating will be included as a dependent variable.

Before analyzing the multilevel model, some exploratory analysis on the performance rating and familiarity was performed. As the performance rating is concerned, the distribution is shown in Figure 4.1. Overall, it can be seen that the concert is liked by most participants, as the majority gave a 4-star rating. In Figure 5, the performance ratings for the 4 different sessions are visualized. It can be seen that the differences in performance ratings in the jazz session were smaller than in the other sessions. Furthermore, it can be concluded that there was no visible pattern that people rated higher at the beginning or at the end of a session.

As the familiarity rating is concerned, the distribution is shown in Figure 4.2. Overall, it can be seen that the majority of the songs were new to the participants. Further, the amount of familiar and very familiar songs is about equal. Finally, only very few people knew a song but never listened to it.

In order to predict user's performance ratings, a multilevel was built. The random intercept was added, since it was not wanted that the effect of the predictor varied between participants. The random intercept explained 27.8% of the variance. Adding other random intercepts like the variation between artists and the variation between songs were not useful as those explained only little variance. The following variables were investigated: rank of a track, MSAE score,, whether the participant chose to explore music out of their comfort zone, gender, the familiarity with a song and possible interaction effects. Note that all independent variables have little correlations with each other, so that the multicollinearity assumption is not violated. The best fitting model is shown in Table 3. The complete model with all second order interaction terms can be found in Appendix D.

Firstly, there was found a small negative effect of MSAE score on the participant's performance rating, $\beta_1 = -0.18, t(32.72) = -2.77, p < .01$. This indicates that people with a higher MSAE score tend to rate the song lower compared to people with a lower MSAE score. The can be considered reasonable, as users with a higher MSAE score have more experience in listening to music, which could them make more critical in rating music performances and therefore give lower performance ratings.

Secondly, a positive effect was found of choosing to explore music out of your comfort zone on the participant's performance rating, $\beta_2 = 0.44, t(244.01) = 2.90, p < .01$. This could indicate that people tend to give higher ratings indicating that they enjoy the song more if they are aware of the fact that they are exploring music.

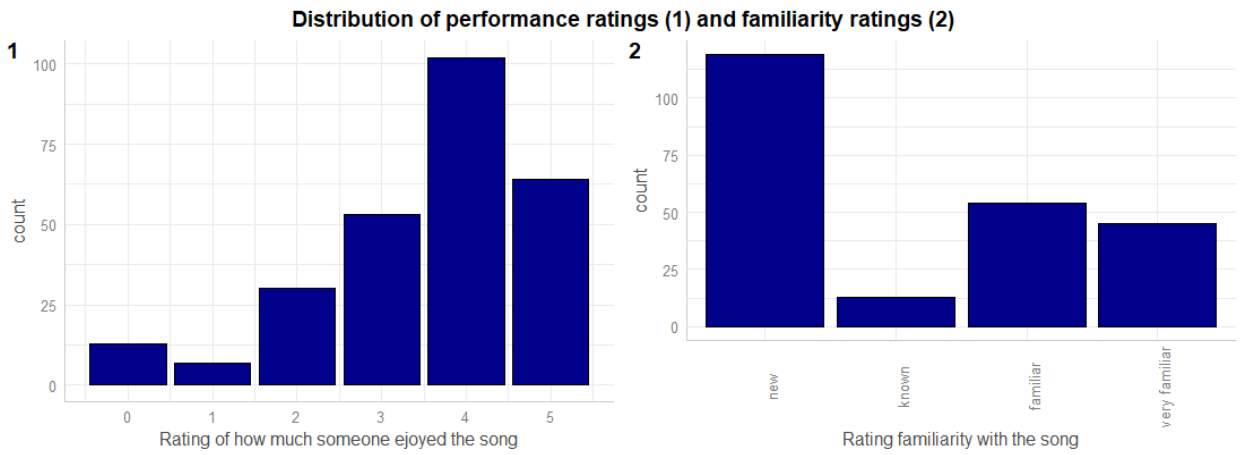

Figure 4: Visualization of the overall distribution of the performance ratings (1) and the familiarity rating (2)

