Explanations in Music Recommender Systems in a Mobile Setting

Alexandra Kimberly Bobrow
Under the advisement of Assoc. Prof. DI Dr. Christoph Trattner

Master of Information Science
Social Sciences Faculty
University of Bergen

Delivery Date: August 7th 2020

Abstract

Every day, millions of users utilize their mobile phones to access music streaming services such as Spotify. However, these ‘black boxes’ seldom provide adequate explanations for their music recommendations. A systematic literature review revealed that there is a strong relationship between moods and music, and that explanations and interface design choices can affect how people perceive recommendations just as much as algorithm accuracy. However, little seems to be known about how to apply user-centric design approaches, which exploit affective information to present explanations, to mobile devices. In order to bridge these gaps, the work of Andjelkovic, Parra, & O’Donovan (2019) was extended upon and applied as non-interactive designs in a mobile setting. Three separate Amazon Mechanical Turk studies asked participants to compare the same three interface designs: baseline, textual, and visual (n=178). Each survey displayed a different playlist with either low, medium, or high music popularity. Results indicate that music familiarity may or may not influence the need for explanations, but explanations are important to users. Both explanatory designs fared equally better than the baseline, and the use of affective information may help systems become more efficient, transparent, trustworthy, and satisfactory. Overall, there does not seem to be a ‘one design fits all’ solution for explanations in a mobile setting.

Acknowledgements

When I first moved to Norway just after receiving a bachelor's degree in Interior Design, I never imagined going back to University. Two years later, after passing the Norwegian language proficiency test, I embarked on another academic journey. It has been seven years since I moved to Norway. I've completed a second bachelor's degree in Information Science and, with the delivery of this thesis, a master's degree in Information Science.

While planning my thesis, I ended up changing both my topic and research method. My mentor Dr. Christoph Trattner stood by me through all my frustrations. Thank you Christoph, for putting up with my procrastination, lack of time management, and sometimes unnecessary need for perfection. I appreciate that you always gave me your honest opinion about my work. Your guidance helped steer me in the right direction whenever I felt like I was veering off the path. You believed in me and pushed me to do better.

I would like to thank Alain Starke, who joined Christoph and myself during our meetings on several occasions. I am grateful that you generously took time out of your busy schedule to answer my questions. Your additional feedback was always very valuable as it helped me gain a different perspective on certain aspects of my work. A big thank you to the group for Intelligent Information Systems (I2S) at the University of Bergen. I appreciate the comments you all gave during my mid-term presentation. Whenever I attended a meeting, I always enjoyed the stimulating conversations that took place afterwards.

A special thanks to all my fellow students, friends, family, and extended family for attempting to understand my thesis topic in addition to patiently listening to me discuss my frustrations.Some of you offered to participate in my user studies, others helped me distribute my surveys. It is imperative to mention how grateful I am to Ida Sekanina and Torstein Thune who agreed to read my thesis and gave me constructive criticism on how to better it. Thank you both for making an effort to help me out.

Furthermore, I want to thank Torstein Thune for his continuous support and encouragement. Thank you for letting me stay up late working, never getting too upset if I didn't come home until five in the morning. Thank you for allowing me to brainstorm new ideas with you. Thank you for constantly being my guinea pig, and agreeing, sometimes reluctantly, to review and test all my prototype designs. Thank you for being my rock when the world erupted into chaos and distractions overwhelmed my ability to concentrate. There are countless reasons to thank you, but most importantly, thank you for always reminding me what I am capable of.

Lastly, I must also acknowledge the continual support of my best listener, Luna. It's hard to imagine completing this without your companionship. Though there were many times when you were the sole reason for my procrastination, your unconditional affection always lifted my spirits.

Dedication

I dedicate this thesis to all those who have not had the opportunity to complete studies in higher education like I have. No one should ever be denied education due to their ethnicity, gender, religious affiliation, social or economic standing, sexual orientation or otherwise. May the protests of 2020 help guide the powers that be towards equality in education and society as a whole.

Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 Motivation and Research Context

Recommender systems (RSs) alleviate the problem of information overload through content filtration. Improving systems by enhancing recommendation algorithms has previously controlled much of the discourse around research topics in this field. However, by aiming sole attention at statistical accuracy, other elements which also have direct implications on how recommendations are perceived may be overlooked. According to Swearingen & Sinha (2001) and Ricci et al. (2015), a user's experience with a recommender system is actually not confined to whether or not they get accurate recommendations. According to Åman & Liikkanen (2010, p. 1), recommendation aids such as *"explanations, interactive elements, and visualizations"* have been shown to greatly improve user experience. Millecamp et al. (2019) asserts that due to the 'black box' nature of recommender system, explanations should be used to increase a user's understanding of a system and its music recommendations. Bostandjiev et al. (2012), Millecamp et al. (2019), and Andjelkovic, Parra, & O'Donovan (2019) are examples of recent studies within the music domain where elaborate explanatory interfaces were developed to tackle issues which can occur in algorithm focused ones such as of lack of acceptance, transparency, usability, and usefulness. According to Nunes & Jannach (2017), researchers are constantly experimenting with what types of explanations should be presented, where when and why they are presented, how they are presented, how they are generated, and how to evaluate them. Visualizing explanations show promising results over text only explanations as far back as Herlocker et al. (2000), where explanatory graphs were preferred over explanatory sentances. Chen & Pu (2013), Kamalzadeh et al. (2016) and Andjelkovic, Parra, & O'Donovan (2019) have experimented with affect-based explanations, an alternative datatype to genres, which also show positive results. Explanations should be used whenever necessary for the elucidation of recommendations, but the design approach to this is currently unclear, particularly when it comes to mobile devices. Issues such as balancing low cognitive load and providing enough information can be especially challenging when designing for smaller screens. If previous research shows that explanations can improve recommender systems in a variety of different domains with larger screens, there is clearly a need to better understand the potential explanations have on improving user experience in the music domain on mobile devices.

With the role of connected devices steadily increasing, they are becoming an integral part of the listening experience. As reported by IFPI (2018, p. 5), *"75% of consumers use smartphones to listen to music"*. This number jumps to 94% for users between the ages of 16 and 24. According to Spotify AB (2020), as of March 31st 2020, Spotify is the world's leading music

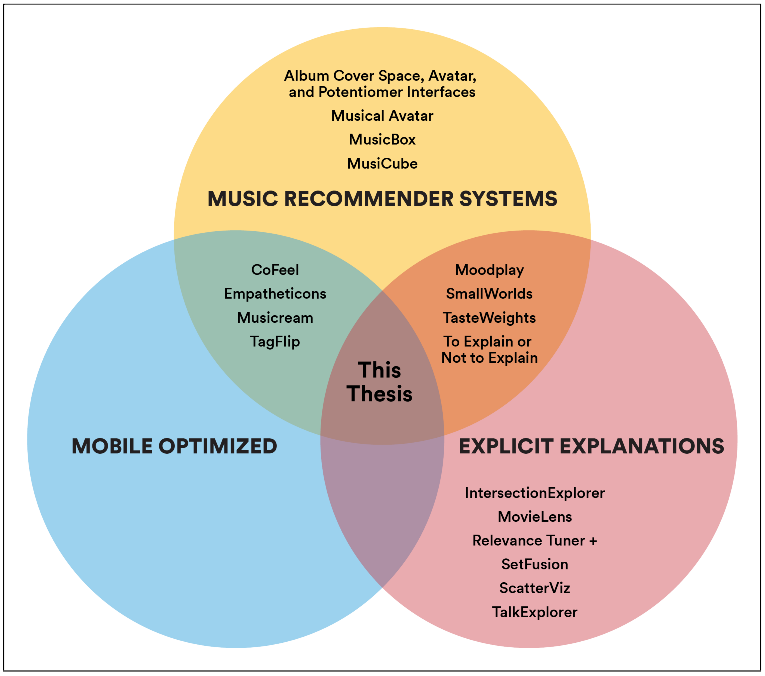


Figure 1: Contextualization of this master's thesis.

streaming service which is exclusively audio-based having around 130 million Spotify Premium users in addition to 156 million Spotify Free users. The business article by Iqbal (2020) indicates that almost 60% of users world-wide primarily listen to Spotify on mobile devices. In terms of design, Spotify's mobile application unfortunately contains noticeably less details and explanations than their desktop application. It seems as though they think they must cut content in order to accommodate smaller screens. As there is not necessarily a correlation between usage and user satisfaction, the upwards trend of people using smartphones to listen to music calls for the additional academic research of mobile interface design for MRSs in general, as well as in the specific context of explanations.

1.2 Problem

The problem domain of this master's thesis is explanations in recommender systems, with a specific focus in a mobile setting. Figure 1 displays the overlap of the most relevant literature pertaining to explanatory interfaces and MRSs, including those which were tested on smaller screens. Searching through pertinent literature accrued no articles which present established design guidelines specifically for music recommender system interfaces to address the presentation of recommendation explanations for mobile devices. Hardly any previous studies have optimized their music recommender system interfaces for handheld devices. The goal of this thesis is to bridge these gaps by investigating what types of affective music explanations are appropriate for smaller screens in addition to examining their importance from a user's perspective. In broad terms, the problem statement for this thesis is:

Making recommendations understandable through explanations in a mobile setting

1.3 Research Questions

This master's thesis looks specifically into what design elements of mood-based music recommendation explanations valued most by users are best suited for mobile devices. Throughout this thesis, the current state-of-the-art and user studies are reviewed and evaluated in order to address the following research questions (RQs):

RQ1: To what extent are explanations in mobile music recommender applications valued by users, and how does music familiarity affect that?

RQ2: How do users evaluate the different design elements of music recommendation explanations in a mobile setting?

RQ3: To what extent do users prefer these explanations to be either textual or visual?

RQ4: To what extent do affect-based explanations influence a user's perception of the system and its music recommendations?

1.4 Contribution

Previous literature has verified that people actually want more explanations in music recommender systems. This thesis first reexamines this notion by completing a systematic literature review. The topic of explanatory interfaces is becoming more prominent in today's research but current academic and commercial approaches to explanations are either too complicated or simple for the average user. The studies completed in this research address how to design music recommendation explanations when being displayed on a mobile device, and user perceptions of explanations both in general, and in terms of music familiarity. Two main design science research artifacts were produced through this research: a literature review and two new explanatory interface designs for mobile devices. Insight is provided into users' personal preferences and perceptions of different explanation designs by providing a statistical comparison of the two novel interfaces against a baseline. Implications of this study show that explanations in a mobile setting may lead to higher or more efficiency, user satisfaction, transparency, trust, and/or use intention.

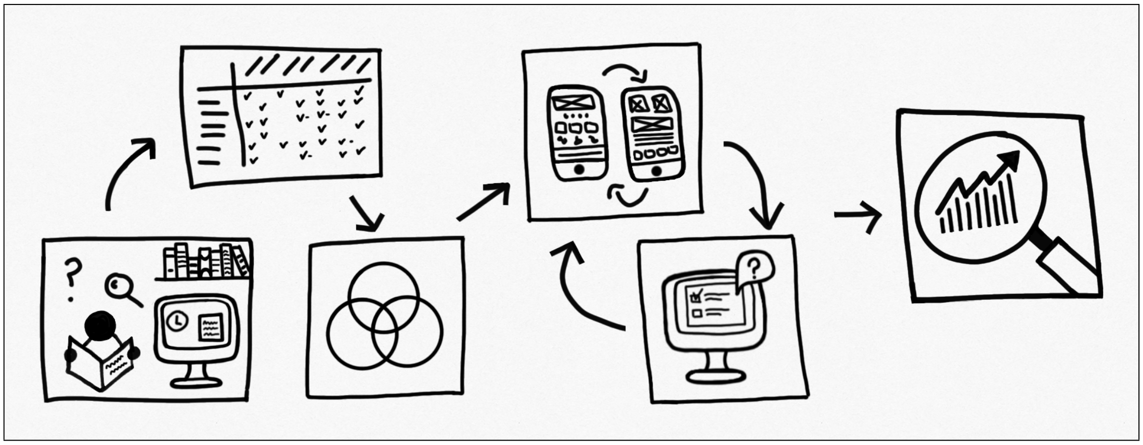


Figure 2: Thesis synthesis.

1.5 Thesis Outline

This paper consists of 5 Chapters: Introduction 1, Background 2, Methods 3, Results 4, Summary and Conclusion 5. These discuss the domain of recommender systems, the methodology behind the design science artifacts, along with an assessment and discussion of the user studies. The research was carried out as shown in Figure 2. First, a thorough literature review was conducted in order to create an organized overview of the current state-of-the-art in the form of a table and venn-diagram, Table 3 and Figure 1 respectively. Kitchenham & Charters (2007, p. 3) explains a systematic literature review as being *"a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest"*. This process is discussed in more detail in B. Based on the literature, a series of different prototype designs and questionnaires were created iteratively in accordance with Design Science principals. The focus then turns to the evaluation and analysis of the findings from the user testing. The conclusion includes a clear presentation of the direct implications these results have on the research questions along with ideas for future exploration.

2 BACKGROUND

The current state of the art in music recommender systems, recommender systems for mobile devices, and recommendation explanations, is discussed here. Section 2.1 touches upon algorithms, interface design tools and frameworks, and interactive music recommender systems. Section 2.2 is an extension of the first, but focuses specifically on mobile interface design for music recommender systems. Section 2.3 examines specific ways of visually representing recommendation explanations in the recommender system domain generally. Section 2.4 provides a synopsis of this Chapter, accompanied by a detailed overview of this information laid out in Table 3. The works mentioned here are meant to provide structure and lay the groundwork for new research pursuits, such as the study explained in this thesis.

2.1 Music Recommender Systems

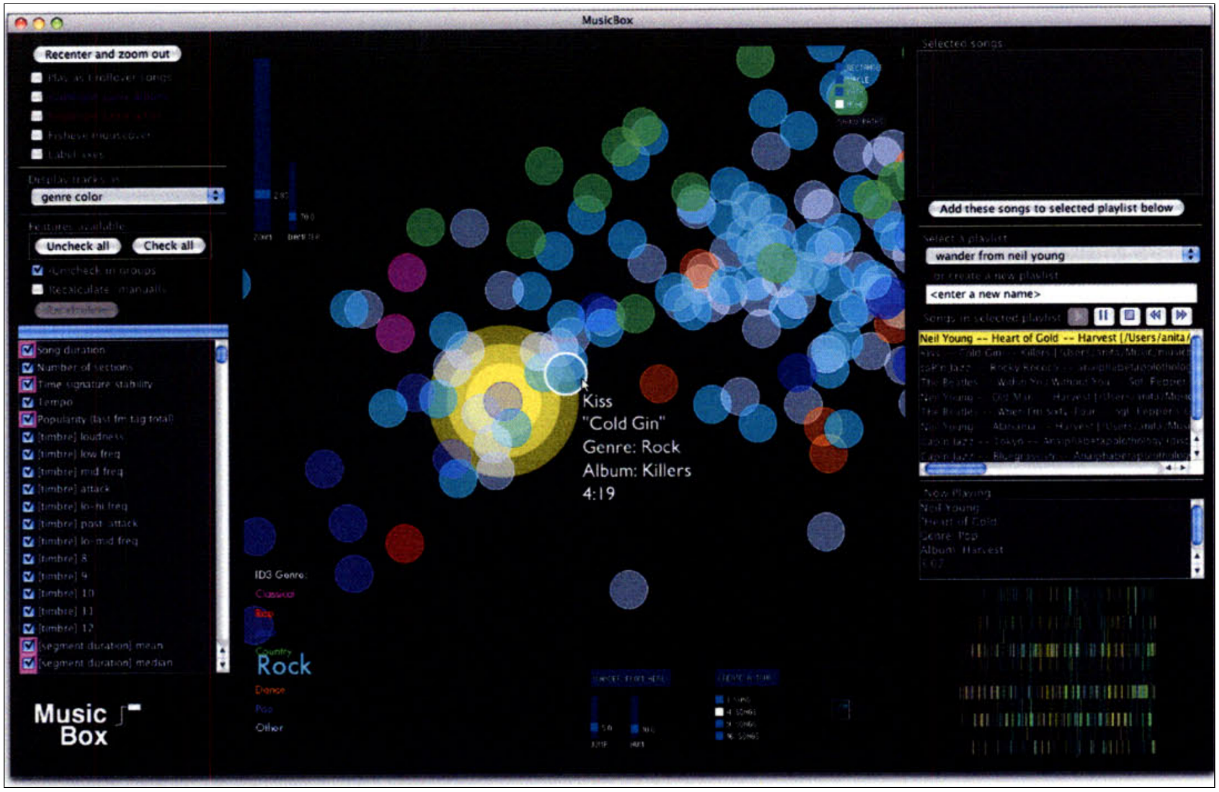
The origin of recommender systems stems from the idea of creating a system which resembles the common human behavior of asking others for assistance in daily decision-making tasks. Ricci et al. (2015, p. 4) defines RSs as *"software tools and techniques"* that aid users in a variety of different decision-making processes by providing *"suggestions for items that are most likely of interest"* to them. According to Ricci et al. (2015, p. vii), this field is composed of a mixture of *"artificial intelligence, human computer interaction, data mining, statistics, decision support systems, marketing, and consumer behavior"*, and only became an independent field of its own in the mid-1990s. When the e-commerce boom occurred, RSs helped lighten peoples burden of trying to navigate the vast amount of information available in the ever-growing world wide web. Today, RSs are vital to web giants like Amazon, Facebook, and Google in addition to popular streaming services such as Apple Music, Spotify, and YouTube.

Ricci et al. (2015), proceeding from Burke (2007), denotes that the main classes of recommendation techniques are collaborative filtering, content-based, demographic, and knowledge-based. It is also becoming more and more common to create hybrid recommender systems which consist of two or more of these techniques. Content-based systems base their recommendations off of items users have previous liked by comparing these items to find other similar items. Collaborative filtering compares users in the same fashion, rather than items, by recommending liked items from other users with similar taste profiles. Demographic RSs make recommendations without needing a long user history by comparing a user's demographics, such as age or gender, to other users with similar backgrounds. As stated in Ricci et al. (2015, p. 13), knowledge-based RSs utilize *"specific domain knowledge about how certain item features meet users' needs and preferences"*. All of these approaches can be used to generate either generalized or personalized suggestions and can give either serendipitous or similar recommendations. Jannach et al. (2011) articulates that personalized recommendations require explicit or implicit user data in order to create a user model or user profile from which the predictions can be based off of. It can be difficult to make accurate predictions when data for a new item or new user is unknown, which is referred to as the cold-start problem. Outside of the main types of RSs, context-based, personality-based, emotion-based, and cross-domain, among others, are being more and more frequently used to create more personalized recommendations.

Music Recommender Systems are typically directed at making tailored predictions not only about one specific item such as a song, but also groups of items such as an album or playlist which contains several songs. In the music domain, Ricci et al. (2015) describes how content-based item to item recommendation approaches are often favored, as information descriptors about items are usually more readily available than explicit user feedback. MRSs therefore rely heavily on the field of Music Information Retrieval (MIR); a field which obtains semantic data about music through either meta-data or audio content. Types of meta-data used include web mined keywords, human made annotations, or social tags. Audio content data consists of acoustic and musical features such as timbre, tempo, or musical key. In addition to what recommendation algorithm is used to generate suggestions, Ricci et al. (2015) also stresses that the design of a RS and how its graphical user interface looks are equally important factors for how useful and effective a system is. These aspects can play a significant role in how users perceive a RS. Music meta-data is not only used in algorithms but can also be visibly presented with text-based views, visual representations, or in combination. Some examples of different visualizations include graphs, icons, maps, node-link diagrams, radial views, sliders, or tables. Visual representations of music may actually contribute to better recommendation explanations. This thesis puts recommendation algorithms aside, and rather aims attention at design-oriented aspects of MRSs in the area of explanations. The following overview of related works build upon these themes.

2.1.1 MusicBox

MusicBox is described by Lillie (2008) as being an innovative music recommendation application for maneuvering through large music libraries which focuses on song navigation, exploration, and recommendation. The idea behind this thesis was to create a program which diverged from the standard, static, text heavy system, which primarily bases recommendations on artist genre generalizations. Artists may produce a broad range of songs with different sound properties. This means that a single album or single artist usually does not confine to a specific genre. This alternative application therefore compares features of individual songs to define their similarity ratings. MusicBox allows users to choose which content-based and context-based descriptors,



genre color

☑ uncheck in groups
□ Recalculate manually

☑ Song duration
☑ Number of sections
☑ Time signature stability
☑ Tempo
☑ Popularity (last fm tag total)
☑ [timbre] loudness
☑ [timbre] low freq
☑ [timbre] mid freq
☑ [timbre] attack
☑ [timbre] lo-hi freq
☑ [timbre] post attack
☑ [timbre] lo-mid freq
☑ [timbre] 8
☑ [timbre] 9
☑ [timbre] 10
☑ [timbre] 11
☑ [timbre] 12
☑ [segment duration] mean
☑ [segment duration] median

Kiss
"Cold Gin"
Genre: Rock
Album: Killers
4:19

wander from neil young

<enter a new name>

Neil Young -- Heart of Gold -- Harvest [/Users/anita/

Neil Young
Heart of Gold
Genre: Pop
Album: Harvest

Figure 3: MusicBox interface from Lillie (2008).

such as mood or time signature stability, should influence the song recommendations they receive, as opposed to just genre. The interface design is interactive, where songs are represented by different colored circles and shown in a two-dimensional visual space.

The interface, shown in Figure 3, is broken up into 5 main components. Interaction tools and feature controls are on the left. The song visualization area is in the middle, where similar songs are shown closer together and dissimilar songs presented further apart. When users add songs to their playlist, they appear on the right-hand side along with basic song information for the song that is currently playing. MusicBox expresses another means to visualizing an entire music library, albeit with a limited sample music library of just between 140 - 500 songs.

MusicBox was tested with 10 people, but testing only really took advantage of the available content-based features and not context-based. It mainly compared iTunes playlist creation to MusicBox playlist creation. All participants were fairly familiar with the music players, although one had never used iTunes before. Users took advantage of the similar artists spaces feature for artists they liked, but also avoided these spaces for musicians they did not like. Ease of use was not as positive as the author had hoped.Since MusicBox gives users many options on how to visualize their music library, which in turn affects their recommendations, 1 hour for using testing may not have been enough time for users to become familiar with the system or get a full enough sense of the program. Despite given more control in the music space with specific music features to choose from to explore music, due to the diverse song base in this serendipitous program, such descriptors seemed unclear to users. It seems users wanted more control rather through the ability to edit song information or by having the option to increase or decrease feature importance.

Some users could see patterns emerging among similar songs but could not articulate or explain the connection between choosing certain descriptors and receiving certain recommendations. This indicates that while searching for songs based on meta-data other than genre may help with music discovery, a lack of explanations can make the interface confusing. Lillie (2008) concludes that it is positive that MusicBox helped users to broaden their perspective about music recommendations by changing their expectations. When confronted with what the user perceived as more serendipitous recommendations than they were possibly used to, Lillie (2008, p. 104) explains that they *"attempted to explain this discrepancy by adjusting their own [mental] model"* about musical relationships. However, Lillie also constructively notes that seeing attributes or song titles etc., will never be enough to guess what a song will sound like. Adding a visual space helped users understand the relationship between songs, but only to a certain extent, as this information overload did not enable users to interpret explanations clearly enough.