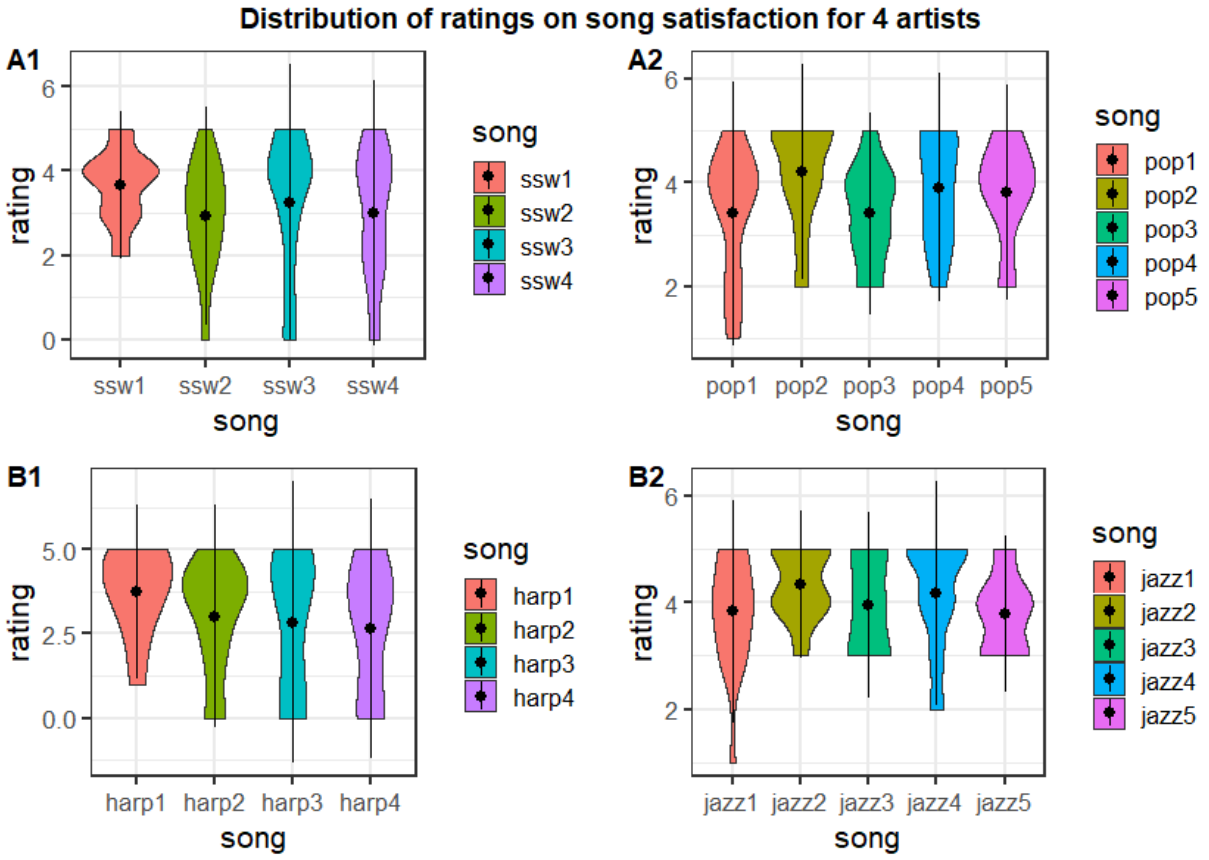

Figure 5: Visualization of rating distributions for the statement: "Rate how much you liked the song" for 4 artists.

Thirdly, a positive effect was found of the user's familiarity rating of a song on their performance rating, $\beta_3 = 0.29, t(239.17) = 6.75, p < .001$. This indicates that people tend to rate a song higher that is already familiar to them.

Fourthly, a negative effect of the interaction term *chosen_out_of_bubble_stream * male* on the user's performance rating was

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.056	0.302	39.472	13.414	0.000 ***
MSAE	-0.184	0.066	32.720	-2.766	0.009 **
chosen_out_of_bubble_stream	0.439	0.151	244.065	2.902	0.004 **
rating_fam	0.292	0.043	239.165	6.746	0.000 ***
male	-0.030	0.252	51.127	-0.119	0.906
chosen_out_of_bubble_stream:male	-0.757	0.278	227.722	-2.721	0.007 **

significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3: Multilevel model that predict the user's performance ratings (how much the participant enjoyed the song)