2.1.2 Avatars, Potentiometers, and Album Covers

Jukka Holm published 9 articles and developed 6 interactive high-fidelity interface prototypes for computers in his PHD thesis *"Visualizing Music Collections Based on Metadata: Concepts, User Studies and Design Implications"* (Holm, 2012). This was in collaboration with Nokia Research Center and Tampere University in Finland. Holm's subsequent publication, *"Designing for Music Discovery: Evaluation and Comparison of Five Music Player Prototypes"* (Lehtiniemi & Holm, 2013), reveals an overview of the main graphical user interface (GUI) prototypes tested. According to Lehtiniemi & Holm, the interfaces ranked from most promising to least promising are as follows: album cover-based design with clip previews, avatar-based, potentiometer-based, animated mood pictures, cube-based, and a virtual world where buildings and characters represent music genres. The top three of prototypes will be the focus of this section. Each GUI listed above was tested individually by 40 Finnish participants and documented in previous research articles. Lehtiniemi & Holm did a comparison study where 40 new participants tested the first five GUIs mentioned above over the course of two to three weeks in order to better compare them in the publication mentioned above. These studies examined the effects of music discovery without textual search fields, but rather through visually pleasing interfaces. All were qualitative and quantitative, where participants were interviewed, observed, and asked to complete several surveys.

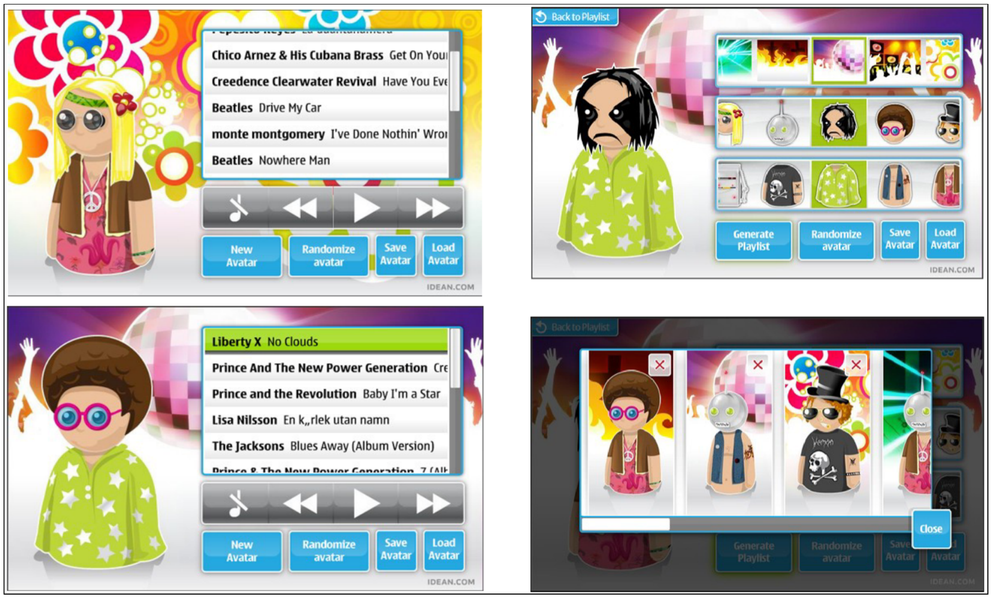
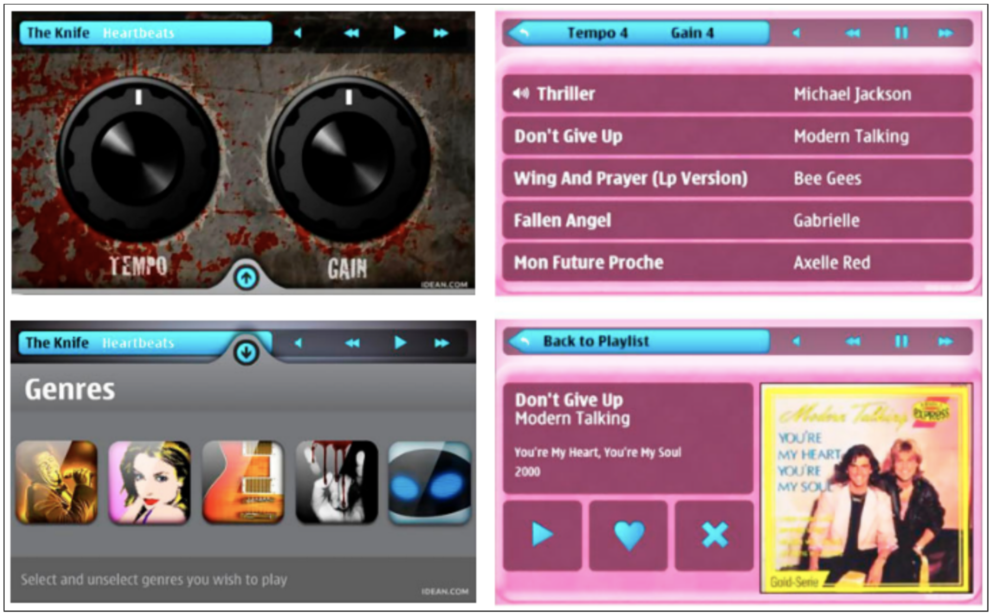

Chico Arnez & His Cubana Brass Get On You
Creedence Clearwater Revival Have You Eve
Beatles Drive My Car
monte montgomery I've Done Nothin' Wro
Beatles Nowhere Man
New Avatar
Randomize avatar
Save Avatar
Load Avatar
IDEAN.COM
Back to Playlist
Generate Playlist
Randomize avatar
Save Avatar
Load Avatar
IDEAN.COM
Liberty X No Clouds
Prince And The New Power Generation Cre
Prince and the Revolution Baby I'm a Star
Lisa Nilsson En k„rlek utan namn
The Jacksons Blues Away (Album Version)
New Avatar
Randomize avatar
Save Avatar
Load Avatar
IDEAN.COM
Close

Figure 4: Avatar-based interface from Holm et al. (2010).

Prior to designing the high-fidelity prototypes, a series of different experiments were completed to determine the best visualizations to use. Tempo and release year were two aspects which were analyzed in Holm (2012, p. i), *"in addition to... five different visualization methods (colors, icons, fonts, emoticons and avatars) for representing musical genres"*. Holm (2012, p. 5) discovered that *"it is not possible to design a globally-accepted color-genre mapping... as the associations between colors and genres are highly subjective and... results indicate that colors alone are not a good general solution for visualizing musical genres"*. Fonts gave slightly better results, seeming like a promising alternative. However, it can be difficult to find the exact fonts needed to accurately represents certain genres. Icons were very promising as the majority of users easily and correctly perceived what icons matched which musical genres.

The findings from the color and icon studies gave a baseline for how to create avatars. The avatar application, shown in Figure 4, automatically creates playlists for music discovery based on avatars. The user can either build an avatar themselves or chose one of the randomly generated avatars. Avatars consist of a head, a body, and a background image. Each part corresponds to one of 5 musical genres Metal, Rock, Electronica, Dance, Hippie, Soul and Disco. That's a total of 125 possible avatar combinations. Results showed that the stereotypical avatars achieved even slightly better results than the icons, however the avatars took up the majority of screen real estate. The author points out that while large detailed visual explanations are helpful to convey more information, they may not be well suited for smaller mobile devices.

In line with Avatars came the Potentiometer-based user interface shown in Figure 5. The top left image displays the start screen. Here, users can turn the circular knobs to change the tempo



Thriller	Michael Jackson
Don't Give Up	Modern Talking
Wing And Prayer (Lp Version)	Bee Gees
Fallen Angel	Gabrielle
Mon Future Proche	Axelle Red

Select and unselect genres you wish to play

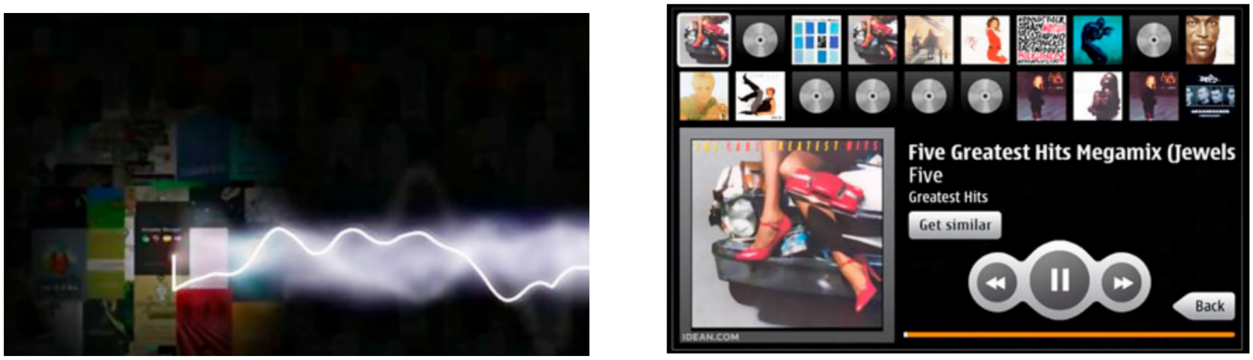
Don't Give Up
Modern Talking
You're My Heart, You're My Soul
2000

Figure 5: Potentiometer-based interface from Lehtiniemi & Holm (2013).

or gain. Turning the tempo knob left correlates to receiving a song with a slower tempo, while turning it to the right presents the user with a song that has a faster beat. The gain knob worked correspondingly, but rather controlled the energy level and aggressiveness of a recommended song. Alternatively, users can click the arrow up button, in the middle on the bottom, in order to choose one or more genres. The application changes its color and style depending upon the genre of music which is currently playing. The genre screen is shown in the bottom left image, where uses click the arrow down button to once again reveal the home screen.

Results show that the potentiometer-based interface is satisfactory for music discovery. Holm (2012, p. 53) state that it is *"innovative, handy, easy to use, and entertaining, and the graphical designs matched well with the musical genres"*, however, it scored slightly worse than the avatar interface. This may be in part due to the fact that users noted that it was too simplistic for continual use, or even just as an additional tool for a standard music player. Feedback suggested that people would be more receptive to this tool if there were a greater amount of specific options for calibrating song recommendations as well as the possibility to search by text.

In the album cover space interface, a user is presented with a wide variety of album covers. Upon clicking a cover, a short snippet of one the songs from that particular album is played. The audio clip is visually represented in a manner similar to an audio-wave, as shown in Figure 6a. Playlists are represented as collections of album covers showing what music is up next. Users can click the 'Get similar' button to retrieve similar music corresponding to the current playing song, as shown in Figure 6b. While the use of interactivity in this model is novel, the ability to preview songs is not. Music imagery first began with album covers, so it is quite reasonable to say that



(a) Choose an album to preview music.

(b) Playlist view.

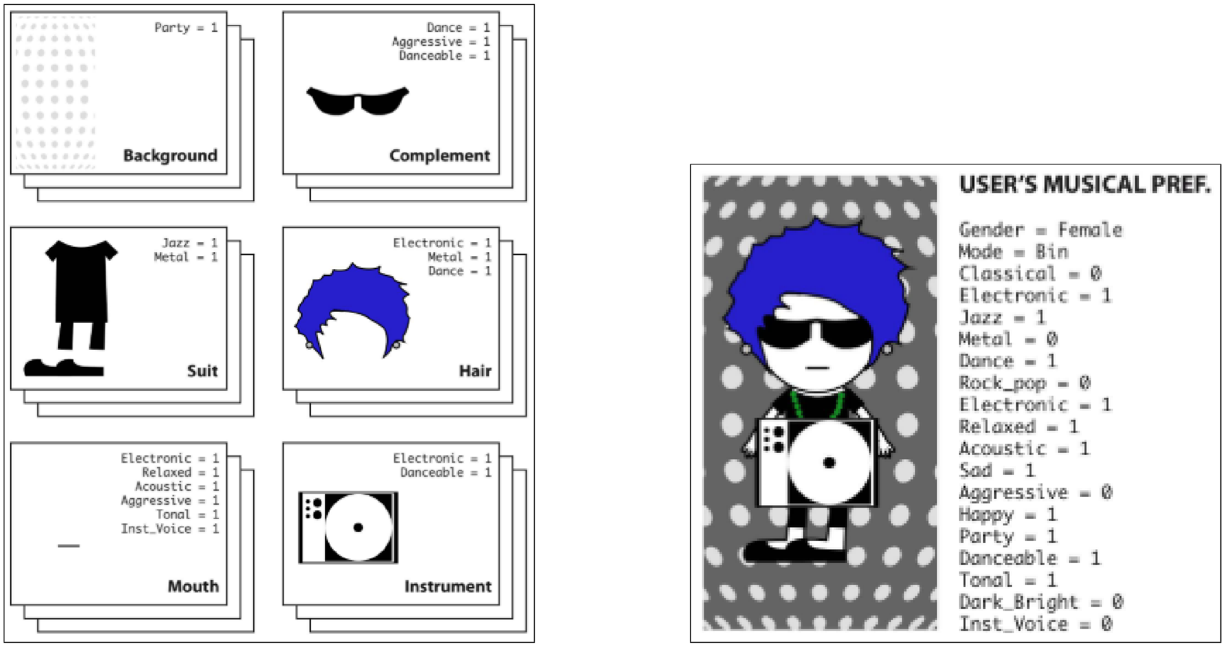
Figure 6: Album cover space interface from Lehtiniemi & Holm (2011).

this is a branch standard. This approach is slightly simpler which is easier to use and also good for music discovery. Users preferred the interactive album cover space to all the other interfaces which exploited different types of musical meta-data. However, while being a familiar and popular interface design choice, users complained that the absence of additional features makes it difficult for them to understand the similarity relationship between albums. In other words, there was a lack of explanations.

The authors posited that the visual interfaces presented would be considered user-friendly and innovative. This hypothesis held true as the interfaces were well received from the majority of users. The results show that visual interactive GUIs help with user engagement as users proactively searched for new music during user testing on their own accord. However, unless specifically designing for music discovery, it would otherwise not be beneficial to remove all textual features as this limits users' ability to search for specific items.

2.1.3 Interactive Music Recommender Systems Survey

A recent survey on interactive recommender systems, He et al. (2016), contained an in-depth analysis of twenty-four different interactive recommender systems, of which seven were from the music domain. Overall, a majority of the twenty-four systems represented data relationships with node-link diagrams or radial visualizations. Additional visualization techniques used were set-based visualizations, icons, flow charts, tables, and scatter plot. Half of these systems had a visualization objective of transparency and/or controllability, a third explored explanations and justifications, while only two experimented with context. Almost all looked at the metrics effectiveness and or usefulness, followed by usability. A few looked at trust, satisfaction and/or engagement, while only two looked at efficiency. In order to evaluate their systems, many chose to compare them to a baseline system. The following will provide an overview of the music



USER'S MUSICAL PREF.

Gender = Female
Mode = Bin
Classical = 0
Electronic = 1
Jazz = 1
Metal = 0
Dance = 1
Rock_pop = 0
Electronic = 1
Relaxed = 1
Acoustic = 1
Sad = 1
Aggressive = 0
Happy = 1
Party = 1
Danceable = 1
Tonal = 1
Dark_Bright = 0
Inst_Voice = 0

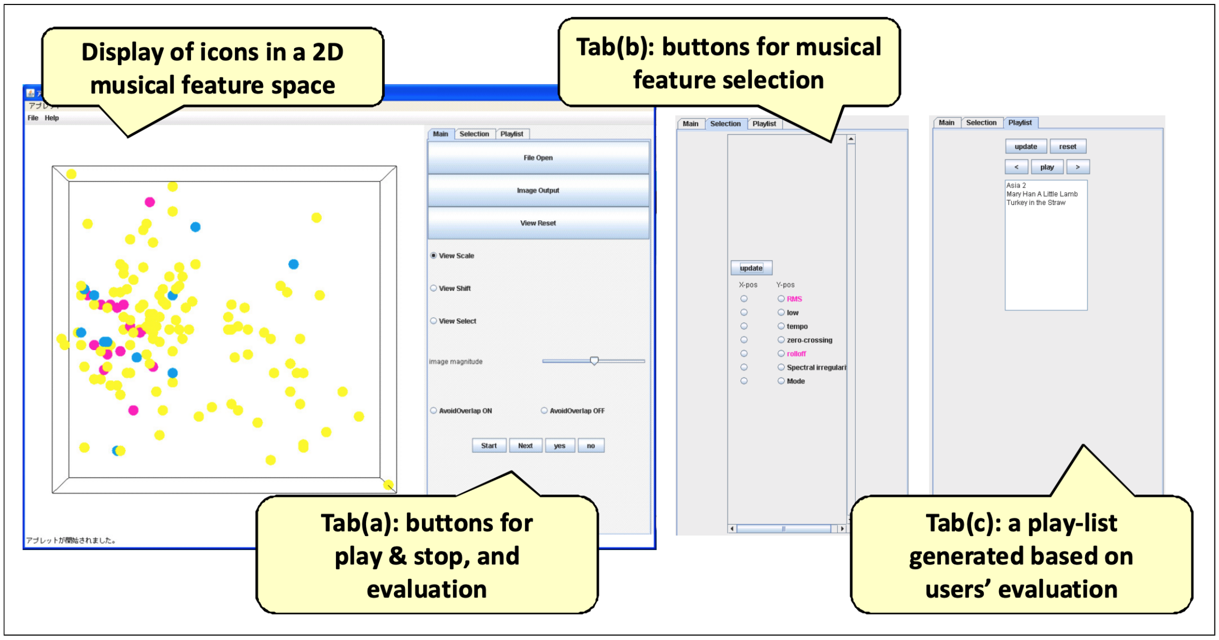
(a) (b)

Figure 7: Musical Avatar interface from Bogdanov et al. (2013).

recommender systems which were tested. They are as follows: CoFeel from Chen & Pu (2013), Empatheticons from Chen et al. (2014), Musical Avatar from Bogdanov et al. (2013), MusiCube from Saito & Itoh (2011), SFViz from Guo et al. (2011), SmallWorlds from Gretarsson et al. (2010), and TasteWeights from Bostandjiev et al. (2012).

CoFeel and Empatheticons were designed for mobile devices and will therefore be explained in more detail in Section 2.2 Mobile Music Interfaces. Seen in Figure 14, they were the only interfaces to have the visualization objective of emotional context in a social group setting. TasteWeights, due to being an explanatory interface, is mentioned in greater depth in Section 2.3 Explanations in Recommender Systems. Seen in Figure 20, this node-link diagram lets users control their music recommendations through contextual data from Wikipedia, Facebook, and Twitter. In addition, this survey discusses the explanatory interface SetFusion by Parra & Brusilovsky (2015), which is also is pertinent literature. However, it has been excluded from this particular section as it is not a MRS, but is brought up in Section 2.3 too.

Musical Avatar and MusiCube both focus on how to justify recommendations and use content-based recommendations. Musical Avatar, as implied in its name, accomplishes this by the use of avatars, using the icon visualization technique. The avatars, which Bogdanov et al. (2013, p. 25) refers to as *"humanoid cartoon-like characters"*, were created by taking semantic descriptors from songs such as musical genres, moods, and instrumentation, and mapping those to different visual aspects of persons style as shown in Figure 7. Figure 7a shows a breakdown of an avatar which uses



Asia 2
Mary Han A Little Lamb
Turkey in the Straw

Figure 8: MusiCube interface from Saito & Itoh (2011).

a binary strategy, where 0 represents no traits and 1 represents one or more traits. So the absence of traits was also seen as valuable information to use when creating a final avatar. Bogdanov et al. (2013, p. 28) explains how the *"descriptor values influence the selection of the different graphic elements used to construct the avatar"*, as shown in Figure 7b. MusiCube also concerns itself with controllability. The visualization method used is a scatterplot, where the X and Y axis represent different musical features, which produce tune recommendations in the form of colored circles, as shown in Figure 8.

The objectives of SFViz, SmallWorlds, and TasteWeights were transparency and controllability. SFViz is a spatial interface for exploring social recommendations using collaborative filtering. Figure 9 is an example of their interface showing top recommended friends. SFViz is composed of a hierarchy of music tags from Last.fm, where users are categorized within specified tag groupings. Instead of user testing, the authors opted for presenting use-cases and was therefore not evaluated further in this interaction survey. SmallWorlds is an interactive explanation interface, also for social recommendations. Gretarsson et al. (2010) compared three different node-link diagram interfaces as shown in in Figure 10. They tested graph-based, tree, and concentric layouts in a Facebook application, where the tree worked best. Items being recommended were either books, movies, or music. In each interface, nodes are color-coded and linked together by lines where larger nodes correlate to a closer neighbor relation. The user's avatar node is a light green color, while other user profile nodes are slightly darker. Similar friends are shown in blue, while dissimilar friends and items are orange. Recommended items are yellow.

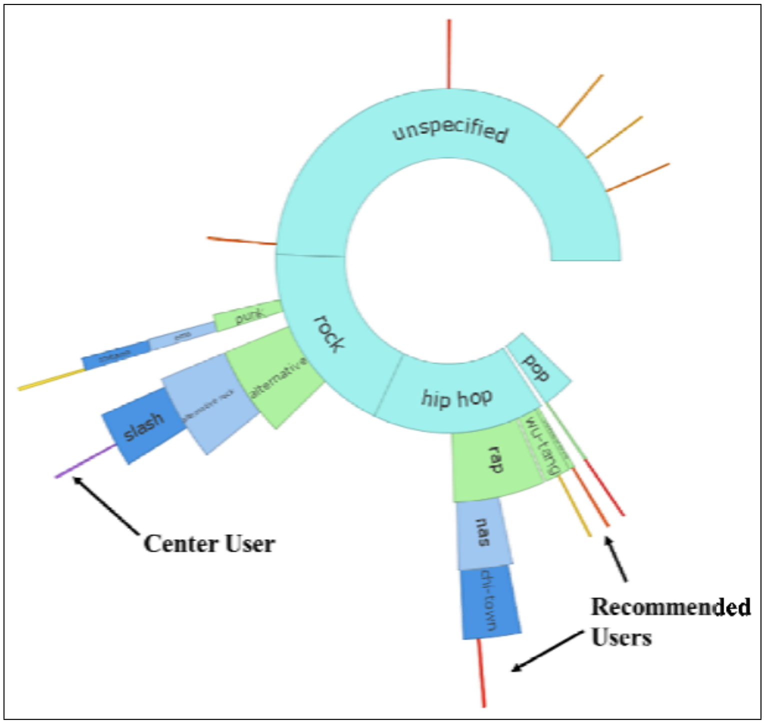
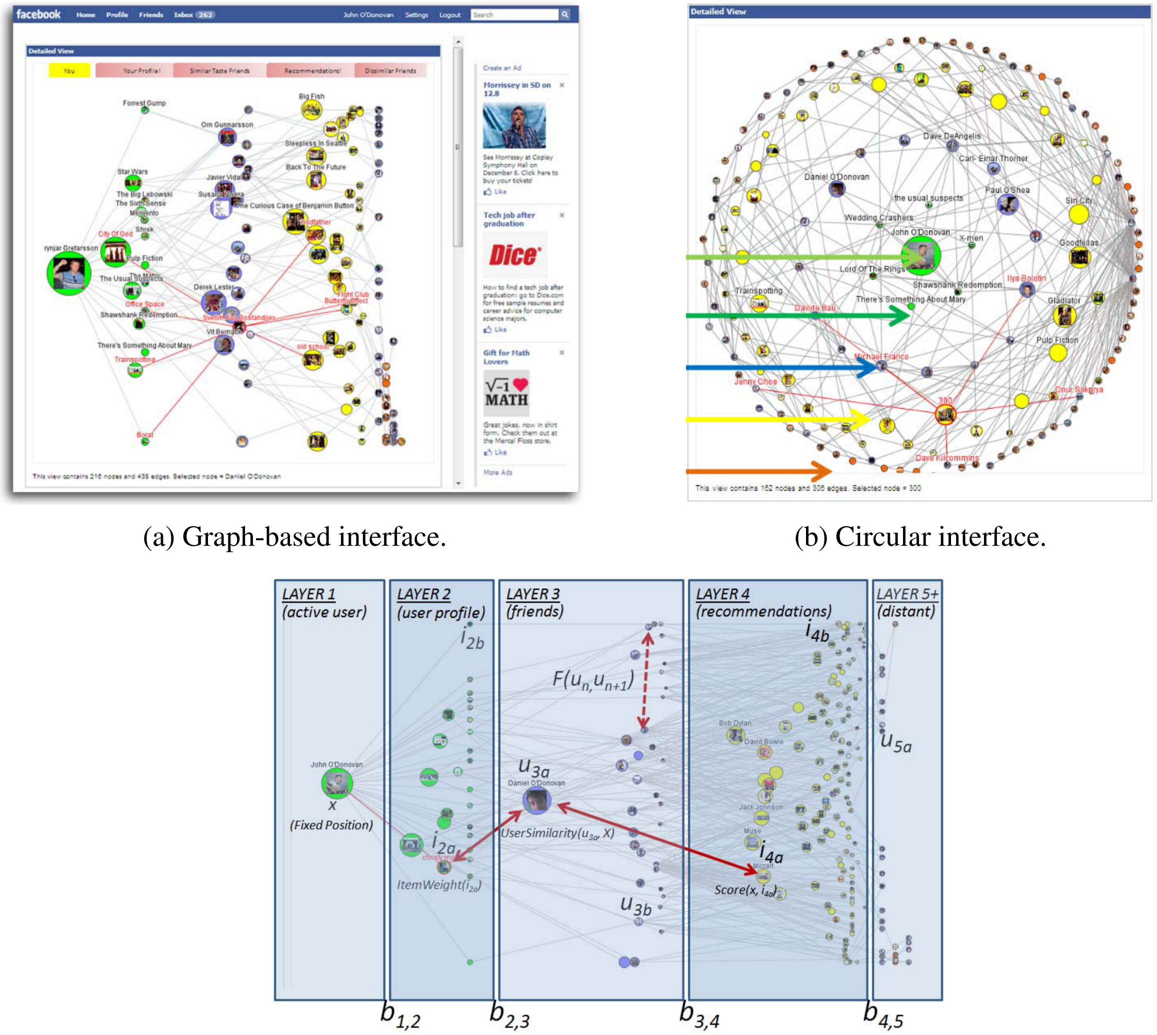


Figure 9: SFViz interface from Guo et al. (2011).

Unlike the other music recommender systems, SmallWorlds, Musical Avatar, and TasteWeights did not ask users to explore freely. The first two approached user testing by comparing different visualizations. The second two compared different recommender algorithms. In terms of metrics, all of the MRSs except for Musical Avatar, explored usability. All expect for MusiCube chose usefulness. Interestingly, CoFeel was the only system out of all systems evaluated in this survey, not just the MRSs, to not use questionnaires during user testing. TasteWeights, SmallWorlds, and MusiCube looked at effectiveness, which was determined by performing a recommendation accuracy test and task performance analysis. MusiCube, SmallWorlds, and TasteWeights took recommendation accuracy in account.

The results were overwhelmingly positive. User testing of all music recommender systems evaluated resulted in positive usefulness and usability feedback. SmallWorlds' evaluation of user satisfaction showed positive results. Users voiced that CoFeel and Empatheticons increased their level of engagement with the system. MusiCube, SmallWorlds, and TasteWeights proved that their systems increased user acceptance. There were mixed reviews on whether MusiCube's task performance was better or not, but other feedback declared that its visualizations increased user acceptance of the system. According to He et al. (2016), overall results show that explanatory interfaces can increase user trust and that explanations can improve user acceptance of recommendations. More control and better perceived user experience can also help with user trust, however too much control can lead to over-fitting. User control has an impact on



See Morrissey at Copley Symphony Hall on December 8. Click here to buy your tickets!

How to find a tech job after graduation: go to Dice.com for free sample resumes and career advice for computer science majors.

Great jokes, now in shirt form. Check them out at the Mental Floss store.

This view contains 216 nodes and 438 edges. Selected node = Daniel O'Donovan

(a) Graph-based interface.

This view contains 162 nodes and 306 edges. Selected node = 300

(b) Circular interface.

(c) Tree interface.

Figure 10: SmallWorlds interfaces from Gretarsson et al. (2010).

recommendation accuracy, and manual exploration is better when a user knows what they are looking for. In terms of design, visualizations help users understand the rationale of recommender systems. However, novice users may benefit from simpler graphics as advanced visualizations can be too complex. In addition, icons can sometimes be misleading. The findings suggest that there is no concrete evidence to theorize that there is one specific design type that is suitable for all users.

While one can make generalizations about groups of users, no two users are the same, so one should keep in mind user disparities when designing new interface visualizations. At the same time, psychological studies such as Rentfrow & Gosling (2007), Rentfrow et al. (2012), and Nave et al. (2018), have explained that peoples personalities and musical preferences can be generalized to a certain extent. It could be said that since stereotypes hold true for the majority, it is acceptable for music recommender systems to stereotype users in order to make recommendations. In the case of avatars, Bogdanov et al. (2013, pp. 25, 30) claims that their use for preference elicitation in music recommendation *"provides a reliable, albeit coarse, visual representation of the user's musical preferences"*. It can therefore be implied that avatars can help users understand the connection between music recommendations and the stereotypical avatars. Avatars can actually function as a type of recommendation explanation.

He et al. (2016, p. 25) suggests that new research could focus on creating custom and flexible systems which can adapt *"visualizations to the knowledge level and interest of user"*. Another direction is how to incorporate emotions and moods in recommender systems. He also expresses that further research should include testing on mobile devices, as multi-touch interactions have the potential to increase search accuracy and provide more efficient information filtering. This thesis has taken that feedback into account by covering both moods and designing for mobiles.

2.2 Mobile Music Interfaces

Before the digital music era, people have carried around radios, cassette players, compact-disk (CD) players etc. with them to listen to music on the go. The concept of a digital audio player gained popularity arguably due to the invention of Apple's iPod in 2001[1]. The iTunes store was launched soon after which was the first place where one could purchase digital music legally. This digital music library synced seamlessly with the iPod to provide users with an on-the-go music listening device. Today, the modern smart phone has for the most part replaced devices such as the iPod. In addition, music streaming services have also replaced the necessity to purchase both digital and physical music.

Technology has revolutionized the music industry yet designing for mobile music applications has not been a priority. As shown in Figure 1, there are hardly any academic works within the field of MRSs which have designed their interfaces for smaller screens in addition to actually evaluating them with user testing. This section discusses these few music applications where research has specifically focused on mobile optimization in addition to the evaluation of some mobile interface designs of commercial streaming systems.

2.2.1 Kugou, Kuwo, and QQ

Despite explicitly focusing on the research gap in China, Hu (2019) is still relevant in this domain due to the general lack of research evaluating music services specifically on mobile devices. Hu uses a combination of ResQue, the Recommender systems' Quality of user experience evaluation

[1] https://www.apple.com/newsroom/2001/10/23Apple-Presents-iPod/

framework by Pu et al. (2011), and Nielsen's 10 usability heuristics[2] to evaluate the user experience and recommendation accuracy of China's three most popular music mobile applications Kugou, Kuwo, and QQ. The first two mentioned focus purely on music. The latter is an all-in-one app, focusing on twitter-like social media functions, where users can utilize a variety of other apps, such as the music application tested in this study, from within QQ. These are shown in the Figures 11, 12, and 13, where a, b, and c refer to the homepage, settings menu and search results page.

While observing the results of this study one must keep in mind that only Chinese user perspectives were studied and that cultural factors may cause bias. All eighteen participants were somewhat acquainted with both the user testing process and the applications tested as they had previously used and evaluated at least one of them prior to this study. Users were observed using a think-aloud protocol while navigating, browsing, searching, and exploring recommendations and their accuracy for all three apps. They were subsequently interviewed about their experience. According to Hu (2019, p. 20), positive criteria included *"Feedback, Metaphor, Consistency and Memory"* from Nielsen's heuristics, while users regarded the criteria of *"Recommendation Accuracy, Interaction Adequacy, Design and Privacy"* from ResQue as being negative.

Overall, users did not respond favorably to homepages which positioned large banners on top as they took up a large portion of their phones screen space. They preferred simplistic icons and most-used shortcut features. At the same time, if the design seemed too minimalistic, it was considered to be a poor composition choice, as it favors style over features and explanations. While participants reacted mostly positive, Hu (2019, p. 23) reported that, in specific regards to QQ's design, there was *"too much information displayed on the screen"* which made the interface feel less intuitive.

Hu recommends that designers should try to find an aesthetic balance between form and function. This is especially important when designing for mobile devices with smaller screens. In addition, Hu noted that while only one participant mentioned the feature of customization, other previous studies published in the western world have shared this sentiment and that personalization also be taken into consideration when designing music recommender systems for mobile devices.

[2] https://www.nngroup.com/articles/ten-usability-heuristics/

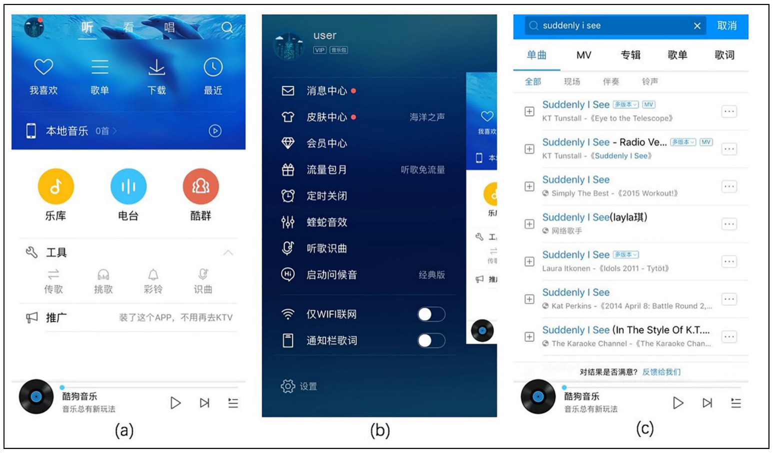


Figure 11: Kugou mobile interface Hu (2019).

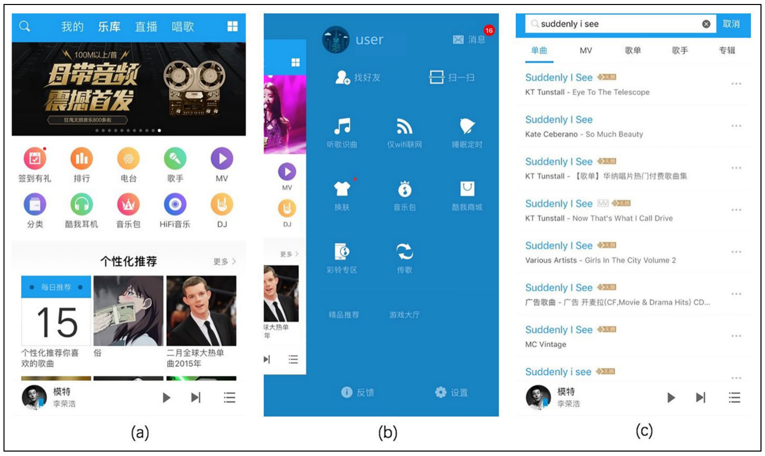


Figure 12: Kuwo mobile interface Hu (2019).

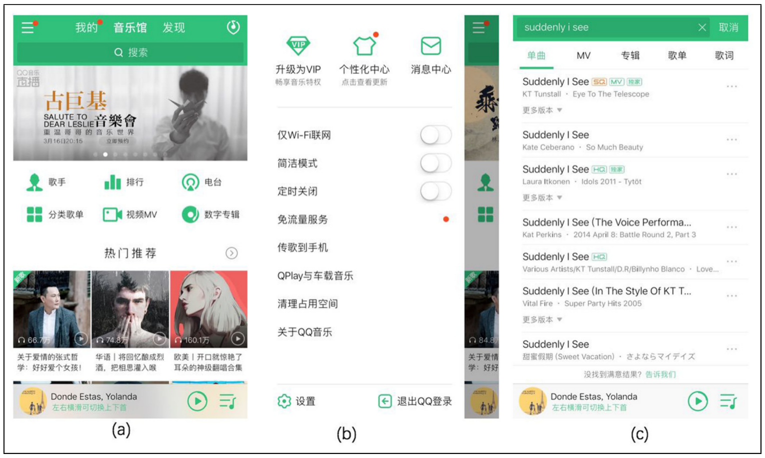

我的 音乐馆 发现
搜索
古巨基 SALUTE TO DEAR LESLIE 音樂會
歌手 排行 电台
分类歌单 视频MV 数字专辑
热门推荐
关于爱情的张式哲学：好好爱个女孩！
华语｜将回忆酿成烈酒，把相思灌入喉
欧美｜开口就惊艳了耳朵的神级翻唱合集
Donde Estas, Yolanda 左右横滑可切换上下首
(a)
升级为VIP 畅享音乐特权
个性化中心 点击查看更新
消息中心
仅Wi-Fi联网
简洁模式
定时关闭
免流量服务
传歌到手机
QPlay与车载音乐
清理占用空间
关于QQ音乐
设置
退出QQ登录
(b)
suddenly i see 取消
单曲 MV 专辑 歌单 歌词
Suddenly I See
KT Tunstall · Eye To The Telescope
更多版本
Suddenly I See
Kate Ceberano · So Much Beauty
Suddenly I See
Laura Itkonen · Idols 2011 - Tytöt
更多版本
Suddenly I See (The Voice Performa...
Kat Perkins · 2014 April 8: Battle Round 2, Part 3
Suddenly I See
Various Artists/KT Tunstall/D.R/Billynho Blanco · Love...
Suddenly I See (In The Style Of KT T...
Vital Fire · Super Party Hits 2005
更多版本
Suddenly I See
甜蜜假期 (Sweet Vacation) · さよならマイデイズ
没找到满意结果？告诉我们
Donde Estas, Yolanda 左右横滑可切换上下首
(c)

Figure 13: QQ mobile interface Hu (2019).

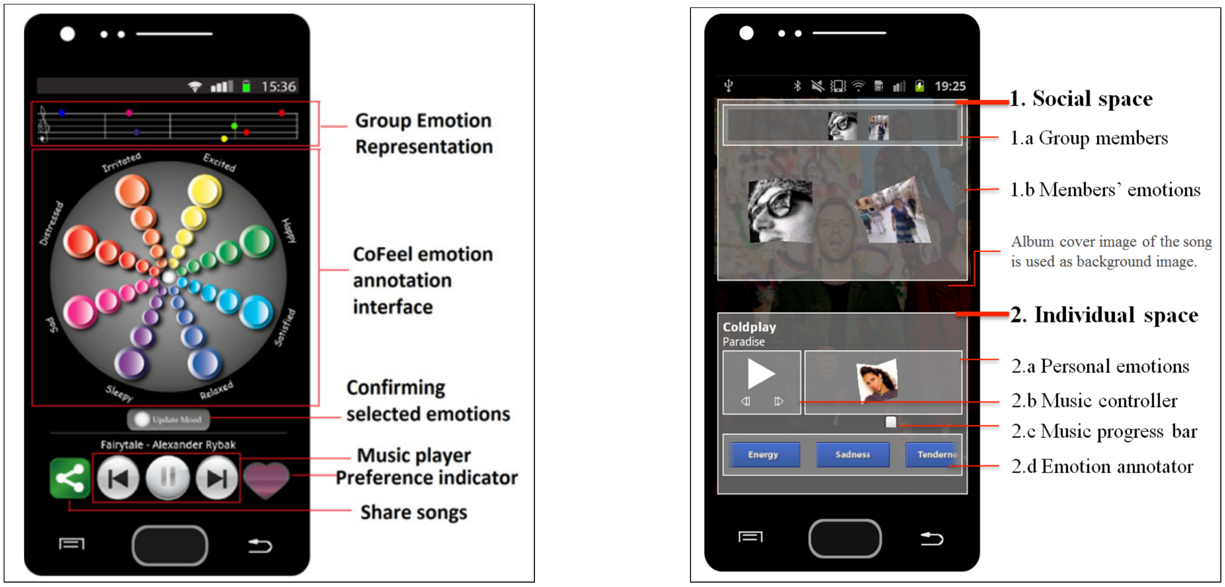
2.2.2 CoFeel and Empatheticons

According to Chen (2015), persuasive social recommendations in a group setting are effective at motivating users towards living a healthier life, and yet few users actually reach their goals. Chen's thesis focuses on improving interface design for social group recommender systems and starts by exploring interactive music recommendation visualization tools. GroupFun was the initial music group recommender system application created to host these tools. It was originally created for desktop computers but was ultimately designed and tested on a Samsung Galaxy SII 9100 with Android OS. The mobile version of GroupFun is a simple application where users log in with their Facebook account to create and join different music listening groups. Users add songs to a group playlist, rate these and other songs in said playlist, and ultimately get to see their groups overall rating for the same songs. According to Chen (2015, p. 23), *"GroupFun generates recommendations by aggregating all users' preferences using Probabilistic Weighted Sum (PWS) to maximize user satisfaction"*.

To enhance GroupFun, three different visualization tools for emotion annotation were researched: a color wheel which uses a radial view visualization technique, as well as two icon designs, hapticons and kineticons. The most promising methods were CoFeel and Empatheticons which are shown in Figure 14. The visualization objective of both CoFeel's color wheel and Empatheticons kineticons, is emotional context in a social group setting. The goals in both experiments were to test for user engagement, usability and usefulness, where both applications received positive feedback in all three areas.

While participating in a group music listening session, CoFeel enables users to input their current mood and the strength of that mood. On this emotional color wheel, moods are represented through a combination of colors and text and are situated in the space based on valence and arousal dimensions. In order to choose an emotion, users must pivot and tilt the phone until the track icon reaches the appropriate spot on the screen. A user can see the moods which other users in their group chose for that song and all the individual moods collectively become the group's total mood score.

A qualitative pilot study was conducted with only four people, therefore the data may be considered to have little statistical significance. User behavior was observed during the sessions, and participants were also recorded due to the additional use of the think-aloud method as well as the unstructured follow-up interviewing of participants. All participants reacted positively to CoFeel and agreed that the design of the interface was helpful in the tagging of emotions. In terms of interaction, having to rotate the phone to register emotions proved cumbersome while in motion, although everyone thought it was fun. Despite the fact that great consideration was given as to how much information CoFeel should display due to the limited space of a mobile screen, all users felt that the listed emotions should be more dynamic as emotions may have different



(a) CoFeel interface.

(b) Empatheticons interface.

Figure 14: CoFeel and Empatheticons mobile interfaces from Chen (2015).

meanings in different contexts.

Empatheticons gets its name from the combination of empathy and motion icons. With Empatheticons, Chen (2015, p. 38) have been able to *"exploit kineticons - an iconographic motion technique - as a means to visualize emotions"*. It differs from CoFeel in that it provides both an individual and group space visualization, as well as being more dynamic with more detailed and personalized emoticons. Users no longer have to try to interpret the emotion annotation of others, as they are now more easily perceived through iconic emotional moving pictures of the individual user. The mood categories used were based off of the Geneva Emotional Music Scale (GEMS) created by Zentner et al. (2008). Zetner's work is perhaps the most extensive research ever done on classifying emotions induced by music. GEMS organizes emotions into 9 major mood categories: wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension, and sadness.

18 people from various countries around the world, divided into 6 groups of 3, partook in the user study for Empatheticons. Users were observed during testing and were asked to fill out a questionnaire postliminary. According to Chen et al. (2014, p. 7), *"users could easily map the empatheticons to their respective emotions in a musical context"*, *"empatheticons could effectively enhance users' perceptions of togetherness"*, and user satisfaction had a high correlation with user annotation activity levels. However, it is interesting to note that in order for users to get acquainted with the system before testing, they were given a piece of paper explaining the different emotions. From an explanatory perspective, this interface is obviously not explicit enough for users to clearly

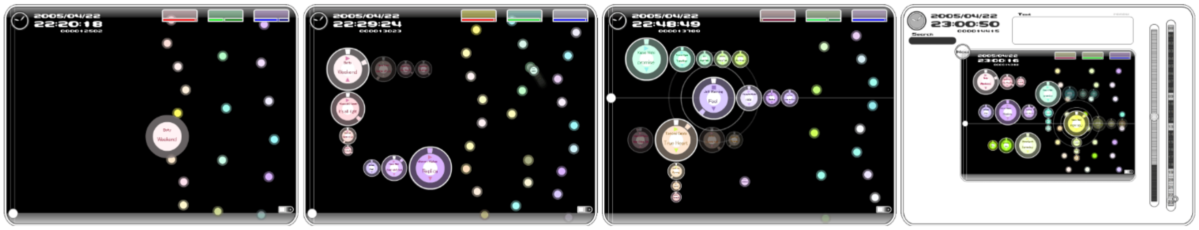
understand the meaning of these emotions just from using the application. At the same time, user feedback showed that participants found the application easy to use, easy to learn, useful and fun.

2.2.3 Musicream

The second earliest interface mentioned in this thesis is Musicream, created by Goto & Goto in 2005, and slightly updated in 2009. Goto & Goto (2009) explain that Musicream's name comes from the combination of 'music' and 'stream'. Even though this was designed 15 years ago with limited technological design possibilities, it was quite remarkable for its time. This creative interactive interface for music discovery, as shown in Figure 15, is reminiscent of a CD collection, with dot-like song nodes representing disks.

The system has four main functions: the music-disc streaming function, the similarity-based sticking function, the meta-playlist function, and the time-machine function. Three different mood colored boxes on the top right of the screen represent faucets or taps which dispense songs in the form of small floating circles. With the music-disc streaming function, new songs are constantly being displayed to a user in order to promote music discovery. The amount of cascading songs released depends upon the weight the user has given to that particular mood feature. The song circles themselves are also colored coded based on the emotional feeling of that particular musical piece. When a user clicks on a circle, it expands into what Goto & Goto (2009, p. 143) refers to as *"maintenance mode"*, which displays a CD looking icon containing the title and artist name with playback controls. Moving a disk to the left-hand side of the interface selects it and allows the user to listen to the song. One can also attach several disks together to make a stack of CDs, or rather create a sort of playlist. This similarity-based sticking function to create a disc series has an effect on what other songs appear next in the streaming cascade of new music. According to Goto & Goto (2009, p. 142), the *"ease of sticking"* is contingent on how alike the moods are in terms of timbre similarity. Goto & Goto explain further that two songs which are considered dissimilar may not stick together at first, though users can potentially change this similarity measure by attempting to bring these two songs into contact multiple times. While a song is playing, users can simply hover their mouse over another song in the CD stack playlist in order to cross-fade into the next song. Playlists can be edited while in maintenance mode, while compressing an expanded playlist brings it back to its original dot-like state, putting it into what Goto & Goto (2009, p. 143) refers to as *"compact mode"*.