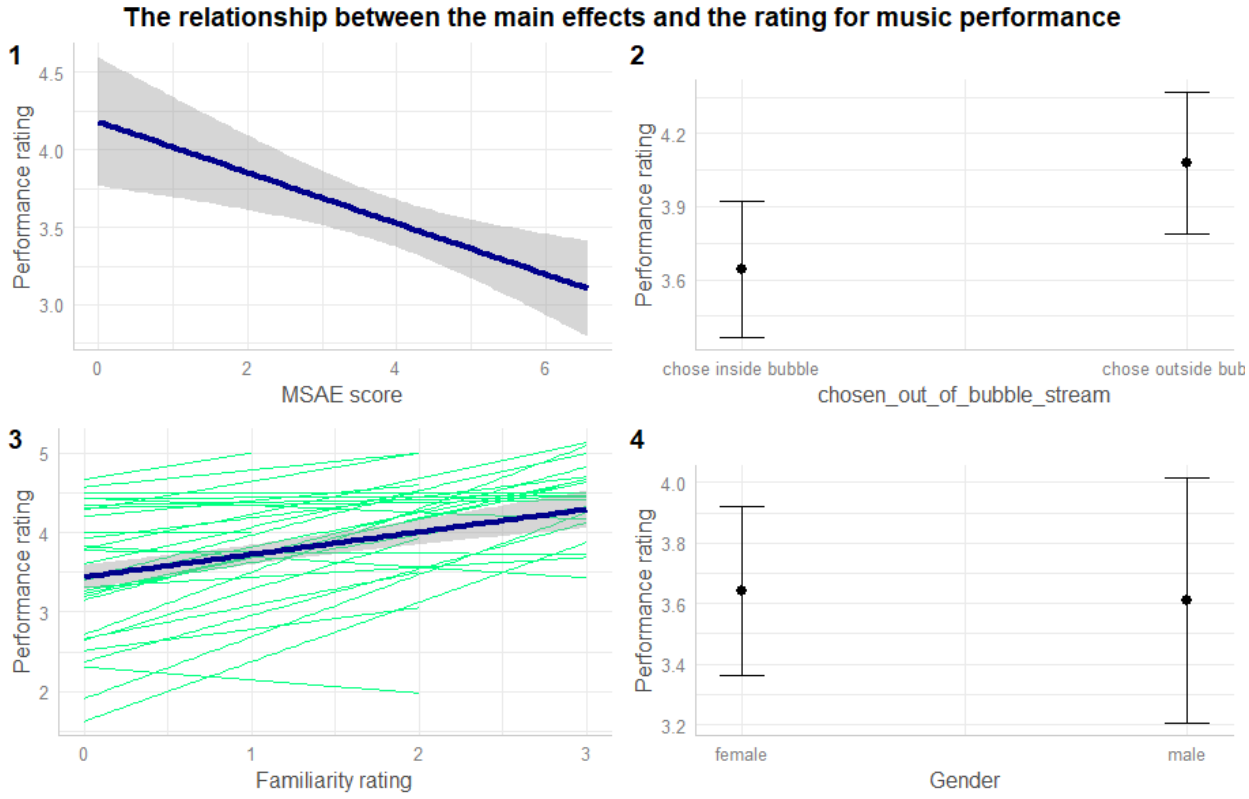

Figure 6: The relationship between the main effects (1) MSAE score (2) whether people chose to explore music (3) familiarity rating and (4) gender and the user's performance rating

found, $\beta_5 = -0.77, t(227.72) = -2.72, p < .01$. This indicates that males that chose the stream that was most out of their comfort zone are more likely to give lower performance ratings than females that chose the stream that was most out of their comfort zone. The effect of gender on its own on the performance rating was insignificant $\beta_4 = -0.03, t(51.13) = -0.12, p = 0.91$. This indicates that gender on its own has no significant influence on how the user enjoys the song performance. The variable gender however needed to be included in the model as it was connected to the higher order interaction effect.

In order to make the multilevel more explainable, the effects of the independent variables as shown in Table 3 are visualized in Figure 6 and Figure 7. In Figure 6 the relationship between the main effects MSAE, *chosen_out_of_bubble_stream*, familiarity rating and gender on the performance rating are visualized. The thicker blue lines show the overall trend given by our estimated fixed effects. The thinner green lines represents the trend for each individual. In Figure 6.1, a negative effect of MSAE score on the performance rating is shown. In Figure 6.2 and Figure 6.3, a positive effect of the familiarity rating and the fact that users decided to explore music out of their comfort zone respectively on the performance rating is observed. In Figure 6.4, a negative effect of gender on the performance rating is observed, however note that this effect is insignificant.

In Figure 7, the relation between interaction effect *chosen_out_of_bubble_stream * male* and the performance rating is visualized. From this visualization, it can be concluded that males tend enjoy the song more compared to females when it is not out of their comfort zone, whereas females tend to enjoy the song more compared to males when it is out of their comfort zone. Note that the differences within categories are large, so future research should try to reproduce this and investigate whether the same conclusion holds for a different sample.