The meta-playlist function generates a playlist of playlists, where each series of discs are played in the order of importance based on where they are located on the screen in relation to the horizontal line playback bar. The bar can be moved, and anything in contact with the bar has precedence from left to right, and then other songs are played from top to bottom. Any playlists in compact mode



Music-disc streaming function: Active and unexpected encounters with interesting musical pieces

Similarity-based sticking function: Controllable and flexible "query-by-example" similarity search

Meta-playlist function: Advanced playlist editing with a high degree of freedom

Time-machine function: Intuitive browsing and perfect recalling of past music listening activities

Figure 15: Musicream tablet interface from Goto & Goto (2009).

are excluded. Musicream also has a non-interactive auto-playback mode. New songs are smoothly transitioned to automatically and chosen from the available on-screen songs based on the mood of the song that is currently playing. The time-machine function keeps screen shots of all activity so that users can revisit a previous playlist setup. A user can not only see previously played tracks but retrace the exact actions they took during a previous session too. Playlists can be taken from a past listening state and be copied and transferred to a new one.

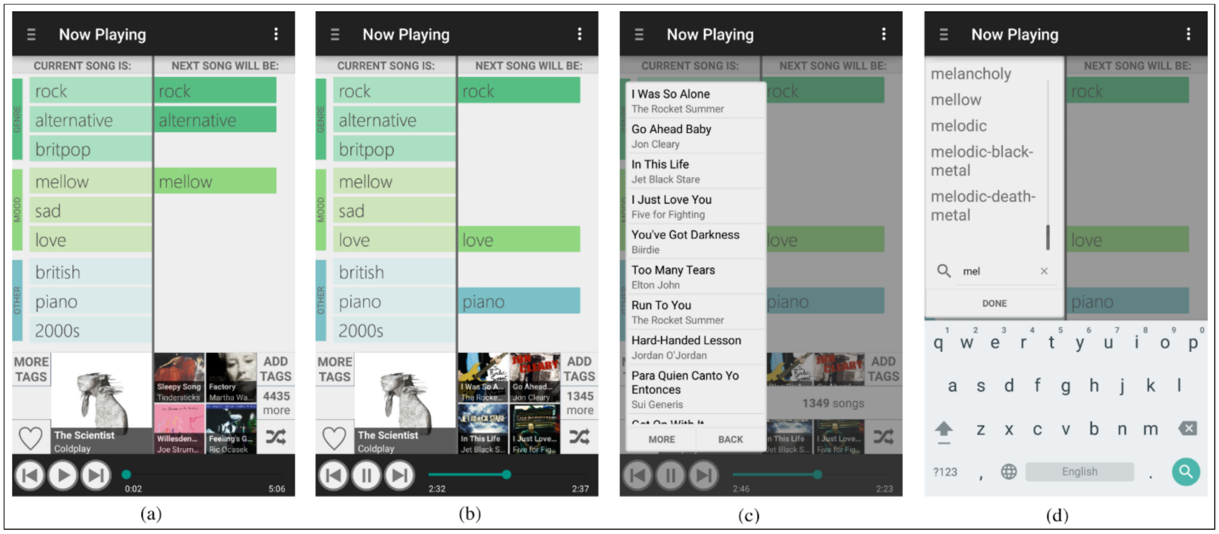
Before using testing, a trial was run on a touch screen windows tablet. Since there was no keyboard or mouse, Musicream was operated with a digital pen. Users had just under 2000 songs to choose from, the majority of which were from Japan's top hits chart, though Musicream can handle larger libraries. For the main study, there were 27 user testing participants who completed a questionnaire after freely exploring the tablet application for 5 minutes. Subjects were then asked to compare Musicream to a baseline standard music player.

Although creating playlists on Musicream was more convenient, it was just as easy to do so in the baseline system. Most users responded favorably to the interface, agreeing that all four features, music-disc streaming, similarity-based sticking, meta-playlist, and time-machine, were interesting. On the contrary, while feature one was seen as effective and four was seen as convenient, user feedback revealed that users did not want to use feature two and that feature three was considered inconvenient. Users anticipated a manual search ability, which was nowhere to be found. Even though Musicream is meant to expose users to unfamiliar music, other studies show that such interfaces should also give users the option to search manually for specific songs. Automatic playback mode was useful when trial users got bored of manually interacting with the system. While there was no concrete evidence of this from user feedback, it is possible that the system had too much functionality, causing therefore users to opt for this non-interactive possibility. While there is no statistical data, Goto & Goto (2009) claims that Musicream is easy to use based on the fact that users did not need a lot of training to get used to the application. However, the interface does not seem to be explicitly informative as there are no indications for how to use the interface or what the different components do. It cannot be intuitive if one must be taught to use it. Goto & Goto explain that when considering the redesign of Musicream, it should include combining the interface with a text-based music player in addition to showing additional bibliographic information about songs. Overall, users preferred Musicream to the baseline system, saying it was much more enjoyable.

2.2.4 TagFlip

Kamalzadeh et al. (2016) is one of the few studies to take into account the ever-growing number of people who listen to music on the go. Kamalzadeh created TagFlip, a tool for mobile devices which tries to find a middle ground between cognitive load and user control by utilizing popular social Last.fm tags for music discovery. Figure 16 displays the user interaction progression of the TagFlip interface. Upon first using TagFlip, users can either search for a specific track or choose a variety of tags to suggest what to listen to. When the current song is playing, the bottom left-hand side of the interface displays the album art of this song and above it appears a list of its top tags from the categories of genre, mood, and other. The system initially chooses three tags to set the basis for the auto playlist creation; two genres and one mood. This list is displayed on the right side of the screen, with 4 album covers from the next four songs in the playlist shown underneath it. While the left-hand side updates with new tags and album cover art whenever a new song plays, the right-hand side stays the same, though the user is free to adjust these tag combinations at any time. This progression of changing tags, as shown in a and b, changes the amount of songs in their automatic playlist from 4435 to 1345. By clicking this number amount, as shown in c, the user has access to the entire playlist. Users can also add tags to songs, as shown in d.

User testing started with three stages of formative usability tests where ten different participants tested nine initial designs and two final prototypes. They helped Kamalzadeh et al. (2016) to create a set of design considerations for future research. Adding complex features can take up a lot of space on a small screen, so the authors needed to find a balance between control, explanations, and simplicity. One design, inspired by DJ mixer faders, had separate boxes for each tag but reorganizing things from left to right proved to be less confusing and more organized, creating a more natural layout flow. They also tested the ability to update the strength of a particular tag to change the recommendation for the upcoming song either via interactive bars, knobs, sliders, or text size. Users were unsure about what strength referred to, and in the end gave the feedback that this level of control was unnecessarily complex. They considered adding visual encodings to tags to explain how the size of the current playlist song set would change based on the addition or removal of a tag, but this idea was also discarded as it too proved not to be needed. In order not to overload a user with too many choices or too much information, it was clear that only a few tags



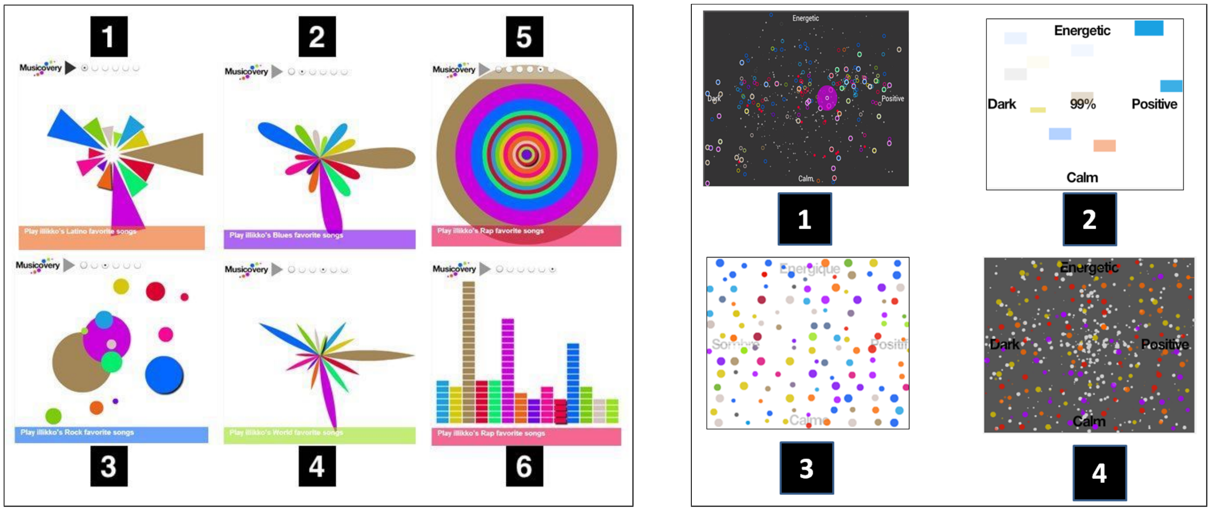
(a) (b) (c) (d)

Figure 16: TagFlip mobile interface from Kamalzadeh et al. (2016).

could be displayed on the main screen simultaneously. Genre and mood were the two most popular ways to classify tags, so the other category was created in order to incorporate additional metadata. They tried to give users access to all the tags for each specific song, but this led to issues of visibility since pinned tags would then disappear from the screen as users scrolled through the list to look at other tags. TagFlip therefore only displays the top three tags from each group, making for a total of nine tags.

A formal within-subjects design of TagFlip's final interface was completed with 16 participants. Users were given 5 minutes to get acquainted with TagFlip before testing both it and the baseline, Spotify. Participants completed questionnaires regarding demographic information, usability in terms of the system usability scale (SUS), and recommendation aspects based on the ResQue framework. Afterwards, they were interviewed. As reported by Kamalzadeh et al. (2016), their interface excelled over the baseline in terms of control, explanations, interaction adequacy, interface design, transparency, and trust. At the same time, satisfaction and the willingness to reuse the apps were the same for both TagFlip and Spotify. Even though Spotify required more effort than TagFlip, it was just as easy to find songs in both interfaces. Though not perfect, users overall preferred TagFlip to Spotify.

TagFlip does not have a single seed song for playlist creation but rather promotes music discovery by allowing the user to update next song tags. According to Kamalzadeh et al. (2016), the user control aspect to this tool is scalable. This means that the application adjusts to both the needs of users with either low or high engagement levels. Active users can fine tune their playlist as often as they would like, while allowing other users to update more seldom. The passive users who may be non-technical or have issues using other systems due to high mental load will enjoy TagFlip's low interaction effort. In terms of explanations, users can see direct correlations



(a) October 8, 2012 Facebook Post.

(b) April 24, 2013 Facebook Post.

Figure 17: Musicovery visualizations from Vavrille et al. (n.d.).

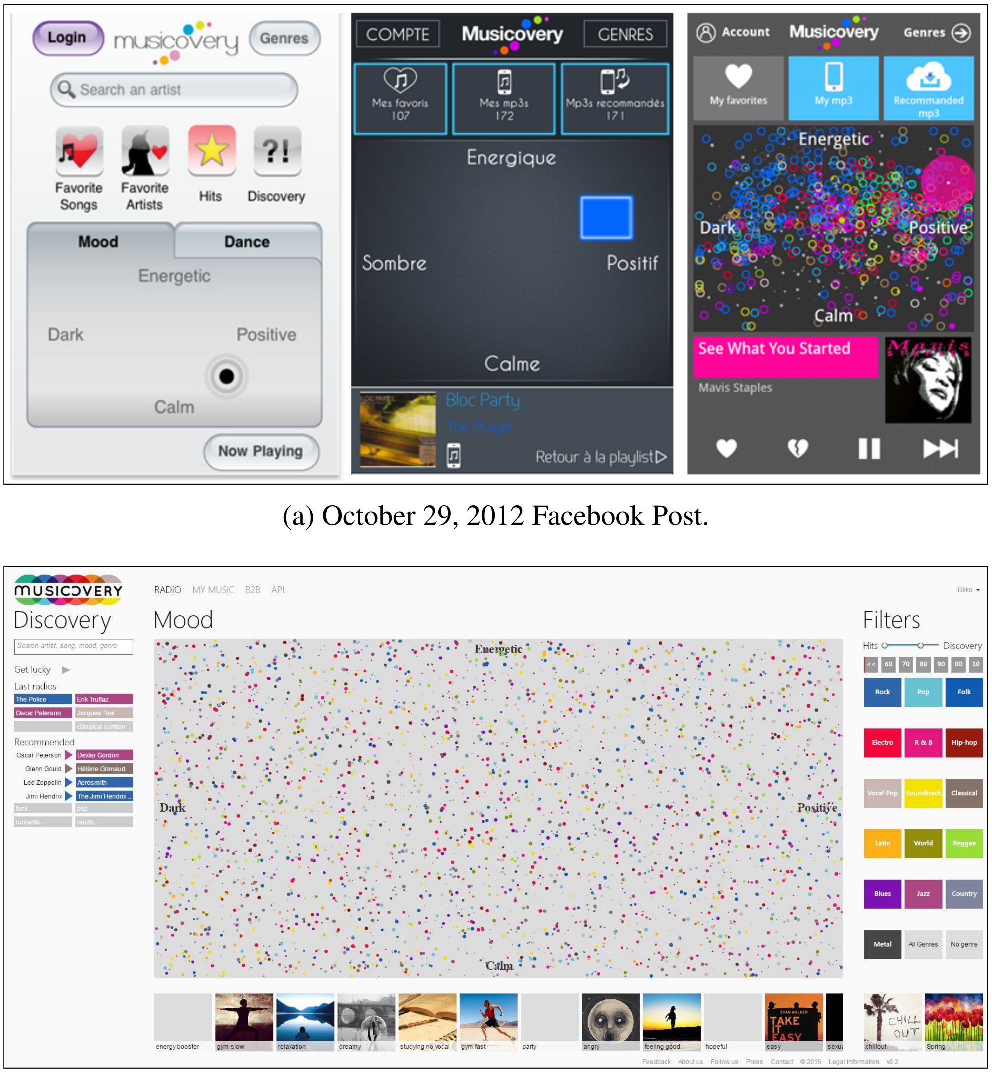
between tags and song recommendations, especially when a user modifies them. This type of transparency increases trust and user satisfaction. The majority of users revealed that there is a need for TagFlip, as nothing else commercially available comes close to what it achieves.

2.2.5 Musicovery

Between 2007 and 2017 the commercial streaming service Musicovery[3] was a popular music recommendation system which allowed users to discover new music by interacting with its 2D affective plane. Playlists were automatically generated based on the mood palette of a user. Their web player's initial colorful interface consisted of a side panel with various filtering methods such as dance, genre, mood, and year. The main area was a 2D node-mapping space showing recommendation results, as seen in Figure 18b. Each song was represented as an individual node with a specific color and flower sunburst style, where nodes were connected in the 2D space to explain to users the musical relationship between each other. The center of each node displayed textual artist information and when clicked upon, brought up a mini music player with the ability to love or hate the song.

Musicovery experimented with a variety of visualizations, asking users through social media for feedback about which they preferred. Halfway through their success in 2012, they wrote a post on their Facebook page asking which type of music visualization style users preferred most out of the six types shown in Figure 17a. Among all of the 22 people that commented, one person said they like them all equally. Out of the rest of the 21 people, the majority preferred number

[3] http://b2b.musicovery.com/



(a) October 29, 2012 Facebook Post.

(b) January 7, 2017 Facebook Post.

Figure 18: Musicovery mobile and desktop interfaces from Vavrille et al. (n.d.).

2, the rounded version of the rose plot display, followed by number 3, the circular dot icons. Further research showed that the circular dot icons were indeed the most popular. Shortly after they launched their application for mobile devices. The progression of the interfaces over the span of one year, as seen in Figure 18a, leads away from the original design to something more simplistic and taking into account user feedback. During this time Musicovery continued with user testing, once again posting on Facebook the image shown in Figure 17b, asking users which mood pad design they liked best. The chosen mood pad design was number 3 from Figure 17b. User feedback suggested that they should keep iterating the design as it lacked usability and some user interface features were difficult to understand. On May 10th 2013, Vavrille et al. (n.d.) announced the release of a new website version of the application. From there they seem to have focused their re-designs on the web-app as some users expressed being hesitant to using such a mobile application for music, preferring to use other music players such as iPods. Musicovery only worked with a user's own music library and at the time mobile devices could not hold a large amount of songs due to lack of storage space, which may have also accounted for the lack of future updates for the mobile application. In time, Musicovery adapted and became a radio application, allowing users to listen to music other than what was in their own personal library.

Musicovery was the first commercial system to implement the use of moods for music discovery. There were scrutable aspects to the application such as giving users the ability to control their recommendations. Not only were users able to unlike songs but the system also prevented those songs from ever being recommended in the future. Musicovery was also the first application to create a transparent user interface without the use of explicit textual explanations. Although their use of the of arousal / valence model used in the 2D space for music visualization was novel and explicit user testing went into developing this application, Musicovery decided to focus on their recommendation algorithms instead and discontinued all applications in 2017.

2.3 Explanations in Recommender Systems

Many recommender Systems are commonly considered to be 'black boxes' by not providing enough information or justification to the user about the suggestions they receive. Tintarev & Masthoff (2015, p. 353) describes a black box system as one that lacks transparency in its recommendation process in addition to not *"offering any additional information to accompany the recommendations beyond the recommendations themselves"*. For example, in regard to music recommender systems, it can be confusing to a user when two music streaming services, such as Spotify and Apple Music, display different recommendation results even if the user has the exact same listening habits on both platforms. Therefore, systems may clarify information through explanations. According to Tintarev & Masthoff (2015, p. 353) *"an explanation can be an item description that helps the user to understand the qualities of the item well enough to decide whether it is relevant to them or not"*. These can be presented as an algorithm or equation, through natural language or with visualizations. Types of explanation purposes have varied slightly from paper to paper, where researchers like Tintarev & Masthoff (2007), Jannach et al. (2011), Tintarev & Masthoff (2015), Nunes & Jannach (2017), Zhang & Chen (2018), and Jannach et al. (2019), have stressed the importance of different categories. Based on the literature mentioned above, the main 12 purposes for using explanations are for Comprehensibility, Debugging, Education, Effectiveness, Persuasiveness, Relevance, Satisfaction, Scrutability, Transparency, Trust, and Validity. Table 1 should effectively cover all of the various explanatory goals.

Table 1: Recommendation explanation definitions.

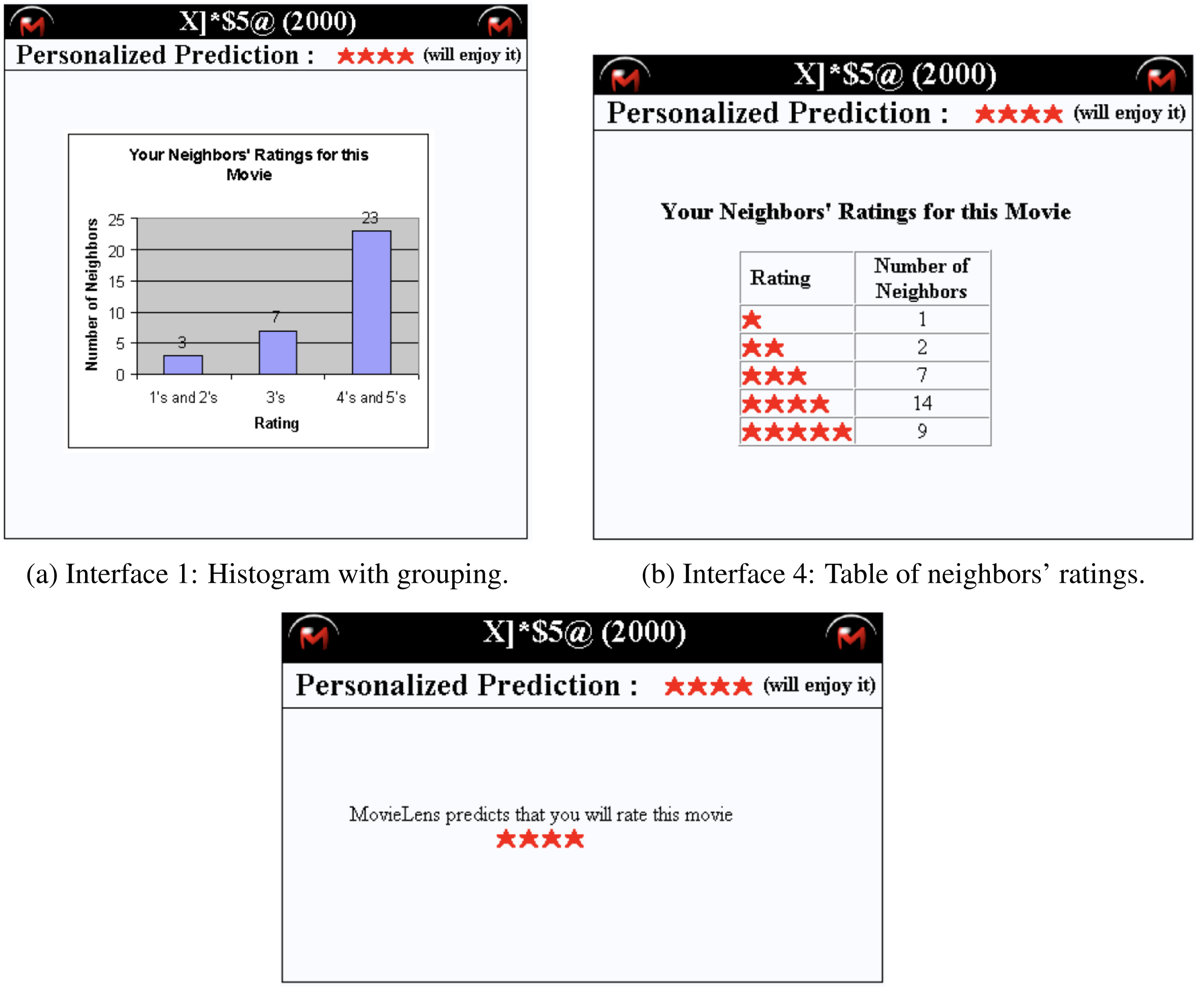
Explanation Type	**Definition**
Comprehensibility	As systems often do not have prior knowledge of a user and do not know their technical ability level, there should be a match between recommender concepts and user concepts known to be generally understood.
Debugging	The system should present adequate information so that a user can identify aspects that should be troubleshooted.
Education	Educational explanations should teach users about the system and inform about the product domain so as to better understand it.
Effectiveness	Effective explanations should help users improve their decision making skills by providing relevant recommendations, thereby allowing users to explore and identify their preferences.
Efficiency	Efficient explanations should curtail the time and cognitive effort needed to make decisions.
Persuasiveness	Persuasive explanations are meant to influence a user's decision making behaviors by convincing them to complete a certain action such as liking or buying a recommended item.
Relevance	Relevance explanations may be needed for the justification of asking users for further information.
Satisfaction	Explanations for satisfaction are used to provide better interaction and user experience with a recommender system.
Scrutability	When users identify aspects that should be troubleshooted, scrutable explanations should allow them to relate this information directly back to the system.
Transparency	Transparent explanations allow users to better understand why they receive certain suggestions.
Trust	Explanations for trustworthiness should enhance the user's belief and confidence in the system by raising the user's certainty in the quality of their recommendations.
Validity	Explanations for validity should allow users to judge and verify their recommendations.