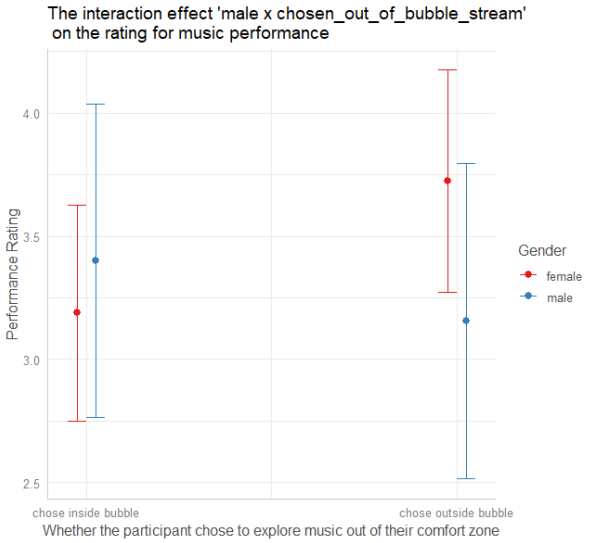

Figure 7: The relationship between the interaction effect of gender and whether people chose to explore music and the user's performance rating

Surprisingly, there was not a strong relationship between the personalized rank of a song and how much the song is enjoyed by that participant. Recall that the rank of the song is calculated by multiple one-dimensional GMMs based on the user's Spotify top tracks (for more detail see Section 1.1.2.). This weak relationship could be explained by the fact that people tend to rate the performance instead of the actual song. Therefore, the rank of a song might be not a good predictor for predicting whether participants enjoyed the song that was performed.

4.4. Helpfulness of the visualization in the registration process (usability question)

In order to answer our usability question of whether the visualization in the registration process (Figure 2) helped users to understand their own music taste and the music to be explored, some exploratory analysis was performed. The main piece of information was extracted from survey that was conducted in the the registration process (see Table 1). The statements of interest were (4) the visualization helped me to understand their own music taste, (5) the visualization helped me to understand the song characteristics of the different sessions, (6) the visualization helped me to choose which session to go to.

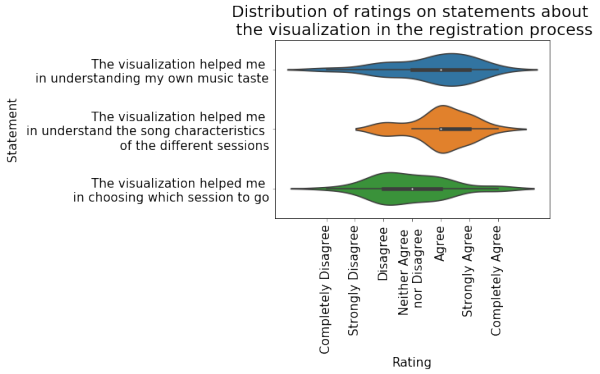

Figure 8: Distribution of rating concerned the statements on the visualization in the registration process

In Figure 8, the distribution of the ratings for the different statements is shown. Overall, users rated relatively positive, which means that the visualization did help them to understand their own music taste and the music to be explored. Especially in statement 5, whether the visualization helped them in understanding the song characteristics of the different session, the majority agreed. However, users gave a neutral rating to the question whether the visualization helped them to choose which session to go to. Hence, the visualization that was shown in the registration process can be considered useful to also show in future experiments, were the main focus point of improvement lies in improving the helpfulness in choosing which session to go to.

Additionally, it is investigated whether the helpfulness of the visualization in the registration process is mediated by the variables age and MSAE score. For all three statements, a Pearson's correlation test is performed for both age and MSAE score. All correlations were insignificant and can be found in Appendix E. In conclusion, whether the visualization helped participants to understand their own music taste and the music to be explored is not mediated by their age and MSAE score.

5. Discussion

This study gave an innovative and exploratory view on how people can be influenced to explore music out of their comfort zone and on the experience of music exploration. In this section, the limitations of this research are explained and opportunities for future work are presented.

As this study was based on a live experiment, there were some issues that could not been solved on the spot and compromises needed to be made. First of all, many participants and people that wanted to participate struggled with the fact that two separate steps needed to be performed to be able to take part in the experiment, namely, the registration process and the actual concert. It was not entirely clear that participants needed to participate in both parts. As a consequence, there were missing data points, for example from participants that

only took part in the registration process, but were not present during the actual concert. Future work could try to integrate both parts of the experiment more by creating one application that participants could walk through in one go. Thereby however, the live component of the experiment will be lost. Secondly, at the night of the concert, there were technical issues with the live stream of the pop musician, which caused a delay of circa 15 minutes. As a consequence, people stopped watching the stream which led again to missing data points. These issues could have been prevented, especially when the experiment would be held in a non-live fashion.

Besides missing data points, there was also the issue of too little data points as not enough people participated in the experiment. The statistical power and thereby the power of the conclusions that were drawn would increase if more data points were to be collected. Multiple attempts have been made to increase the number of participants. The most successful attempt was to give people the possibility to take part in the experiment after the experiment had already passed, so to take part in the non-live version. The number of participants increased by 23.4%. However, a total of 58 participants (which included the participants that had missing data) is not enough to draw generalizable conclusions. Future work could try to increase the number of data points by lowering the effort, for example decreasing the duration of the concert and by making the experiment non-live.

Another issue that should be addressed is that only people that had access to a Spotify account were able to take part in the experiment. Therefore, conclusions about exploring music recommendation cannot be generalized to the population as a whole. Moreover, it is questionable how representative a Spotify account is for the participant's music preference, as some participants made a Spotify guest profile specifically to take part in this experiment. Therefore, the top tracks that were used as an input for the recommendation algorithm provided by Liang & Willemsen (2019) could be not representative for the participants actual preference. In addition to that, it might be the case that users share a Spotify account with family or friends, which makes the top tracks of a user also not representative for their music taste. Future research could try to limit the effect of a non-representative Spotify account by including survey questions about their experience and expertise with Spotify in the experiment. For example by defining another musical sophistication index score, specifically for experience with Spotify, which could then be included as a control variable.

In our data analysis, only the Musical Sophistication Index for Active Engagement (MSAE) is taken into account. However, in the conducted experiment, also the Musical Sophistication Index for Emotion (MSE) was measured. The reason for excluding the MSE was the high correlation between someone's MSAE score and MSE score. If both the MSAE score and the MSE score were included in the logistic regression model, the assumption of no multicollinearity would be violated, as logistic regression requires to be little or no multicollinearity among the independent variables. Future research could investigate whether the MSE score has an effect on either whether people want to explore music out of their comfort zone, or whether people with a higher MSE score are more

susceptible to nudging to explore music.

Due to limited time, the questions that were asked during the concert on a session level (see Table 1) were not included in the data analysis. Future work could investigate how ratings differ on session level compared to song level.

In order to evaluate hypothesis 1, the binary variable *chosen_out_of_bubble_stream* was created, on which a logistic regression model was built. Therefore, a classification problem is solved, whereas the extent to which people want to explore music can also be interpreted as a continuous variable. In the current study, the songs that artists performed were still quite different from each other in terms of energy and valence (see the visualization in Figure 2). Those differences are currently not taken into account due to the decision to make exploring music out of your comfort zone a binary variable. Future research could model the experiment in a different way, for example to let participants choose multiple times between different songs, instead of choosing one time between two playlists. In this way, the extent to which users are likely to explore music out of their comfort zone can be interpreted as a continuous variable. This solves the problem of differences in audio features within playlists. In the current experiment, users needed to compare the audio features of the different songs to be able to understand the differences in audio features for each playlist. This made it hard for participants to understand which session is more close to them. This issue can be prevented by letting participants choose multiple times between two songs, one song that is inside their comfort zone and one that is outside their comfort zone.

In order to evaluate hypothesis 2, the performance ratings of different songs were investigated. It should be noted that the addition of the familiarity rating as an independent variable in the multilevel model caused the random intercept to explain only 27.8% of the variance. Without including the familiarity rating, the random intercept explained 95.3% of the variance. This is due to the fact that the familiarity rating differs per individual.

Further, a surprising finding was the weak relationship between the rating and the the rank of the track as calculated by the multiple one-dimensional GMMs. It would be expected that if the track fits the user's current music preference (based on the top tracks of the user's Spotify profile), it would lead to a high rating. A possible explanation could be that users were inclined to rate the performance of the artist, and not the song with their corresponding audio features. Future work could investigate how exploring novel music is enjoyed in a non-concert setting, for example by playing the Spotify tracks one by one and then asking to rate the songs.

Despite the experiments' limitations, this experiment can be considered as a preliminary exploratory view of trying to explain what influences people to explore music out of their comfort zone and the experience of music exploration.

6. Conclusion

In this study, recommender systems are used to explore objects out of someone's comfort zone. This research specifically focuses on music recommendation by making use of the Spotify API. The main question of interest was: what influences people to explore music out of their comfort zone and how is this mu-

sic exploration experienced? This research question has been answered by investigating two main hypotheses and one sub hypothesis. In addition, a usability question is also investigated. In this section, we will draw conclusions part by part and we end with the main conclusion.

To recapitulate, the first hypothesis was: making a recommendation list the default list will increase the probability of selecting that list. From the logistic regression model, it can be concluded that there is a strong effect of the default preselected playlist on the decision to explore music out of their comfort zone. This indicates that the nudge to set defaults has been successful in this experiment. This result can be used in future experiments that want to let people explore music out of their comfort zone.