In accordance with Kouki et al. (2017), the six major explanation styles are social, content-based, user-based, item-based, item average rating, and user average rating. A system can provide social recommendations for music by stating to the user that song A was recommended since their friend likes song A. A content-based approach could for example explain that song A was recommended to the user since the user likes a particular genre of music and song A also falls under that genre category. In essence, new items are recommended to a user-based items which they previously liked and are explained as such. To explain a user-based recommendation, a system would clarify that song A was recommended based on other user profiles with similar tastes. For item-based recommendations a system would explain to the user that song B was recommended since people who like song A, also like song B and the user likes song A. If the system explains to a user that song A was recommended since it is popular, or highly rated, then that is item average rating. A system can also use user average ratings to explain a recommendation. It may state that song A was recommended since it has high ratings and the user often gives high ratings to songs.

Explanations are not only necessarily for end-users. They can serve a variety of functions, being either user-centric or used for the benefit of developers, a product owner, or a company's stakeholders. Nunes & Jannach (2017) describe that the first type of explanations utilized were system logs which provided developers with information to fix computer bugs. These event reports, while informative for developers, are too detailed and complex for normal users to understand. Information density is always important to keep in mind since information overload tends to lead to negative results. There are a lot of factors to consider when creating explanations. Explanations need to be adapted to different situations and different domains. Currently, according to Nunes & Jannach (2017, p. 169), "there is... no clear consensus on what constitutes a good explanation". The following explores these topics through prior prominent research within the domain of recommendation explanations.

2.3.1 Explaining Recommendations with MovieLens

One of the first major works to explore the importance of explanations in recommender systems was Herlocker et al. (2000), published by the GroupLens research organization at the University of Minnesota. An assessment of different explanatory interfaces for movie recommendations was done with the help of automated collaborative filtering (ACF) and the MovieLens system. Herlocker et al. (2000) states that ACF is a recommendation technique which follows a white box model, as opposed to black box, and calculates recommendations based on neighbor similarity through user feedback, such as ratings. MovieLens was created in 1997 by GroupLens as a non-commercial system for personalized movie recommendations, similar to the commercial



X]*$5@ (2000)
Personalized Prediction : ★★★★ (will enjoy it)
Your Neighbors' Ratings for this Movie
Number of Neighbors
25
20
15
10
5
0
3
7
23
1's and 2's
3's
4's and 5's
Rating

(a) Interface 1: Histogram with grouping.

X]*$5@ (2000)

Personalized Prediction : ★★★★ (will enjoy it)

Your Neighbors' Ratings for this Movie

Rating	Number of Neighbors
★	1
★★	2
★★★	7
★★★★	14
★★★★★	9

(b) Interface 4: Table of neighbors' ratings.

X]*$5@ (2000)

Personalized Prediction : ★★★★ (will enjoy it)

MovieLens predicts that you will rate this movie ★★★★

(c) Interface 11: Baseline recommendation without extra data explaining why.

Figure 19: MovieLens explanatory interfaces from Herlocker (2000).

system IMDb[4].

Two different studies were completed as well as an initial pilot test to find out the most effective models and techniques for supporting explanations in ACF systems, as well to see if these explanations lead to increased acceptance and better performance. As stated in Herlocker et al. (2000, p. 7), interviews conducted in the initial pilot study revealed that explanations changed participants' *"internal conceptual model"*. Therefore, in the two main studies that followed, participants were explicitly told that all explanations were generated using the same model. Visually, 17 of the 21 interfaces used textual explanations, while 3 presented graphs and 1 showed a table. Figure 19 shows the three main types of interfaces.

After the pilot test, the earlier study had participants evaluate the acceptance and filtering performance of explanations by filling out pre and post surveys and testing the system over the

[4] https://www.imdb.com/

period of one month. Each participant received a special link to the online website MovieLens. Upon entering the site, they were either presented with the original layout, which was the control, or a random new explanatory interface. As user testers may have been familiar with the system already, the layout for the control group, while not including additional explanations, was still changed ever so slightly to give the user the impression that it could potentially be one of the new experimental interfaces. Out of 210 users, 97 who were not in the control group completed the post survey. According to Herlocker et al. (2000), 86 percent of the 97 preferred the explanatory interface and would like to see this feature added to MovieLens. The majority of additional feedback was positive, with some mentioning negative aspects to the algorithms and accuracy of the recommendations. At the same time, results regarding the effect of explanations on decision performance were ambiguous.

The later study investigated explanation techniques by having participants compare all 21 explanatory interfaces. In order to prevent a bias as to whether or not a user liked or disliked a particular movie, each explanatory interface recommended the same movie, and this title was scrambled. Results from Herlocker et al. (2000) uncovered that most compelling way to present the explanatory reasoning behind recommendations is through histograms. Specific content features, such as favorite actor or actress, were received positively by many, however there was a high variance, showing a clear division between those who evaluate movies based on these specific factors and those who don't. People have a desire to search or filter recommendations differently using various and diverse features. In terms of trust, personalized recommendations, like through comparing user data to other users, were perceived to be more accurate than ratings from other sources such as critics, which was also validated in previous studies.

Both studies showed that while detailed graphs were popular among expert users, the authors advise that one should avoid creating interfaces which are too complex as too much additional information by way of explanations can be confusing and result in information overload. This sets a precedence for finding a middle ground and stresses the importance of good design, as Herlocker et al. (2000, p. 7) states that *"poorly designed explanations can actually decrease the effectiveness of a recommender system"*.

2.3.2 TasteWeights

TasteWeights, as described in Bostandjiev et al. (2012), is a hybrid social music recommender system intended to help users gain control over their recommendations for better system transparency. Its visual explanation interface, as shown in Figure 20, uses interactive features to elicit preferences from the end user for better recommendations. As its name implies, users are encouraged to adjust their musical tastes via interactive slider-weights and other UI components

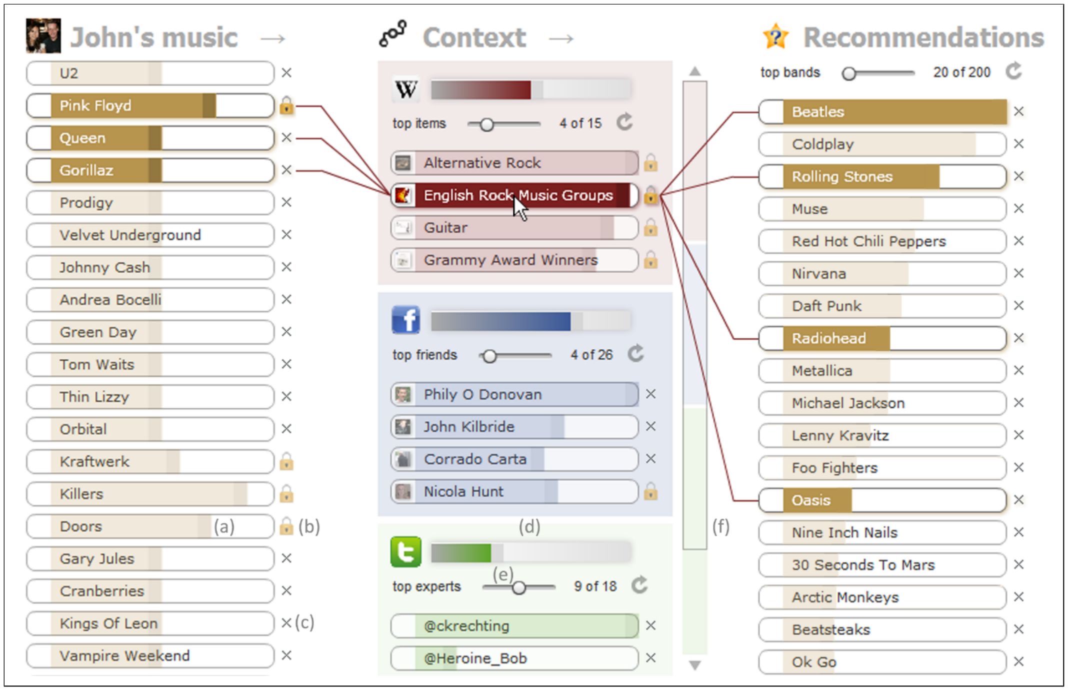


Figure 20: TasteWeights explanatory interface from Bostandjiev et al. (2012).

for a more personalized user experience. Hybrid recommendations were generated using a variety of meta-data and social media information about users' friends' preferences gathered from Facebook, Twitter, and Wikipedia. The left-hand side of the interface displays music the user already likes. The middle section, context, explains how much information from each source is being inputted when recommending new music, allowing users to explicitly modify the importance of specific factors. Bostandjiev et al. (2012, p. 3) describes how each context source has a particular color, where its opacity changes based on *"the weight of the source expressed through its source slider"*. In real time, recommendations are shown in the last column on the right.

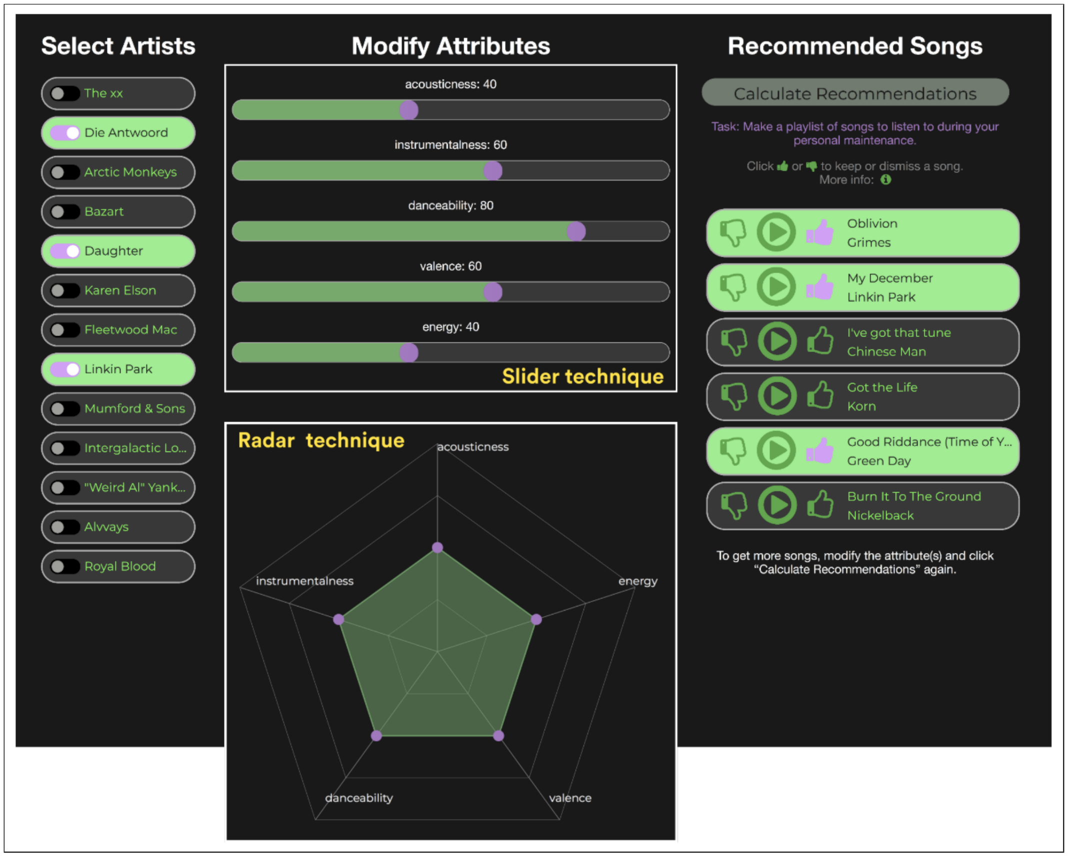
The TasteWeights system was tested in numerous ways with varying levels of interactivity to explore the effects of system controllability. After a pilot study with 7 user trials uncovered that the cross-hybrid was the best recommender, the team went ahead with a controlled user study with 32 people. A within-subjects experiment was performed to conduct a quantitative analysis on the accuracy of the different recommendation methods. Bostandjiev et al. (2012, p. 8) came to the concrete conclusion that *"hybrid strategies combining different social APIs can provide better recommendations than traditional CF (over Facebook music preferences)"*. Subjects were also given a post-questionnaire to make a qualitative analysis about user experience. User feedback from Bostandjiev's post questionnaire showed that 88% of participants said that the system helped them understand their recommendations. Through interacting with the system, 84% of users were able to receive better recommendations. 80% of participants found the system to be informative, stating it was useful to see how items were connected. 82% of participants also said the system was fun to use. However, 78% of participants said it was easy to use, and only 64% felt that the system was intuitive. The system was predominately well received and liked by the majority. System ease of use was perceived to be fair, but was only deemed intuitive by just over half of all participants. Previous research has shown that providing the average user too many options can be confusing, and that may be the case here.

As reported by (Bostandjiev et al., 2012, p. 3), explaining recommendations keeps a user *"in the loop"* by educating them about the recommendation process so they can understand the reasoning behind a recommendation. This helps with user acceptance, confidence, justification, and transparency (understanding) of their recommendations and of the recommender system itself. Tasteweights is a great example of the benefits of building an explanatory interface as it was shown to increase overall user satisfaction, recommendation accuracy, and user experience.

2.3.3 To Explain or Not To Explain

The black box effects of recommender systems encourage designers to create more visually pleasing and personalized interfaces which help users understand recommendations as well as take better control of them. Both Millecamp et al. (2018) and Millecamp et al. (2019) discuss how they designed interactive interfaces with different visualization methods to analyze controllability, explanations and the effects of personal characteristics on music recommender interfaces. Figure 21 depicts the initial 3-column formatted interface with two different 'Modify Attributes' variations, the slider technique and the radar technique. Users select seed artists, modify the musical attributes acousticness, danceability, energy, instrumentalness and valence using one of the techniques, preview a song for 30 seconds, and finally click either thumbs up on a song to keep it in the playlist or thumbs down to remove it. The second interface, as shown in Figure 22, incorporated positive aspects of these different features as well as adding a scatterplot and why buttons for explanations. In addition to choosing a seed artist, users can click the '?' icon to understand the specifics behind each musical attribute preference, recommendations are explained by comparing the user's modified preferences to those of the current selected song, users can compare individual attributes in a scatterplot layout, and disliked songs are now displayed in addition to liked songs.

The Millecamp et al. (2018) study compared two different visualization techniques, simple sliders against the more advanced radar chart. This was done to see what user-related characteristics influence perception and performance of the user, given that visualizations seem to be more effective than textual rank-lists, though more difficult for some users. Out of the 80



acousticness: 40

instrumentalness: 60

danceability: 80

valence: 60

energy: 40

Task: Make a playlist of songs to listen to during your personal maintenance.

Click 👍 or 👎 to keep or dismiss a song.
More info: ⓘ

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

Figure 21: Millecamp's first explanatory interface from Millecamp et al. (2018).

personally procured user testing participants, 40 were approved to be evaluated and finish the study. In addition to routine demographic questions about age and gender, before testing the system, Millecamp et al. (2018, p. 104) also asked users about their *"music sophistication, visual working memory, tech-savviness, Spotify usage, familiarity with recommender systems and attitude towards recommender systems"*. For each interface, users were asked to create a playlist and evaluate their experience afterwards. Evaluation questions were based on the recommender systems framework ResQue, to assess user experience quality. Participants were also asked to rate the different musical attributes as well as answer some open-ended questions about control and visualizations.

Overall, Millecamp et al. (2018) discovered that visual explanations gave users a better understanding of their recommendations as users were able to recognize how song suggestions and selected musical attributes were related. Participants favored musical attributes in the following order: energy, acousticness, danceability, instrumentalness, tempo and valence. Users unsure of their music preferences seemed to have some difficulties utilizing the 'Modify Attributes' section. In terms of Spotify Usage, avid or frequent users of music recommender systems were more willing to interact with complex visual interfaces than casual users. Users

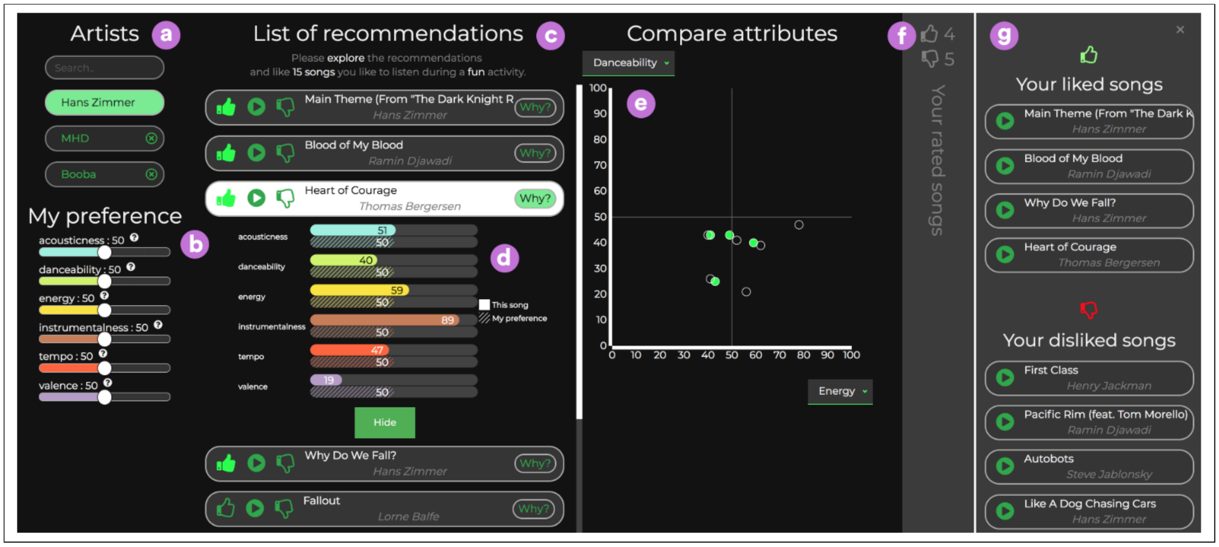


Figure 22: Millecamp's second explanatory explanatory interface from Millecamp et al. (2019).

with high musical sophistication, Spotify usage, and visual working memory interacted more with the radar chart interface. On the other hand, users with low musical sophistication and visual working memory interacted more with the sliders interface. Through evaluating this study, Millecamp also found out that a user's opinion of the different interfaces was not dependent upon their age, gender, recommender system familiarity or stance towards recommender systems. According to Millecamp et al. (2018) *"personal characteristics do not seem to play much role on perception of the visual control techniques in music recommendation"*. The next interface focused more specifically on the effects of personal characteristics in regard to recommendation explanations by comparing two interface versions, one without explanations and one with explanations. Millecamp et al. (2019) expanded upon the user characteristics tested in the previous study. In addition to visual working memory, they tested two other cognitive skills, need for cognition (NFC) and visualization literacy. The first tested how willing the user was to engage with different components of the interfaces, and the second tested the user's ability to understand different visual aspects of the interfaces such as the scatterplot. The locus of control personality trait was also tested, which according to Millecamp et al. (2019, p. 2), *"measures the degree of control individuals perceive towards outcomes"*. Of the 105 users recruited via the online crowdsourcing website Amazon Mechanical Turk, 71 answered questionnaires that were considered to be consistent and valid enough to be evaluated. The study set up the tasks and questionnaire more or less in the same fashion, although this time they did not use the ResQue framework for evaluation. Instead they used a questionnaire based on the metrics of recommender effectiveness, good understanding, trust, novelty, use intentions, satisfaction, and confidence, as well as user perception of explanation components. The questionnaire consisted of questions which followed a Likert scale[5] format.

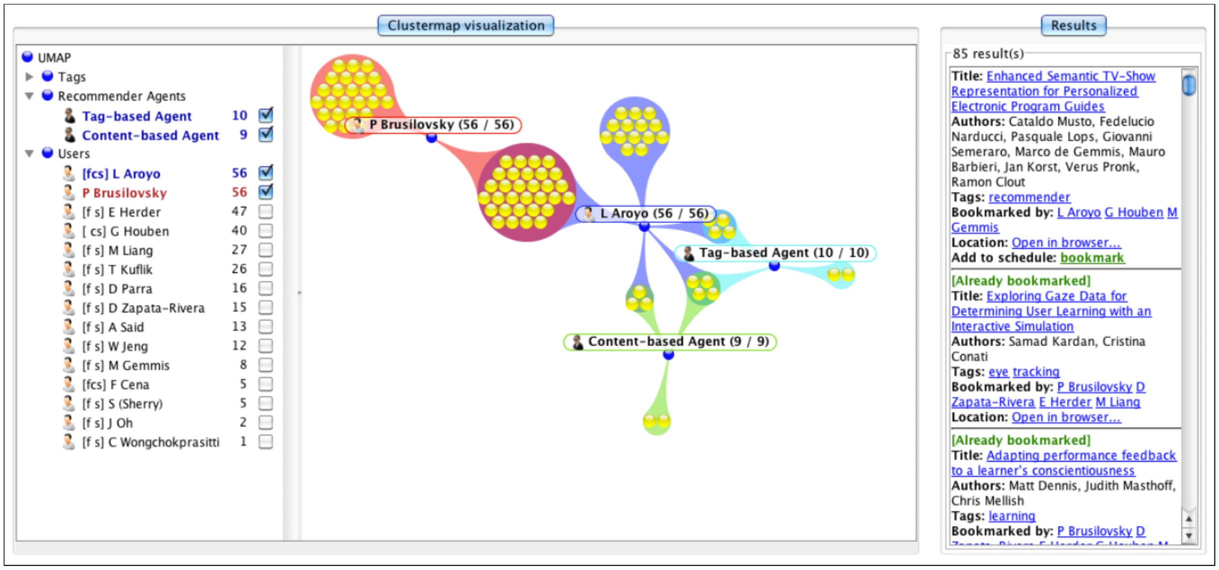
Millecamp et al. (2019) wrote that the user study revealed that participants spent less time listening to songs when testing the explanation interface than they did when they tested the baseline. This shows that explanations helped participants understand song recommendations, since without explanations, users were forced to listen to songs more often in order to make decisions. Interestingly, high NFC users preferred the baseline and the use of explanations did not increase confidence, whereas the opposite was true for low NFC users. In terms of personal characteristics, Millecamp states that there did not seem to be any correlation between tech-saviness and explanation preferences. In terms of effects of personal characteristics, Millecamp found in this study that there was a correlation between the way users interacted and perceived the recommender systems which was also affected by the addition of explanations. In closing, Millecamp et al. (2019, p. 10) asserts that when designing for explanatory interfaces, *"like the recommendations themselves, explanations should be personalized for different groups of end-users. Secondly, to reduce information overloading, users should be able to choose whether or not they wish to see explanations. Thirdly, explanation components should be flexible enough to present varying level of details depending on a user's preference"*.

2.3.4 TalkExplorer, SetFusion, and IntersectionExplorer

In order to combat information overload, conventional recommenders tend to be straightforward, creating interfaces which are not very interactive by only displaying textual ranked lists. On the other hand, users feel as though they have more control over their recommendations when they can interact with the system through visualizations of multitudinous data. TalkExplorer, SetFusion, and IntersectionExplorer are all interactive article recommender system interfaces for the recommender system Conference Navigator to help academic conference attendees discover lecture information and research papers while at a conference. Figures 23, 24, and 25 shows the design evolution. In different ways, they each try to incorporate features from both traditional and ocular systems by itemizing and visualizing recommendation results.