In hypothesis H1.1, we hypothesized that whether people are susceptible to nudging is dependent on their MSAE score, age, the difference in recommendation score between streams and gender. Based on our results, we can conclude that all variables that were investigated did not have a significant effect on whether people are susceptible to nudging to explore music out of their comfort zone.

The second main hypothesis was: ratings for music performance will depend on someone's rank of a track, MSAE score, whether they chose to explore music out of their comfort zone, gender and their familiarity rating. The performance rating was defined by whether people liked the song that was performed. From the multilevel model, six conclusions can be drawn. Firstly, people with a higher MSAE score tend to give lower performance ratings. Secondly, people that are aware of the fact that they were exploring music out of their comfort zone tend to give higher performance ratings. Thirdly, people tend to give higher performance ratings to songs that were more familiar to them. Fourthly, females and males tend to rate the performance similarly, as no effect of the variable gender on its own was found. Fifthly, males tend to give higher performance ratings compared to females when it is not out of their comfort zone, whereas females tend to give higher performance ratings compared to males when it is out of their comfort zone. Finally, it can be concluded that when participant's are asked to rate a song in an environment of a performance, they are more likely to rate the performance instead of the song.

After investigating the main hypothesis, the usability question whether the visualization in the registration process helped users to understand their own music taste and the music to be explored is answered. Based on the exploratory analysis, users were of the opinion that the visualization helped them to understand their own music taste and the music to be explored. However, they did not necessarily agree or disagree to the statement that the visualization helped them to choose which session to go to. Hence, the visualization that was shown in the registration process can be considered useful to show in future experiments, where the main focus point of improvement lies in improving the helpfulness in choosing which session to go to. Further, the helpfulness of the visualization to understand their own music taste and the music to be explored was not mediated by the participant's age and MSAE score.

In conclusion, the default option to select the playlist that was out of the peo-

ple's comfort zone nudged people to explore music. The susceptibility of nudging to explore music is not mediated by any of the investigated factors. As the experience of music exploration is concerned, people tend to enjoy a song more when they are aware of the fact that they are exploring music, but also tend to enjoy a song more when it is familiar to them.

Acknowledgements

I would like to thank the following people for helping with this research project. Firstly, I want to thank M.C. Willemsen and Y. Liang for providing guidance and a sounding board when required. They have helped building the interfaces and helped setting up the experiment overall, by contacting artists and the Jheronimus Academy of Data Science (JADS). Secondly, I want to thank JADS for promoting this event and the artists that performed live during the concert. Finally, I want to thank all participants for participating in this experiment.

Appendix A: Overview of variables in dataset 1 which is used to investigate H1

Variable	Categories	Explanation
email	string	The email address participants filled in during the registration process.
user_id	string	The unique user id for each participant.
MSAE	range 1-7	Musical Sophistication Index for Active Engagement
MSE	range 1-7	Musical Sophistication Index for Emotion
age	integer	age in years
male	1 or 0	1 if male 0 if female
session1	ssw or harp	Session 1 is the first half of the concert. Participants could choose between listening to a singer-songwriter (ssw) or to a harpist (harp)
session2	pop or jazz	Session 2 is the second half of the concert. Participants could choose between listening to a pop musician (pop) or to a jazz musician (jazz)
default_outbubble	1 or 0	1 if default stream is the out of your comfort zone stream 0 if the default stream is the in your comfort zone stream
chosen_stream	a or b	The stream that the participant picked after the registration process. a = the stream consisting of the sing-songwriter and the pop musician. b = the stream consisting of the harpist and the jazz musician.
recommended_stream	a or b	The stream that was recommended to the participants based on the recommendation algorithm. a = the stream consisting of the sing-songwriter and the pop musician. b = the stream consisting of the harpist and the jazz musician.
chosen_out_of_bubble_stream	1 or 0	Binary value indicating whether the user picked the stream that was most out of their comfort zone. 1 if the user picked the stream that was out of their comfort zone 0 if the user picked the stream that was in their comfort zone
rec_score_difference	integer	Integer indicating the difference in recommendation scores between the two playlists for a particular user.

Appendix B: Overview of variables in dataset 2 which is used to investigate H2

Variable	Categories	Explanation
email	string	The email address participants filled in during the registration process.
user_id	string	The unique user ID for each participant.
rec_id	string	Unique recommendation ID for each user that is created when recommendations are generated in the registration process
male	1 or 0	1 if male 0 if female
MSAE	range 1-7	Musical Sophistication Index for Active Engagement
session_id	harp, jazz, ssw, pop	Id indicating which session the rating belongs to
rating _performance	range 0-5	Performance rating which corresponds to an answer to the statement: "Rate how much you liked the song (5 star scale)"
song	harp1, harp2, harp 3, harp 4 jazz1, jazz2, jazz3, jazz4, jazz5 pop1, pop2, pop3, pop4, pop5 ssw1, ssw2, ssw4	String indicating which song is played at which session and when exactly.
chosen_out_of _bubble_stream	1 or 0	Binary value indicating whether the song was inside or outside the user's comfort zone 1 if the song was out of their comfort zone 0 if the song was in their comfort zone
rank	range 1-20	integer that gives an order of which song fits the user best based on user's Spotify top tracks 1 indicates the song with the highest recommendation score 20 indicates the song with the lowest recommendation score
switched	1 or 0	Value indicating whether the user switched in the middle of the concert 1 if switched 0 if not switched
rating_fam	0, 1, 2 or 3	Familiarity rating which corresponds to the answer to the question: "How familiar is this song to you?" The categories are: 0: completely new to me 1: know it but never listened to it 2: familiar to me 3: listened many times

Appendix C: Complete logistic regression model with second order interaction terms

Model:	Logit	Pseudo R-squared:	0.096
Dependent Variable:	chosen_out_of_bubble_stream	AIC:	83.4449
Date:	2021-01-09 13:26	BIC:	104.0493
No. Observations:	58	Log-Likelihood:	-31.722
Df Model:	9	LL-Null:	-35.085
Df Residuals:	48	LLR p-value:	0.66579
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
const	-2.5115	2.2142	-1.1343	0.2567	-6.8513	1.8282
rec_score_difference	-0.0354	0.3027	-0.1168	0.9070	-0.6287	0.5580
default_outbubble	1.7780	3.3030	0.5383	0.5904	-4.6958	8.2518
age	0.0162	0.0444	0.3640	0.7158	-0.0708	0.1031
male	0.2866	1.0694	0.2680	0.7887	-1.8094	2.3826
MSAE	0.1157	0.3111	0.3718	0.7100	-0.4941	0.7254
rec_score_dif x default_outbubble	0.1110	0.4722	0.2350	0.8142	-0.8146	1.0365
male x default_outbubble	-0.7386	1.3283	-0.5561	0.5782	-3.3421	1.8648
age x default_outbubble	-0.0151	0.0577	-0.2615	0.7937	-0.1281	0.0980
MSAE x default_outbubble	-0.0199	0.3918	-0.0507	0.9596	-0.7877	0.7480

Appendix D: Complete multilevel model with all second order interaction terms

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.535	0.531	91.369	8.546	0.000 ***
rank	-0.030	0.027	216.136	-1.109	0.269
MSAE	-0.257	0.120	85.652	-2.135	0.036 *
chosen_out_of_bubble_stream	-0.131	0.532	166.867	-0.246	0.806
rating_fam	0.355	0.147	225.070	2.414	0.017 *
male	-0.453	0.668	40.339	-0.678	0.502
rank:MSAE	0.001	0.006	210.528	0.138	0.890
rank:chosen_out_of_bubble_stream	0.030	0.020	216.072	1.543	0.124
rank:rating_fam	0.010	0.009	216.503	1.176	0.241
rank:male	-0.016	0.019	216.650	-0.862	0.390
MSAE:chosen_out_of_bubble_stream	0.092	0.108	170.365	0.849	0.397
MSAE:rating_fam	-0.025	0.030	226.315	-0.843	0.400
MSAE:male	0.147	0.162	35.636	0.907	0.370
chosen_out_of_bubble_stream:rating_fam	-0.079	0.092	226.554	-0.856	0.393
chosen_out_of_bubble_stream:male	-0.758	0.303	230.145	-2.505	0.013 *
rating_fam:male	-0.017	0.095	228.331	-0.183	0.855

significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Appendix E: Results of Pearson's correlation tests between the rating on the given statement and age and MSAE score respectively

	Statement	Pearson's correlation test for age	Pearsons's correlation test for MSAE score
4.	The visualization helped me in understanding my own music taste.	$r(56) = -.13$, $p = .386$	$r(56) = -.18$, $p = .244$
5.	The visualization helped me in understanding the song characteristics of the different sessions.	$r(56) = -.23$, $p = .121$	$r(56) = -.06$, $p = .710$
6.	The visualization helped me in chosing which session to go.	$r(56) = -.02$, $p = .871$	$r(56) = -.09$, $p = .534$

References

Bauer, C., & Schedl, M. (2017). Introducing surprise and opposition by design in recommender systems. In *Adjunct publication of the 25th conference on user modeling, adaptation and personalization* (pp. 350–353).