Figure 23 presents the TalkExplorer interface, which lets users explore talks through both content and social relevance perspectives to improve a user's ability to grasp the context of an item and why it is relevant for them. The left column is the entity section where users can select what information, tags or users or recommender agents. they want to compare their relationships and receive recommendations from. The middle is the canvas area where their possible recommendations are visualized in a cluster-map to compare said entities interrelationships based on their previous personal choices from the entity section. Each entity is presented as a large

[5] https://www.britannica.com/topic/Likert-Scale



- UMAP
  - Tags
  - Recommender Agents

Tag-based Agent	10	☑
Content-based Agent	9	☑

  - Users

[fcs] L Aroyo	56	☑
P Brusilovsky	56	☑
[f s] E Herder	47	□
[cs] G Houben	40	□
[f s] M Liang	27	□
[f s] T Kuflik	26	□
[f s] D Parra	16	□
[f s] D Zapata-Rivera	15	□
[f s] A Said	13	□
[f s] W Jeng	12	□
[f s] M Gemmis	8	□
[fcs] F Cena	5	□
[f s] S (Sherry)	5	□
[f s] J Oh	2	□
[f s] C Wongchokprasitti	1	□

85 result(s)

Title: Enhanced Semantic TV-Show Representation for Personalized Electronic Program Guides
Authors: Cataldo Musto, Fedelucio Narducci, Pasquale Lops, Giovanni Semeraro, Marco de Gemmis, Mauro Barbieri, Jan Korst, Verus Pronk, Ramon Clout
Tags: recommender
Bookmarked by: L Aroyo G Houben M Gemmis
Location: Open in browser...
Add to schedule: bookmark

[Already bookmarked]
Title: Exploring Gaze Data for Determining User Learning with an Interactive Simulation
Authors: Samad Kardan, Cristina Conati
Tags: eye tracking
Bookmarked by: P Brusilovsky D Zapata-Rivera E Herder M Liang
Location: Open in browser...

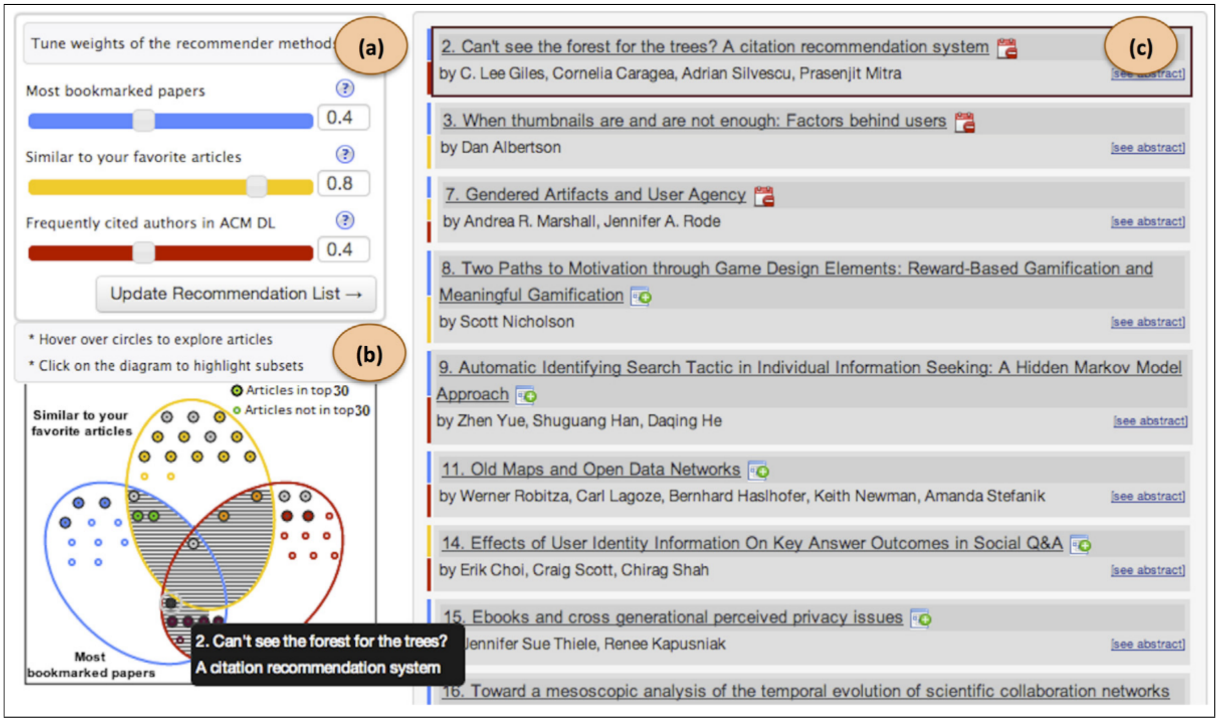
[Already bookmarked]
Title: Adapting performance feedback to a learner's conscientiousness
Authors: Matt Dennis, Judith Masthoff, Chris Mellish
Tags: learning
Bookmarked by: P Brusilovsky D Zapata-Rivera E Herder G Houben M

Figure 23: TalkExplorer from Verbert et al. (2013).

colored bubble which contains a number of Individual items presented as small yellow circles. Overlapping circles show which items two entities have in common. Rightly so, item suggestions are shown in the results panel on the right.

User studies were performed at two conferences in 2012 with 14 and 7 participants using the think-aloud method. Afterwards, users were asked to answer survey questions about the usefulness and effectiveness of the interface visualizations to help them navigate their conference reference needs. Users liked the interactive components to the system, however minimal, which lead to increased user control. The comparison agent of multiple entities increased user trust and transparency, although there were scalability issues. For example, tags were more effective when used in conjunction with either a recommender agent or user rather than when used alone. However, some found the visualization to be complicated and inconvenient to use which resulted in decreased usability. Verbert et al. (2013, p. 352) tells that overall, *"the interface serves to both explain the provenance of recommendations in a transparent way and to support exploration and control by end users"*.

Parra et al. (2014) and Parra & Brusilovsky (2015) both explore SetFusion, which elaborates upon TalkExplorer by allowing users to fuse and inspect different recommendation methods. The interactive clustermap from TalkExplorer was replaced with a simplified and more user friendly venn-diagram. The social features from the previous interface were also discarded. Section (a) in Figure 24 shows how users can easily control the three algorithms, bookmarking popularity, a content-based method, and author-based popularity, by way of sliders. Section (b) helps users learn more about the different recommender approaches through the interactive set-based venn-diagram visualization, and users receive detailed textual information about these recommendations in (c). In



Tune weights of the recommender methods

Most bookmarked papers 0.4

Similar to your favorite articles 0.8

Frequently cited authors in ACM DL 0.4

* Hover over circles to explore articles
* Click on the diagram to highlight subsets

2. Can't see the forest for the trees? A citation recommendation system

2. Can't see the forest for the trees? A citation recommendation system
by C. Lee Giles, Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra [see abstract]

3. When thumbnails are and are not enough: Factors behind users
by Dan Albertson [see abstract]

7. Gendered Artifacts and User Agency
by Andrea R. Marshall, Jennifer A. Rode [see abstract]

8. Two Paths to Motivation through Game Design Elements: Reward-Based Gamification and Meaningful Gamification
by Scott Nicholson [see abstract]

9. Automatic Identifying Search Tactic in Individual Information Seeking: A Hidden Markov Model Approach
by Zhen Yue, Shuguang Han, Daqing He [see abstract]

11. Old Maps and Open Data Networks
by Werner Robitza, Carl Lagoze, Bernhard Haslhofer, Keith Newman, Amanda Stefanik [see abstract]

14. Effects of User Identity Information On Key Answer Outcomes in Social Q&A
by Erik Choi, Craig Scott, Chirag Shah [see abstract]

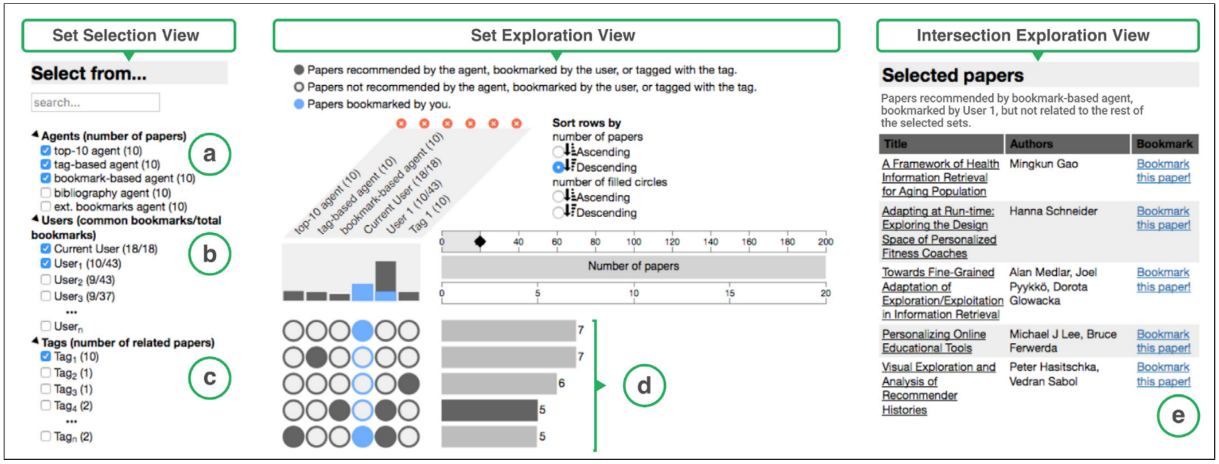
15. Ebooks and cross generational perceived privacy issues
Jennifer Sue Thiele, Renee Kapusniak [see abstract]

16. Toward a mesoscopic analysis of the temporal evolution of scientific collaboration networks

Figure 24: SetFusion from Parra & Brusilovsky (2015).

addition to testing the interface at a conference like Parra et al. (2014) did with TalkExplorer, Parra & Brusilovsky (2015) had SetFusion tested by PhD students and graduates who had specifically not attended more than one iConference in the previous 3 years. Both studies examined user experience by objectively, subjectively and behaviorally evaluating controllability, interface design, and user characteristics.

Results from these two studies revealed positive incites and nearly all users preferred this interface over the baseline. Users found it more useful than the baseline, easy to use, would use the system again and recommend it to a colleague. Not surprisingly, users were more engaged with the system when testing occurred during a conference, as the system was more relevant to use at that particular moment. The interactive design elements helped to explain recommendations which in turn increased user trust in those recommendations, though this proved to be true much more-so for women than for men. Familiarity and prior system knowledge highly affected user experience. Users who had previously used Conference Navigator were less likely to interact with any of the new features, and those already acquainted with interactive sliders were more reluctant to test the venn-diagram. In contrast, those familiar with recommender systems were not as interested in reading abstracts, but rather drawn towards the explanatory and controllable aspects of the interface. Even with design changes, there was some cognitive overload. Parra & Brusilovsky (2015) disclosed that some users found it difficult to concurrently interact with more than three entities, and others mentioned that certain aspects of



Selected papers

Papers recommended by bookmark-based agent, bookmarked by User 1, but not related to the rest of the selected sets.

Title	Authors	Bookmark
A Framework of Health Information Retrieval for Aging Population	Mingkun Gao	Bookmark this paper!
Adapting at Run-time: Exploring the Design Space of Personalized Fitness Coaches	Hanna Schneider	Bookmark this paper!
Towards Fine-Grained Adaptation of Exploration/Exploitation in Information Retrieval	Alan Medlar, Joel Pyykkö, Dorota Glowacka	Bookmark this paper!
Personalizing Online Educational Tools	Michael J Lee, Bruce Ferwerda	Bookmark this paper!
Visual Exploration and Analysis of Recommender Histories	Peter Hasitschka, Vedran Sabol	Bookmark this paper!

Figure 25: IntersectionExplorer from Cardoso et al. (2018).

the sliders and the venn-diagram were redundant.

The newest approach in the evolution of these conference talk recommendation interfaces is IntersectionExplorer by Cardoso et al. (2018). A combination of the two prior interfaces, it introduces an innovative set-based visualization technique called UpSet. According to Cardoso et al. (2018, p. 76), UpSet is scalable, able to explore big complex data loads for *"the analysis of sets, their intersections, and aggregates of intersections"*. The interface may seem slightly simpler with easier to follow visualizations like line graphs, but it also gives more power to users by providing more options like by reintroducing social features.

Figure 25 exposes its three main parts: Set Selection View, Set Exploration View, and Intersection Exploration View. While its setup is keeping in line with the two previous models, there is now a search box for explicit searches and there are added explanations. The largest change in Cardoso et al. (2018, p. 79) is the center area, which *"allows users to explore the intersections between the selected sets of papers as rows"*, where *"currently explored intersections/rows are colored in darker gray and the intersections with the current user's bookmarked papers are highlighted in blue"*.

Three separate user testing sessions were completed in different contexts to study performance, user experience and user interaction. A small study was initially done after a conference where a baseline system was compared to IntersectionExplorer. Due to study limitations, two additional studies during conferences were completed, one with less and one with highly technically oriented participants. A think-aloud protocol was also used, along with a questionnaire based on ResQue and Knijnenburg's evaluation framework, (Knijnenburg et al., 2012), which used a Likert scale from 1 to 5.

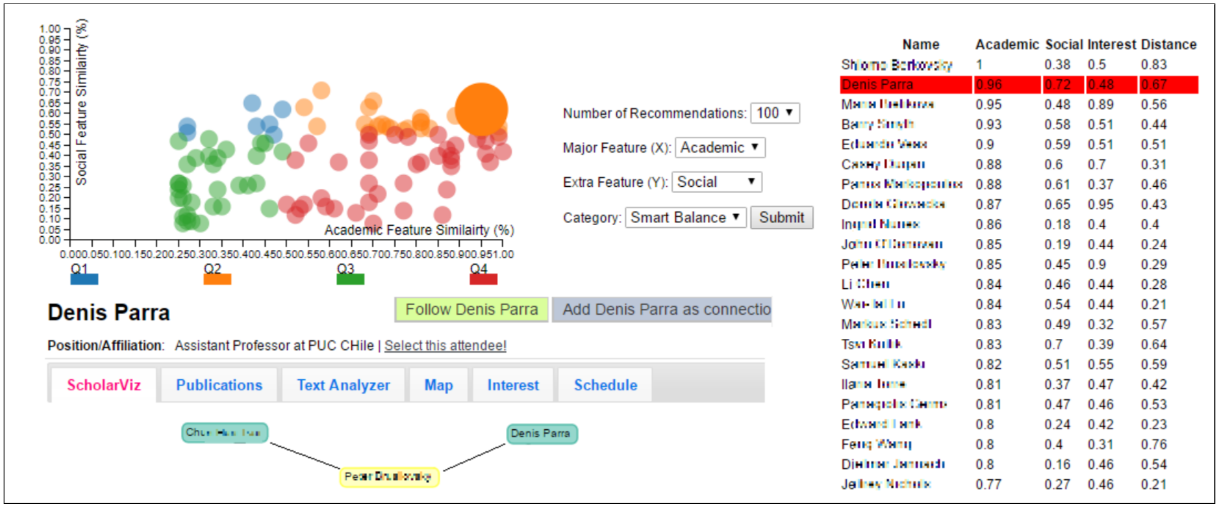
In regard to how IntersectionExplorer held up against the baseline system, Cardoso et al. (2018) declared that users perceived it to be more effective, dispensing higher quality results.

Users were more willing to try this tool over the baseline. It led to higher levels of satisfaction and trust. ‘Interface Adequacy’, ‘Information Sufficiency’, and ‘Control’ had a mean rating of 4 out of 5, however ‘Interaction Adequacy’ scored only 3.5 out of 5, indicating that the system could benefit from including more interactive components. Overall, it took a bit more time to learn how to use IntersectionExplorer than conference navigator, though overtime time efficiency and interaction effort are proportionate for both interfaces. The majority of metrics evaluated received a positive median result of 4 out of 5 for all three studies. Areas to improve upon which were given slightly lower overall ratings from highest to lowest were ‘Control’, ‘Use Intention’, ‘Information Sufficiency’, ‘Interaction Adequacy’, ‘Interface Adequacy’, ‘Fun’, ‘Choice Difficulty’, and ‘Effort’. Cardoso et al. (2018, p. 91) says in conclusion, that a *“multi-perspective approach to recommendation exploration has great promise as a way of addressing the complex human-recommender system interaction problem”*, even though user experience for less technically-oriented users may not be high.

TalkExplorer, SetFusion, and IntersectionExplorer were developed as interactive set-based visualization tools for conference talk recommendation. Through multiple perspectives of relevance, one can explore collections of items simultaneously. Cardoso et al. (2018) reports that these collections are organized meaningfully and expressed textually and visually, where the visual aids insure that users can see information without the need to read it. This can potentially increase user trust. All these studies have shown that transparency can be improved by visually combining multiple sources of recommendations and by providing explanations about those recommendations. However, they have also shown that complex visualizations can be too difficult for non-technical users, and that people may prefer a traditional list format as it is more familiar due to its widespread use in mainstream recommender systems.

2.3.5 Scatter Viz and Relevance Tuner

Just like TalkExplorer, SetFusion and IntersectionExplorer, Scatter Viz (Tsai & Brusilovsky, 2018) and Relevance Tuner (Tsai et al., 2019; Tsai & Brusilovsky, 2019) have also relied on Conference Navigator to test social recommendation explanations in the academic conference domain. These projects were erected to investigate similar issues like controllability, explanation visualizations, serendipity, and transparency. Previous studies show that explanations and controllability can both lead to improved transparency and perceived serendipity through interactive components such as visualizations. As reported by Tsai & Brusilovsky (2019), explanation visualizations support a user’s understanding of how their actions impact the system and can help expose the reasoning behind a recommendation which contributes to the overall inspectability and transparency of the recommendation process and system as a whole.



Number of Recommendations: 100 ▾

Major Feature (X): Academic ▾

Extra Feature (Y): Social ▾

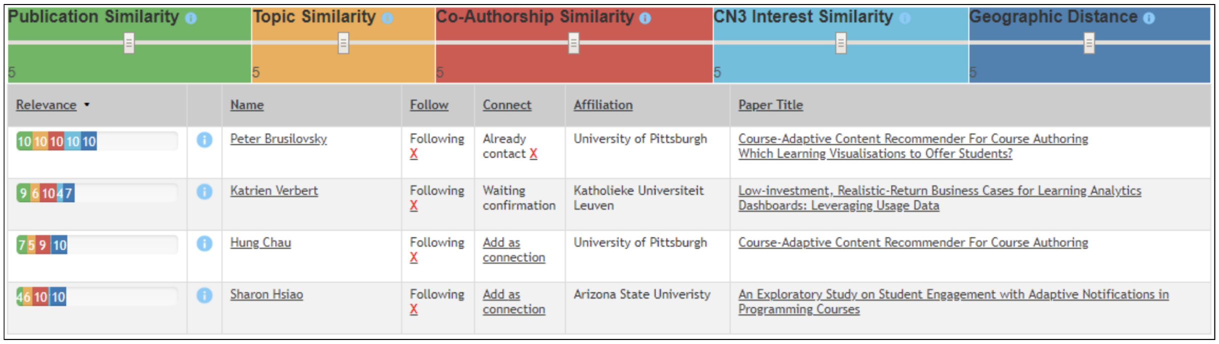
Category: Smart Balance ▾

Denis Parra

Position/Affiliation: Assistant Professor at PUC CHile | Select this attendee!

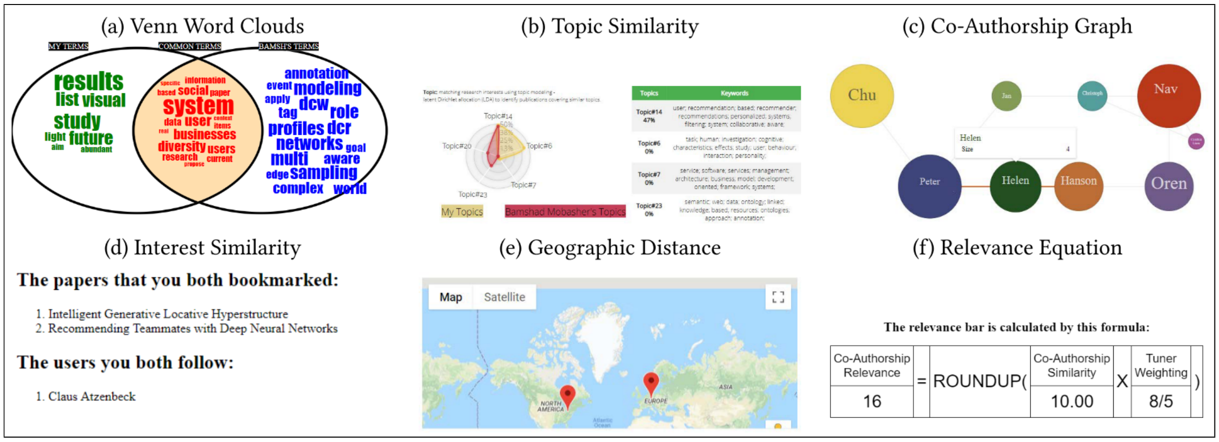
Name	Academic	Social	Interest	Distance
Shlomo Berkovsky	1	0.38	0.5	0.83
Denis Parra	0.96	0.72	0.48	0.67
Maria Bielikova	0.95	0.48	0.89	0.56
Barry Smyth	0.93	0.58	0.51	0.44
Eduardo Veas	0.9	0.59	0.51	0.51
Casey Dugan	0.88	0.6	0.7	0.31
Panos Markopoulos	0.88	0.61	0.37	0.46
Dorota Glowacka	0.87	0.65	0.95	0.43
Ingrid Nunes	0.86	0.18	0.4	0.4
John O'Donovan	0.85	0.19	0.44	0.24
Peter Brusilovsky	0.85	0.45	0.9	0.29
Li Chen	0.84	0.46	0.44	0.28
Wei-Hu Lu	0.84	0.54	0.44	0.21
Markus Schedl	0.83	0.49	0.32	0.57
Tsvi Kuflik	0.83	0.7	0.39	0.64
Samuel Kaski	0.82	0.51	0.55	0.59
Ilaria Torre	0.81	0.37	0.47	0.42
Panagiotis Germanakos	0.81	0.47	0.46	0.53
Edward Lank	0.8	0.24	0.42	0.23
Feng Wang	0.8	0.4	0.31	0.76
Dietmar Jannach	0.8	0.16	0.46	0.54
Jeffrey Nichols	0.77	0.27	0.46	0.21

Figure 26: Scatterviz from Tsai & Brusilovsky (2018).



Relevance ▾		Name	Follow	Connect	Affiliation	Paper Title
10 10 10 10 10	ⓘ	Peter Brusilovsky	Following X	Already contact X	University of Pittsburgh	Course-Adaptive Content Recommender For Course Authoring Which Learning Visualisations to Offer Students?
9 6 10 4 7	ⓘ	Katrien Verbert	Following X	Waiting confirmation	Katholieke Universiteit Leuven	Low-investment, Realistic-Return Business Cases for Learning Analytics Dashboards: Leveraging Usage Data
7 5 9 10	ⓘ	Hung Chau	Following X	Add as connection	University of Pittsburgh	Course-Adaptive Content Recommender For Course Authoring
4 6 10 10	ⓘ	Sharon Hsiao	Following X	Add as connection	Arizona State Univeristy	An Exploratory Study on Student Engagement with Adaptive Notifications in Programming Courses

Figure 27: Relevance Tuner+ from Tsai et al. (2019).



The papers that you both bookmarked:

1. Intelligent Generative Locative Hyperstructure
2. Recommending Teammates with Deep Neural Networks

The users you both follow:

1. Claus Atzenbeck

The relevance bar is calculated by this formula:

Co-Authroship Relevance	=	ROUNDUP(	Co-Authorship Similarity	X	Tuner Weighting	)
16			10.00		8/5	

Figure 28: Relevance Tuner+ explanation styles from Tsai & Brusilovsky (2019).

The first study was with ScatterViz and is explained by Tsai & Brusilovsky (2018). It had 25 scholarly participants compare a scatterplot interface to a ranked list baseline during the 2017 Intelligent User Interfaces Conference. The SCATTER interface called ScatterViz, as shown in Figure 26, consists of a scatter plot, control panel, ranked list, and user profile page. For the baseline interface RANK, the scatter plot component was removed. User testers filled out a demographic questionnaire, were asked to perform three tasks with each interface, and then filled out a post questionnaire. Tsai & Brusilovsky (2018) notes that ScatterViz considerably outperformed the baseline in the areas of 'Trust', 'Supportiveness', 'Interest', 'Satisfaction', 'Intention to Reuse' and 'Enjoyment', and received slightly higher scores regarding the facilitation of diversity exploration and on how beneficial the interface was. Though not statistically significant, RANK interestingly received marginally better marks for explanations. This implicates the notion that visualizations were particularly convincing in getting across to the user why certain items were recommended and helped them make better choices, and therefore other explanations were viewed as less important. The absence of this in RANK had the opposite effect of leading users to investigate other options more so than was needed in SCATTER, so explanations were seen as more important. So, while the majority of users preferred ScatterViz, RANK was easier to use and felt more well-known to them.

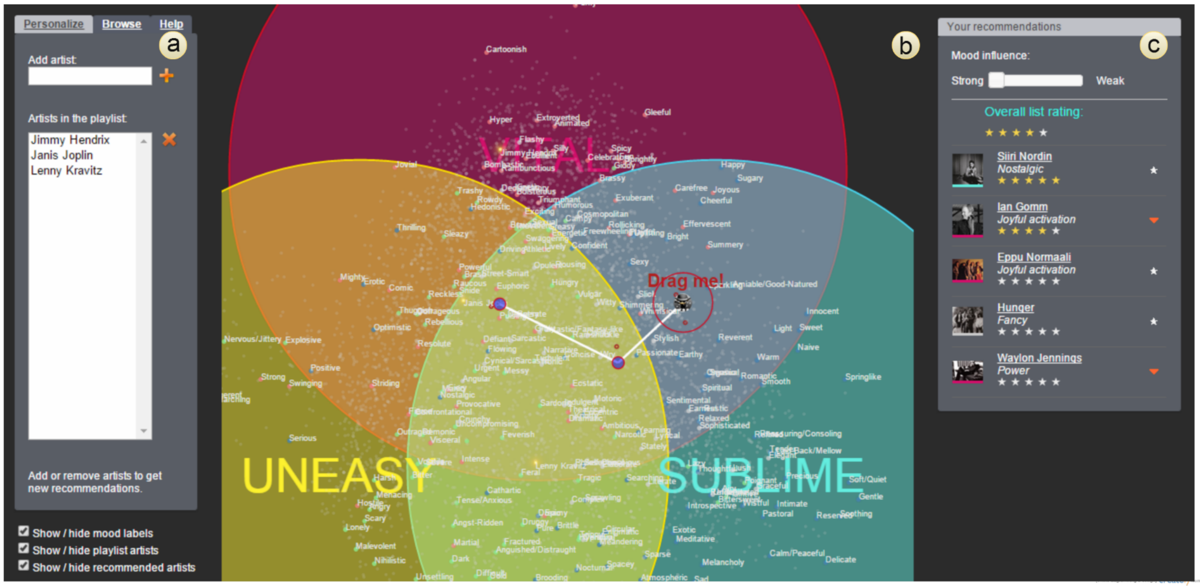
The results of the first study led the team to create Relevance Tuner. Relevance Tuner gives users more control over their recommendations as one can easily tune the different categories depending on how relevant that feature is to them. Tsai & Brusilovsky (2018) chose the color-coded features adjustable through sliders to be 'Publication Similarity', 'Co-Authorship Similarity', 'Conference Navigator Interest Similarity', 'Geographic Distance', and 'Social Context'. Each item listed shows its relevance in relation to the chosen preferences in the form of a stackable score bar with white numbers, where each bar color correlates with its corresponding feature. 20 scholars who attended the 25th Conference on User Modeling, Adaptation and Personalization in 2017 compared Relevance Tuner to ScatterViz, using the same procedures as in the first study. While both interfaces were given positive feedback, it is clear that users have different preferences and will explore different features depending on interface functionality and the task they need to achieve. TUNER fared much better in terms of 'Supportiveness', 'Interest', 'Facilitation', 'Satisfaction', 'Intend to Reuse', and 'Usefulness'. According to Tsai & Brusilovsky (2018, p. 247), users relied *"more on the explanation function and multi-relevance visualization to explore diversity-oriented tasks"* in SCATTER, whereas *"instead of checking the explanation functions"* in TUNER, more time was spent *"inspecting the information on each row"*. Even though ScatterViz was more enjoyable to use and due to the visualization, the interface was determined to be more trustworthy, people preferred to use TUNER. This was due to the fact that it required less learning effort, had an integrated design, and was more familiar to users.

The next stage of this project focused on modifying Relevance Tuner by adding explanations. This controllable interactive social recommender interface with explanations was named Relevance Tuner+ and is described by Tsai & Brusilovsky (2019) and Tsai et al. (2019). As shown in Figure 27, users can now click an explanation icon for more information about the similarity breakdown of each category. A within-subjects design study was completed with 33 voluntary participants who had attended 1 of 3 different conferences. Subjective user feedback and behavior were evaluated across all design components. The method for testing as well as the framework for evaluation was based off of the previous studies. Participants compared Relevance Tuner+ to a baseline, which was Relevance Tuner+ without explanations. When the user clicks on the explanation icon '(i)', depending on which social recommendation they want explained, a different interface is presented. These six different possible explanatory interfaces which were tested are shown in Figure 28.

Overall, the Tuner+ interface was significantly better in providing explainability which indicates better system transparency. In terms of explanation visualizations, for the metric *understandable visualization*, the explanation styles that worked best were word cloud, graph, and map. For the metrics *persuasion and acceptance*, graph came out on top. The word cloud and graph visualizations were the most satisfying to participants. The mere addition of the explanation icon showed no impact on UX dimensions, however only half of the participants actually clicked it. Sliders were used to a greater extend when completing tasks, and tasks took longer to complete for those who viewed explanations. In the report Tsai & Brusilovsky (2019, p. 394), results show a significant decrease in perceived ease-of-use and control *"if the explanations were presented"*, implying that *"users might experience difficulties with a possibly overwhelming amount of information"*. Yet at the same time Tsai & Brusilovsky (2019, p. 394) notes that the *"participants who perceived more transparency will positively associate this with perceived control and satisfaction"*, implying that the *"extra amount of information would not impair those who perceived higher system explainability and understanding"*. The results of this study indicate that information completeness is not always the best approach when providing explanations as it may lead to reduced levels of user confidence and satisfaction. In consonance with Tsai & Brusilovsky (2019, p. 395), it is imperative to filter excess information by only displaying explanations when of relevance to the user *"based on the context of information-seeking tasks"*.

2.3.6 Moodplay

According to Andjelkovic et al. (2016), Moodplay is an interactive affect-based and audio content-based music recommendation web application. By letting users explore artists using affective meta-data in a latent space, this interactive experience gives users more control which


Personalize Browse Help

Add artist:

Artists in the playlist:

Jimmy Hendrix
Janis Joplin
Lenny Kravitz

Add or remove artists to get new recommendations.

☑ Show / hide mood labels
☑ Show / hide playlist artists
☑ Show / hide recommended artists

Your recommendations

Mood influence:

Strong Weak

Overall list rating:

Siiri Nordin
Nostalgic

Ian Gomm
Joyful activation

Eppu Normaali
Joyful activation

Hunger
Fancy

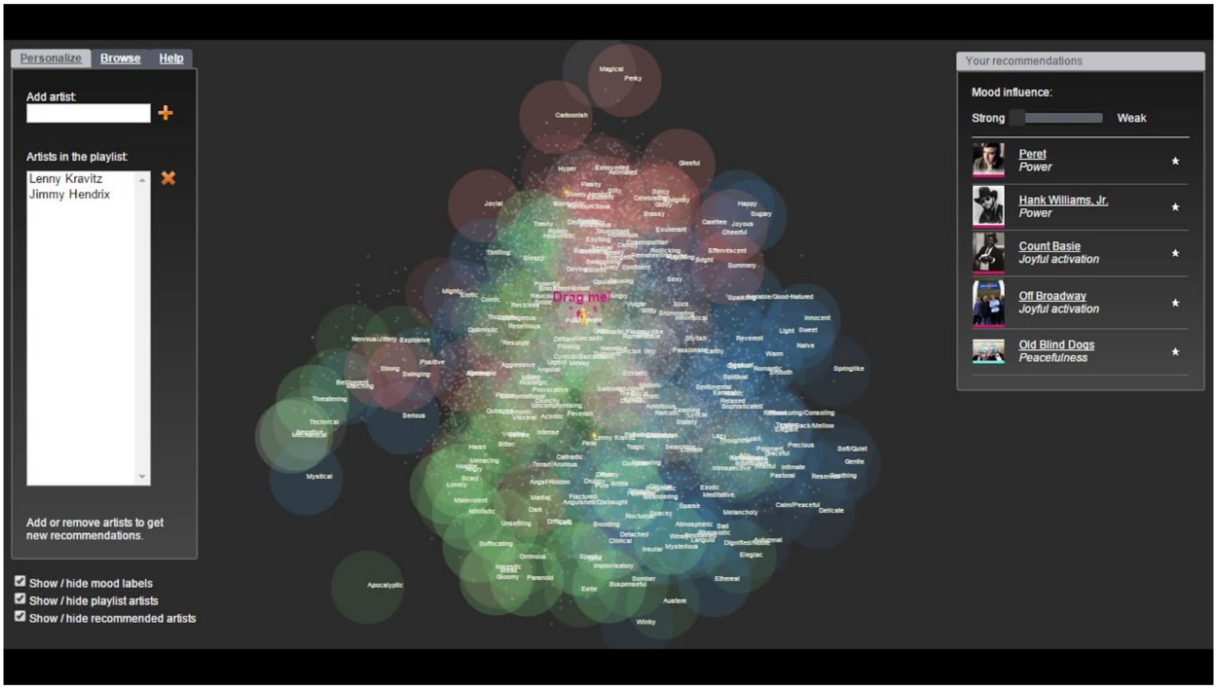
Waylon Jennings
Power

Figure 29: Moodplay's first interface from Andjelkovic et al. (2016).

helps with both the acceptance and understanding of recommendations. This research analyzed how to visualize music using affective data, how users interacted with these visualizations in regard to explanation and control, as well as user experience to discover how to find a balance in the amount of interaction needed in a music recommender system.

Figures 29, 30, and 31 depict Moodplay's interface design evolution. Andjelkovic et al. (2016, p. 276) states that the interface of the two first figures is setup from left to right as such: *"a pane for entering artist names, a latent mood space visualization, and a recommendation list, along with slider for adjusting mood influence"*. Each individual node displays the name of a distinct affect, with artist recommendation nodes within. An artist can be inside of several different mood nodes. Based on the weakness or strength position of the mood slider, the circumference around this node becomes larger or smaller respectively. A stronger mood influence produces a smaller portion of select highly similar artists to be recommended to the user based on the seed artist(s), whereas a weaker influence produces more but less similar artists to be recommended. Similarity scores created for artists in order to categorize affect, were calculated differently than the typical arousal / valence scope. Artists were categorized into either 'Sublimity', 'Unease', 'Style', or 'Vitality', based on comparing data from the similarity tool WordNet[6], Rovi[7] mood meta-data, and 45 GEMS mood words. In the right-side panel, recommendations are displayed with a picture of the artist, a colored line underneath this picture representative of its mood category, the artists name, and the name of the particular mood category.

[6] https://wordnet.princeton.edu/
[7] http://developer.rovicorp.com/



Add artist:

Artists in the playlist:
Lenny Kravitz
Jimmy Hendrix

Add or remove artists to get new recommendations.

☑ Show / hide mood labels
☑ Show / hide playlist artists
☑ Show / hide recommended artists

Your recommendations

Mood influence:
Strong Weak

Peret
Power

Hank Williams, Jr.
Power

Count Basie
Joyful activation

Off Broadway
Joyful activation

Old Blind Dogs
Peacefulness

Figure 30: Moodplay's second interface from Andjelkovic, Parra, & O'Donovan (2019).

The newest interface, Figure 31, eliminated the playlist creation feature in favor of a simpler design focusing more on explaining artist mood distribution. Whereas the previous interfaces displayed nodes representing mood categories, here each node represents a specific artist. Each artist node is clearly colored either green for sublimity, red for unease, white for style, or blue for vitality. When a node is selected the right-hand column comes into view displaying artist name, their mood distribution in either a line graph or pie chart, their current top three songs based on data from Spotify, Spotify's related artists, and Moodplay's mood based related artists. While Spotify's related artists section always remains, the mood based related artists change depending upon the strength of the artist relativity slider, now on the left-hand side. Each artist which is clicked on by a user is also saved in the bar on the bottom, so a user can see their music choice history. Users can rearrange the node layout in the 2D space by changing artist similarity through either Linear Discriminant Analysis (LDA) or T-distributed Stochastic Neighbor Embedding (t-SNE). Both of these are dimension reduction methods for displaying data. LDA, as its name suggests, is linear, placing the four main mood groups into individual sections with a slight overlap in the middle, as shown in Figure 31, similar to the layouts of the two previous interfaces. t-SNE on the other hand is non-linear, and distributes artist nodes in a well spread out scatter-plot, rather than a venn-diagram type style, while still keeping the relative distance between mood groups. Even though music is usually grouped into generalized categories such as genre, music similarity, is not as straight forward or linear as it may seem. Given the additional factor that most people generally listen to

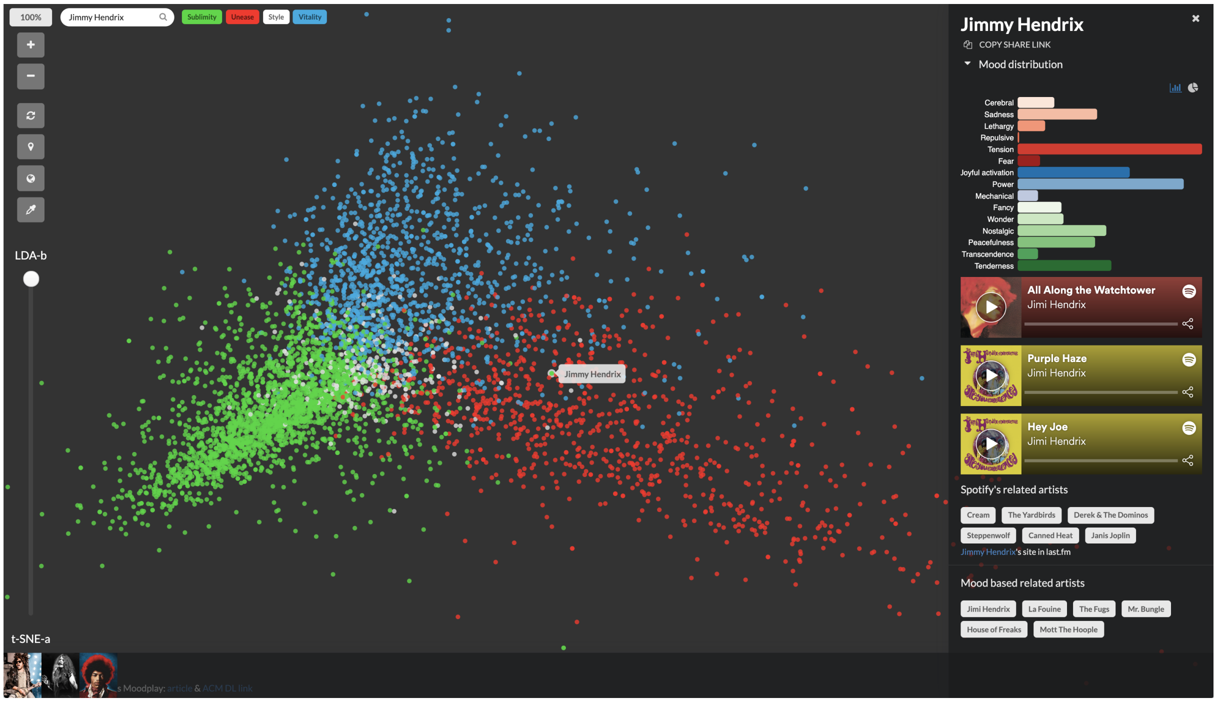


Figure 31: Moodplay's third interface from Andjelkovic, Parra, O'Donovan, & Herrera (2019).

multiple types of music it could be said that the newer t-SNE method produces more pleasing and diverse recommendations.

The most interesting detail of this interface is the explanation of the mood distribution, which is broken down into 15 different sub-sections which represent the overall 4 main mood categories. The line graph begins with 'Cerebral', a very light peach color, and the color becomes progressively redder and darker throughout the next categories of 'Sadness', 'Lethargy', 'Repulsive', 'Tension', and 'Fear'. It then jumps to a medium blue color for 'Joyful Activation', a lighter blue for 'Power', and a very light grayish periwinkle color for 'Mechanical'. Lastly comes 'Fancy' in an almost white green color, getting greener and darker through the categories of 'Wonder', 'Nostalgic', 'Peacefulness', 'Transcendence', and 'Tenderness'.

For the first experiment from Andjelkovic et al. (2016), 240 user testing sessions from Amazon Mechanical Turk were deemed valid. Users filled out a pre-survey to collect demographic information, performed main tasks, and filled out a post-survey to collect qualitative feedback. Participants deemed that 37 percent of artists were assigned to incorrect mood meta-data. Evaluation therefore focused on user characteristics, interaction with the interface and experience rather than recommendation accuracy. However, users perceived recommendation accuracy to be better due to latent space visualization. This could be because by visualizing music in a 2D-space, users were able to understand the system better. An evaluation of user testing revealed that older users spent more time on tasks. Andjelkovic et al. (2016, p. 278) also noted that *“gender influences degree of interaction”*, that there is *“some relation between level of interaction and cognitive strain”*, and one should be weary that *“too much information can lead to cognitive overload”*.

For the second experiment from Andjelkovic, Parra, & O’Donovan (2019), 279 new user tests were deemed valid. The testing was performed in the same manner as before. They reiterated that when mood is used to explain artist relationships in an interface which properly balances design and interactivity, user satisfaction and perceived recommendation accuracy was greater than the baseline. When designing the interface, Andjelkovic, Parra, & O’Donovan (2019, p. 155) kept in mind the importance of *“choice of colors, item sizes and transparency, dynamic labeling of mood nodes and node filtering based on mood categories”*. All of these features were implemented to better clarify mood classifications, which in turn aided in comprehension of recommendation explanations. Andjelkovic, Parra, & O’Donovan (2019, p. 155) also echoed the same sentiment from the previous one that *“increased system complexity beyond a comfortable threshold caused cognitive overload”*.

During user testing, users were not specifically asked about the performance of features. According to Andjelkovic, Parra, & O’Donovan (2019, p. 153), feedback did reveal that users would like the design to be *“more dynamic”* yet *“simpler and more concise”*, with users stating that *“there is a lot of text on the page and it’s a little overwhelming”*, and that the system was *“slow and laggy”*. There were also some comments on moods not matching up, similar to feedback from the previous study. Overall, participants in both studies liked MoodPlay. In general, it may still be too difficult for basic or typical users. The need to design a visually cleaner and simplistic interface is most likely what led to the creation of the newest interface shown in Figure 31. The introduction of the mood distribution chart helps explain to users the mood makeup for the specific artist chosen in much more detail than before. It also gives a better overview of how this artist fits into the latent space based on the current algorithm chosen, which can be changed by moving the left-hand slider. While the system removed the playlist creation ability, the bottom of the screen shows a history of artists previously visited in the form of an artist’s picture which gives the user quick access to revisit their profile. In terms of mobile devices, Moodplay has never been optimized for a small screen.

2.4 Summation of Prior Research

As previously mentioned, most past academic research has relied on enhancing algorithms to advance recommender systems, simplifying interfaces for minimal user interaction by focusing more on back-end solutions. Commercial music streaming services seem to favor this approach, as evident from their interface design. The opposite perspective, concentrating more on creative

Table 2: Recommendation explanation category clarifications for Table 3.

Category	**Clarification**
Education	A user's ability to learn from the system.
Effectiveness	The systems ability to help the user make better decisions, in addition to personalization.
Efficiency	Users can complete tasks fast and a low cognitive effort is needed to use the system.
Transparency	The system is able to explain the how and why of things.
Usability	User satisfaction and ease of use.
Usefulness	Use intention and value.

and interactive front-end solutions, is becoming more and more prevalent in recommender systems research today. These elaborate interfaces tackle issues such as of lack of transparency, control, and personalization which occur in algorithm focused ones. This becomes apparent after reading through this literature review.

An analysis of related works is shown in Table 3, which compares visualization types, system types, data collection methods, and results in terms of explanations. A check mark means that that particular requirement has been met. A check minus mark means that it is unclear whether or not a particular requirement was met. Therefore, a check minus is not necessarily negative. For example, maybe results showed that almost a majority of user testers agreed that this requirement was met, however a few were quite adamant that this requirement was definitely not met. If a requirement was met but was not statistically significant, in comparison to other results from within the same test, but not compared to results from other research papers, then no check mark was given. In regard to the last section, there is a need to expand upon some the categories listed in the table as they are slightly different than the explanatory goal definitions stated previously. Different explanation types were grouped together to make them more easily relatable to this thesis. An overview of these category clarifications is found in Table 2.

Overall, all systems received overwhelmingly positive results where the majority of user testers in every study preferred at least some sort of minimal explanation over no explanation. The usability and usefulness requirements were more or less met in every case, and questionnaires were utilized in every single system evaluation. The main visualization types used were icons, node-link diagram, node map, scatterplot, set-based, sliders, and text. None of the articles included interfaces which explicitly researched explanations which were optimized for mobile devices in the music domain. Therefore, it is difficult to speculate as to which types of explanation designs work better than others based solely on this literature review. Textual explanations proved beneficial in text-based interfaces whereas they were not as important when used together with visualizations. Explanations though in general seemed to improve the user's overall experience with a system, to a certain extent. A common sentiment amongst users was that many of the interactive interfaces were overly complex and text heavy interfaces lead to information overload. These types of systems were not suited for casual listeners. Concurrently, if the interface was too minimalistic and not interactive enough it was seen as mundane and lacking features. There needs to be a balance between user-centricity and simplicity. There appears to be a strong correlation between aspects such as having a high visual working memory / previous familiarity with the system / high musical sophistication / a low need for cognition and being able to understand and want to use more complex explanation and interactive elements. As such, users should be able to decide for themselves the level of explanation information they wish to receive. Designs where explanations were hidden until explicitly chosen by the user were favored over designs where explanations took up more screen real estate.