Bikker, Y. (2020, Jul). *The 7 most creative examples of habit-changing nudges.* The Startup. Retrieved from `https://medium.com/swlh/the-7-most-creative-examples-of-habit-changing-nudges-7873ca1fff4a`

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Biswas, D., Lund, K., & Szocs, C. (2016). Ambient music and food choices: Can music volume level nudge healthier choices? *ACR North American Advances*.

Bothos, E., Apostolou, D., & Mentzas, G. (2015). Recommender systems for nudging commuters towards eco-friendly decisions. *Intelligent Decision Technologies*, *9*(3), 295–306.

Bothos, E., Apostolou, D., & Mentzas, G. (2016). A recommender for persuasive messages in route planning applications. In *2016 7th international conference on information, intelligence, systems & applications (iisa)* (pp. 1–5).

Carrasco, O. C. (2020, Feb). *Gaussian mixture models explained.* Towards Data Science. Retrieved from `https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95`

Dinner, I., Johnson, E. J., Goldstein, D. G., & Liu, K. (2011). Partitioning default effects: why people choose not to choose. *Journal of Experimental Psychology: Applied*, *17*(4), 332.

Fort, T. L. (2018). A response to olivier urbain and an exploration of how music may serve as a nudge for more ethical and peaceful business behavior. In *College music symposium* (Vol. 58, pp. 1–14).

Jesse, M., & Jannach, D. (2020). Digital nudging with recommender systems: Survey and future directions. *arXiv preprint arXiv:2011.03413*.

Kapoor, K., Kumar, V., Terveen, L., Konstan, J. A., & Schrater, P. (2015). " i like to explore sometimes" adapting to dynamic user novelty preferences. In *Proceedings of the 9th acm conference on recommender systems* (pp. 19–26).

Kapoor, K., Srivastava, N., Srivastava, J., & Schrater, P. (2013). Measuring spontaneous devaluations in user preferences. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1061–1069).

Koren, Y. (2009). Collaborative filtering with temporal dynamics. In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (pp. 447–456).

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37.

Kumar, J., & Tintarev, N. (2018). Using visualizations to encourage blind-spot exploration. In *Intrs@ recsys* (pp. 53–60).

Lee, M. K., Kiesler, S., & Forlizzi, J. (2011). Mining behavioral economics to design persuasive technology for healthy choices. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 325–334).

Leonard, T. C. (2008). *Richard h. thaler, cass r. sunstein, nudge: Improving decisions about health, wealth, and happiness.* Springer.

Li, Q., Myaeng, S. H., & Kim, B. M. (2007). A probabilistic music recommender considering user opinions and audio features. *Information processing & management*, *43*(2), 473–487.

Liang, Y., & Willemsen, M. C. (2019). Personalized recommendations for music genre exploration. In *Proceedings of the 27th acm conference on user modeling, adaptation and personalization* (pp. 276–284).

McAlister, L., & Pessemier, E. (1982). Variety seeking behavior: An interdisciplinary review. *Journal of Consumer research*, *9*(3), 311–322.

McKenzie, C. R., Liersch, M. J., & Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. *Psychological Science*, *17*(5), 414–420.

Meeuwisse, T. (2019). Effects of visualizing recommendation for music genre exploration on understandability and helpfulness.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS one*, *9*(2), e89642.

Noggle, R. (2018). The ethics of manipulation.

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*.

Robinson, J. (2014). Likert scale. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 3620–3621). Dordrecht: Springer Netherlands. Retrieved from `https://doi.org/10.1007/978-94-007-0753-5_1654` doi: 10.1007/978-94-007-0753-5_1654

Taramigkou, M., Bothos, E., Christidis, K., Apostolou, D., & Mentzas, G. (2013). Escape the bubble: Guided exploration of music preferences for serendipity and novelty. In *Proceedings of the 7th acm conference on recommender systems* (pp. 335–338).

Thaler, R. H. (2015, Oct). *The power of nudges, for good and bad.* The New York Times. Retrieved from `https://www.nytimes.com/2015/11/01/upshot/the-power-of-nudges-for-good-and-bad.html`

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness.* Penguin.

Tintarev, N., Rostami, S., & Smyth, B. (2018). Knowing the unknown: visualising consumption blind-spots in recommender systems. In *Proceedings of the 33rd annual acm symposium on applied computing* (pp. 1396–1399).

Weinmann, M., Schneider, C., & Vom Brocke, J. (2016). Digital nudg-

ing. *Business & Information Systems Engineering*, *58*(6), 433–436.

Wendel, S. (2020). *Designing for behavior change: Applying psychology and behavioral economics*. " O'Reilly Media, Inc.".

Zhang, Y. C., Séaghdha, D. Ó., Quercia, D., & Jambor, T. (2012). Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth acm international conference on web search and data mining* (pp. 13–22).