Table 3: Analysis of the current state-of-the-art comparing system types, evaluation methods, and results in terms of explanations.

	Visualization Type	Explicit Explanations	Mobile Optimized	Music Domain	Interview	Questionnaire	Task Performance	Think-aloud	User Behavior	Acceptance	Education	Effectiveness	Efficiency	Engagement	Transparency	Usability	Usefulness
MovieLens Herlocker et al. (2000)	Textual, Graph, Table	✓				✓				✓					✓		✓
MusicBox Lillie (2008)	2D Space Node Map			✓		✓						✓			✓-	✓	✓
Musicream Goto & Goto (2009)	2D Space Icon - Disc		✓	✓		✓						✓	✓			✓	✓
SmallWorlds Gretarsson et al. (2010)	Node-Link Diagram	✓		✓		✓				✓	✓				✓	✓	✓
MusiCube Saito & Itoh (2011)	Scatterplot			✓		✓	✓			✓		✓-				✓	
TasteWeights Bostandjiev et al. (2012)	Node-link Diagram	✓		✓		✓				✓	✓					✓	✓
CoFeel Chen & Pu (2013)	Radial View		✓	✓		✓		✓	✓					✓		✓	✓
Empatheticons Chen et al. (2014)	Icon		✓	✓		✓		✓	✓					✓		✓	✓
Musical Avatar Bogdanov et al. (2013)	Icon			✓		✓										✓	
Album Cover Space Lehtiniemi & Holm (2013)	Icon			✓	✓	✓										✓-	✓-
Avatar Lehtiniemi & Holm (2013)	Icon			✓	✓	✓				✓						✓	✓-
Potentiomer Lehtiniemi & Holm (2013)	Sliders			✓	✓	✓									✓-	✓	✓-
Talk Explorer Verbert et al. (2013)	Set-based (clustermap)	✓				✓	✓	✓	✓	✓		✓			✓	✓	✓
SetFusion Parra & Brusilovsky (2015)	Set-based (sliders/diagram)	✓				✓	✓		✓	✓		✓	✓	✓	✓	✓	✓
Intersection Explorer Cardoso et al. (2018)	Set-based (matrix)	✓				✓	✓	✓		✓		✓	✓-			✓	✓-
TagFlip Kamalzadeh et al. (2016)	Textual		✓	✓	✓	✓	✓						✓	✓	✓	✓-	✓
Scatter Viz Tsai & Brusilovsky (2018)	Scatterplot	✓				✓	✓			✓		✓-	✓-			✓-	✓-
Relevance Tuner + Tsai & Brusilovsky (2019)	Set-based (various)	✓				✓	✓		✓	✓	✓	✓	✓-	✓	✓	✓	✓
Moodplay Andjelkovic, Parra, & O'Donovan (2019)	2D Space Node Map	✓		✓		✓				✓			✓-		✓	✓-	✓
To Explain or Not to Explain Millecamp et al. (2019)	Sliders, Radar	✓		✓		✓	✓			✓-		✓	✓		✓	✓-	✓-

3 METHODS

This Chapter presents the specific procedures and techniques which were used to conduct this research. The specifics of this philosophical framework allow for the eventual replication of this study in addition to demonstrating research reliability. Section 3.1 illustrates the design science research cycle and explores the type of design science artifact created from this research. Section 3.2 clarifies the specific methods used for creating the pre-studies. Section 3.3 discusses the approach for creating the prototype. Section 3.4 explains the data sources used in the prototype. Section 3.5 describes the approach to the main study including which frameworks were used when designing user surveys.

3.1 Design Science

The field of design science helps to provide solutions to problems through not just research but also through the acquisition of research data and creation of innovative and purposeful artifacts. According to (Simon, 1996, p. 6), *"an artifact can be thought of as a meeting point - an 'interface' in today's terms - between an 'inner' environment, the substance and organization of the artifact itself, and an 'outer' environment, the surroundings in which it operates"*. Artifacts can be constructs, instantiations, methods, or models, as stated in Hevner et al. (2004). The design science research artifact created from this thesis' research is an explanatory interface for mobile devices. This model is a new visualization for a graphical user interface of a music recommender system. The outer environment to this artifact is the music recommender system itself, whereas the inner environment here refers to the explanation designs which lie within the outer environment. The research from the artifact may result in a meta-artifact containing design guidelines for music recommender systems in a mobile setting.

As explained in Hevner (2007), design science research focuses on three main cycles: relevance, design, and rigor. A model of the cycles is shown in Figure 32. In the relevance cycle, one must first determine real life applications for the research, what problem is being observed, and in what environment. As evident by previous research there is a great need for more and better recommendation explanations in the context of music recommender systems. The design cycle calls for a very iterative process, and is considered to be the heart of the project according to Hevner (2007). It was unfortunately outside the scope of this thesis to complete multiple large-scale iterative user testing studies for each design. However, a number of different design styles were explored and iterated upon in accordance with the design science guidelines as laid out in (Hevner, 2007, p. 4), which state to iterate design based on *"input from the relevance cycle"*. The most important factor in this research is that the artifact created will be a good

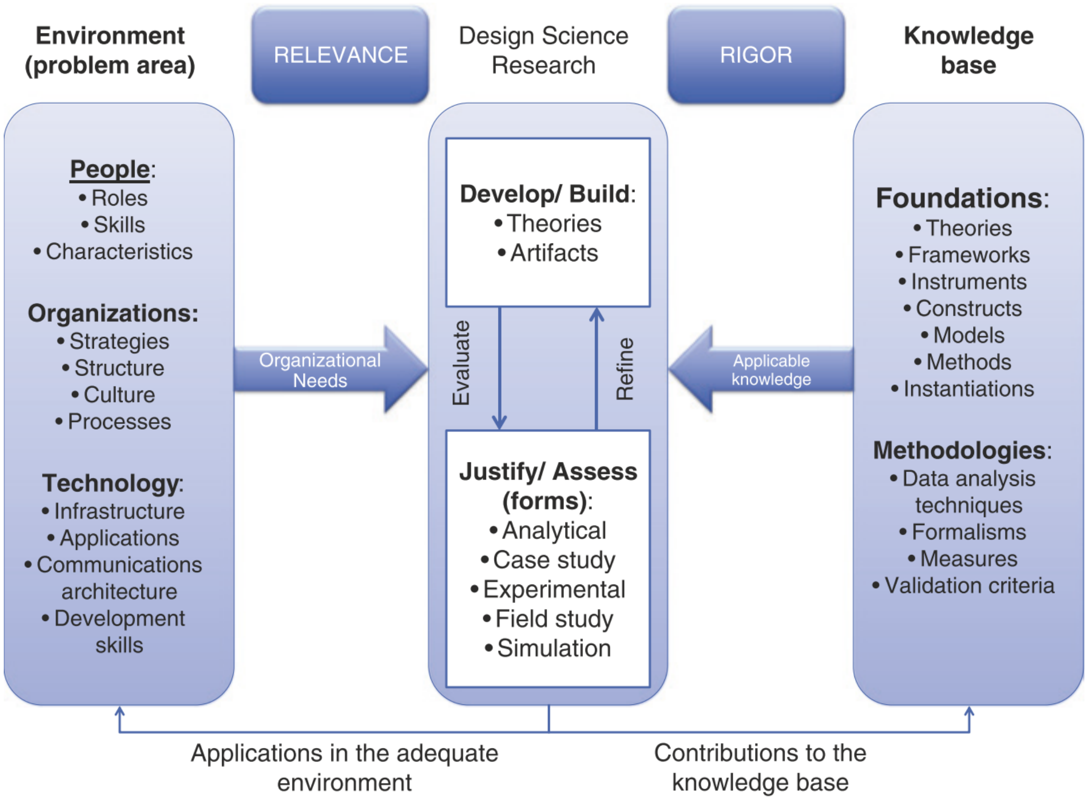
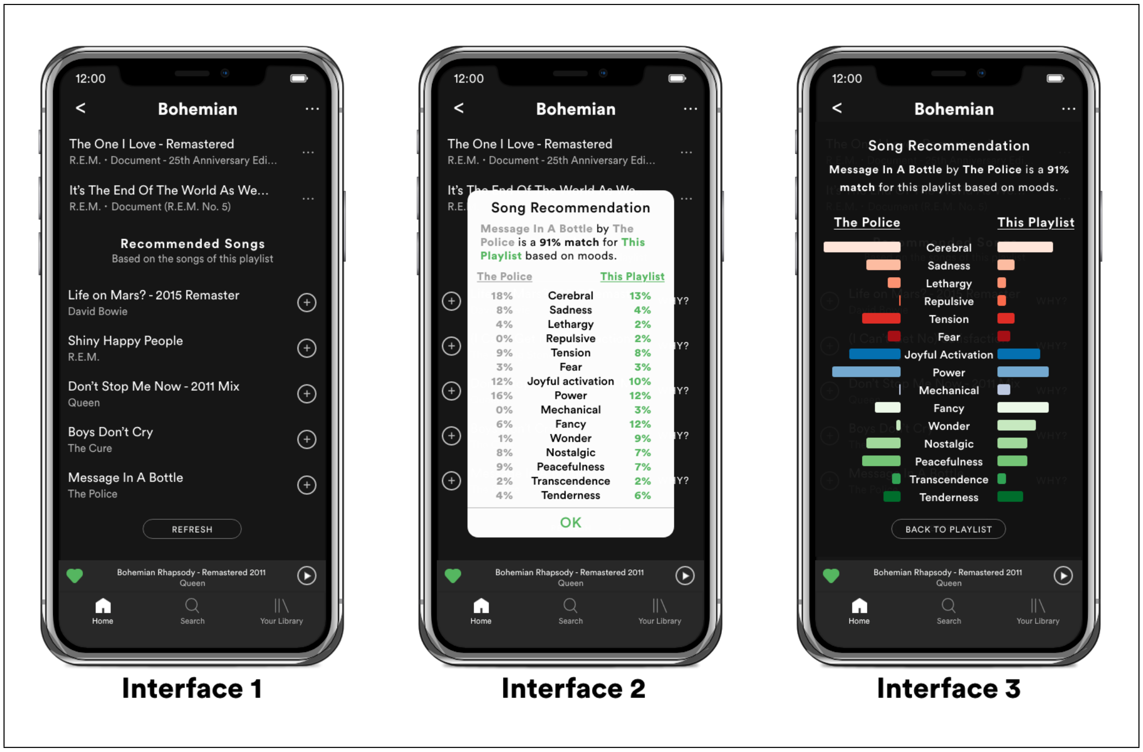


Figure 32: Design science research cycles from Dresch et al. (2015).

starting point for others to build upon in further studies. This is the idea behind the rigor cycle, that this research will contribute to the current knowledge base. As explained in Dresch et al. (2015), this knowledge base is then used to improve upon existing artifacts. Through the design cycle it became clear that the best course of action would be to create explanations based off of existing artifacts. By continuing to revamp previous designs created by others, this research is expanding the current knowledge base within this field of study.

3.2 Approach to Pre-Studies

Questionnaires for the two pre-studies were distributed via the social networking platform Facebook and the instant messaging platform Slack. The first used google forms and asked people about music recommendation explanations. The second used SurveyXact and asked people to rank different textual explanation designs. These questionnaires were distributed through the author's personal and work networks. Additionally, certain individuals also shared them further within their own networks. None of the questionnaires collected implicit or explicit, private or personal information. Based on the networks in which these were distributed, it can still be assumed that the majority participants were between the ages of 18 and 80, and were either


12:00
Bohemian
The One I Love - Remastered
R.E.M. · Document - 25th Anniversary Edi...
It's The End Of The World As We...
R.E.M. · Document (R.E.M. No. 5)
Recommended Songs
Based on the songs of this playlist
Life on Mars? - 2015 Remaster
David Bowie
Shiny Happy People
R.E.M.
Don't Stop Me Now - 2011 Mix
Queen
Boys Don't Cry
The Cure
Message In A Bottle
The Police
REFRESH
Bohemian Rhapsody - Remastered 2011
Queen
Home
Search
Your Library
Interface 1
Interface 2
Interface 3
OK
BACK TO PLAYLIST

Song Recommendation

Message In A Bottle by The Police is a 91% match for This Playlist based on moods.

The Police		This Playlist
18%	Cerebral	13%
8%	Sadness	4%
4%	Lethargy	2%
0%	Repulsive	2%
9%	Tension	8%
3%	Fear	3%
12%	Joyful activation	10%
16%	Power	12%
0%	Mechanical	3%
6%	Fancy	12%
1%	Wonder	9%
8%	Nostalgic	7%
9%	Peacefulness	7%
2%	Transcendence	2%
4%	Tenderness	6%

Figure 33: Baseline, textual, and visual explanation designs from MTurk Survey 1.

Norwegians or expats living in Norway. There is also the possibility than a small minority were American participants. It should be noted that some people who filled out the first pre-study questionnaire also completed the second one, though there is no explicit data to suggest how many.

A sequential within-subjects study approach was taken. Budiu (2018) explains that in a sequential within-subjects study, all subjects answer the same questions, in the same order, under the same conditions. In Survey 2, which asked users to rank different textual interfaces from best to worst, all of the different interface designs were grouped together and shown at the same time. Ranking the interfaces as a group instead of using a Likert scale for each individual interface reduced the likelihood for participants to rate their favorite interface as being best and all the others as being the worst. The group ranking system essentially forced participants to compare the interfaces to each other as opposed to assessing them separately.

3.3 Prototyping Overview

Dozens of preparatory sketches for the prototypes were made while researching related works. This led to creating three interfaces, a baseline and two explanatory interfaces. The three conditions tested were minimal explanations, textual explanations, and visual explanations. The data used for the explanations was mood data appropriated from Moodplay, and recommendations were taken from Spotify. As explanations are at the forefront of this research, it was not in the best interest of this study to create a novel interface. It was most important for users to rate the explanation designs themselves, as opposed to the overall interface design of the system. As the majority of people use the music streaming service Spotify, the baseline essentially just replicated this design. Design choices were based off of their mobile application from November 2019 and earlier. The hypothesis is that it will be easier for users to immerse themselves in the prototype and focus on the explanations if they are already used to the look and feel of its interface. In terms of explanations, the visual explanation design was based off of Moodplay's latest website design (Andjelkovic, Parra, O'Donovan, & Herrera, 2019). This visualization was then converted into text to create the textual design.

Eiband et al. (2018) explains that recommendation explanation designs are either normative or pragmatic. The former briefly refers to displaying explanations in a separate window, whereas the latter integrates them into the user interface. The baseline's minimal explanations are shown directly in the interface and are therefore considered to be pragmatic. In order not to overcrowd the interface, both explanatory interfaces have normative explanations where a user must click on a button labeled 'WHY?' in order to view the recommendation explanations. The baseline looks most similar to Spotify, as stated above, and is therefore more or less without explanations. It may be inappropriate to generalize that Spotify has no explanations, as they do explain certain features at a bare minimum. For example, at the bottom of a playlist under the title 'Recommended Songs', they include the phrase 'based on the songs of this playlist'. That in itself is an explanation, albeit arguably inadequate. For the purposes of this study the baseline as compared to the other two interfaces is considered to have minimal or no explanations.

The initial idea was to create a fully functioning high-fidelity prototype in Adobe Experience Design (XD)[8] which users could then physically test on an iPhone XS. Adobe XD is an application for designing interactive wireframes and prototypes which can then be tested on desktop, tablet, or mobile devices. Everything is designed manually without the use of programming. Since the prototype was not connected to Spotify's application programming interface (API) in any way, a finite amount of music available for the user to choose from during testing had to be predetermined. Originally, the prototype had a total of 6 playlists, 3 artists, 3 albums, and 1 radio station. The music chosen to be included was primarily based on what was currently popular on Spotify in Scandinavia and included a variety of genres. Preliminary prototypes were quickly tested by a handful of friends and colleagues during the initial design phases without any instructions to observe their reactions to the application. Initial feedback revealed that people were perplexed and asked why they were being asked to use Spotify. This clearly indicated that it was interchangeable with Spotify's mobile

[8] https://www.adobe.com/no/products/xd.html

application and that users who were already familiar with Spotify did not need any instructions on how to use the prototype.

In March of 2020 the Norwegian government began imposing social distancing guidelines and travel restrictions in addition to closing all universities due to the world-wide corona virus pandemic. As a result, there was no way to complete a lab-study. An online interactive study was considered, however it was deemed to be too difficult. The prototype was specifically made for the iPhone XS and did not scale well to other devices. For instance, an iPhone XS does not have a physical home button or a specific back button. It is therefore intuitive for iPhone users to click the arrow back button on the top left hand side of the screen. Some android phones are designed with a physical button for specifically going back one screen, and the prototype design did not include this functionality. As the design does not take that into account, in addition to other factors, all study participants would have had to have their own iPhone XS in order to test the prototype properly. In addition, while Adobe XD can work in a web-browser with prototypes shared to user testers via link, it would still be difficult to accurately test usability for a mobile phone if the user was testing the system on a computer. Unfortunately, the application was also unable to properly scale to a mobile phone's screen size. A quick test showed that if a user opened the link on their mobile instead of a computer, it opened the prototype in a web-browser. Instead of showing a full window application, a scaled down version of the prototype was displayed in the middle of the web-browser. This significantly reduced the user experience and actually made it harder to use the prototype.

Due to the aforementioned difficulties, 3 non-interactive surveys were created in SurveyXact[9]. SurveyXact is a Danish company managed by Rambøll Management Consulting which takes data protection very seriously and is currently being used by not only the University of Bergen, but also by companies such as the Norwegian Labour and Welfare Administration (NAV) and the Norwegian Police Directorate (Politidirektoratet). These surveys were distributed online through the crowdsourcing website Amazon Mechanical Turk (MTurk)[10] where participants were paid $1 USD for completing a survey. MTurk is a web service that gives tools to researchers who require large numbers of participants and are capable of collecting data for their experiment in an online setting. All the turkers, what Amazon calls their workers, had a HIT[11] Approval Rate of at least 98%, had more than 500 HITs previously approved, and live in the United States. The survey task announcement on MTurk and the SurveyXact survey in its entirety can be viewed in Appendix B. MTurk has been studied extensively for validity. According to both Buhrmester et al. (2011) and Komarov et al. (2013), MTurk participants answer questions just as well as participants answering

[9] https://www.surveyxact.com/

[10] https://www.mturk.com/

[11] HIT stands for Human Intelligence Tasks. More information can be found at https://blog.mturk.com/understanding-hit-states-d0bc9806c0ee

the same questions in laboratory experiments. More recently, (Simon, 1996, p. 20) concluded that while there is some truth in saying that there are issues with online crowdsourcing, *"the evidence bearing upon these concerns does not demonstrate that MTurk samples are fatally flawed"*. Overall, the quality of data harvested from MTurk can be comparable to what might be collected from lab studies, as long as researchers keep in mind its limitations in addition to correctly addressing these issues by utilizing proper research methods appropriate for this domain. Given the circumstances, it was deemed that the MTurk study could be appropriate for this research.

The Norwegian Centre for Research Data AS (NSD) accessed that these anonymous surveys to be in accordance with current data protection legislation. Therefore, a formal ethics approval was not needed as none of the data being processed could directly or indirectly identify individual persons. The number of voluntary participants for Survey 1, Survey 2 and Survey 3 were 71, 70, and 72 people respectively. 65 out of 71 people passed the attention checks of Survey 1, 50 out of 70 for Survey 2, and 63 out of 72 for Survey 3. This made for a total of 178 participants with viable results to analyze. The average survey completion time for Survey 1 was 13.5 minutes, 12.5 minutes for Survey 2, and 12.3 minutes for Survey 3. Each survey was identical; however, each presented a different playlist. As Amazon Mechanical Turk broadens the spectrum of user testers, it was deemed necessary to use music which was not only specifically popular to the Scandinavian market. 2 of the previous 6 playlists were used, and 1 new playlist was created. Even though the original high-fidelity prototype included song, artist, album, and playlist recommendations, the final prototype only included playlists with song recommendations. Users had the opportunity to view a short 30 second video to see what it would be like to interactively use the system. This was made in Adobe XD and hosted on YouTube. It should be noted that the final prototype as shown in the questionnaires only include a very small portion of the entire system.

3.4 Moodplay and Spotify Data

All music data was taken from either Spotify or Moodplay (Andjelkovic, Parra, O'Donovan, & Herrera, 2019). Table 4 describes the moods used in this study in more depth. These represent the types of moods that are evoked when listening to different music genres. Moodplay based their mood categories on the Geneva Emotional Music Scales (GEMS) from Zentner et al. (2008), with a few minor changes. In order to correctly create the precise mathematical percentages displayed in the textual explanations, Denis Parra Santander generously shared the entire Moodplay dataset after the approval of all those who worked on the project. Raimundo Herrera granted this study permission to reuse and edit his colorful bar-graph design from Moodplay's latest web player for the visual explanations. As each playlist contained three different artists, mood percentages for the playlists were calculated by finding the mean values between the three artists. Once the mean

Table 4: Specifications of mood categories from Andjelkovic, Parra, & O'Donovan (2019).

Category	**Sub-category**	**No. of moods**	**Example moods**
Sublimity	Tenderness	24	Delicate, romantic, sweet
	Peacefulness	22	Pastoral, relaxed, soothing
	Wonder	24	Happy, light, springlike
	Nostalgic	9	Dreamy, rustic, yearning
	Transcendence	10	Atmospheric, spiritual, uplifting
Vitality	Power	29	Ambitious, fierce, pulsing, intense
	Joyful activation	32	Animated, fun, playful, exciting
Unease	Tension	32	Nervous, harsh, rowdy, rebellious
	Sadness	18	Austere, bittersweet, gloomy, tragic
	Fear	10	Spooky, nihilistic, ominous
	Lethargy	8	Languid, druggy, hypnotic
	Repulsiveness	10	Greasy, sleazy, trashy, irreverent
Style	Stylistic	19	Graceful, slick, elegant, elaborate
	Cerebral	12	Detached, street-smart, ironic
	Mechanical	7	Crunchy, complex, knotty

values for the given playlist were found, they were compared against each recommended artist to find out how well they matched, in percentages. Jannach et al. (2011) explains that the standard metric for determining item similarity is the cosine similarity measure. Therefore, this method was used in order to find the similarity score between a playlist and the recommended artist.

As Moodplay has a finite music library, all artists in the playlists had to be chosen from there, as opposed to finding them in Spotify. As show in Table 5, the playlists are representative of different genres, moods, and popularity levels. These choices were made in order to see if differences in playlists would affect people's responses. According to Schedl et al. (2018, p. 102), one of the current challenges in MRSs is that a user's *"familiarity with a playlist's genre or theme influenced their judgment of its quality"*. It was of particular interest to see how familiar users considered themselves to be with the artists in the playlists and whether or not playlist familiarity affected their need for explanations. The popular seed artist 'Queen' was chosen for the first playlist 'Bohemian'. This was the playlist used in the Survey 1. In 2018 the movie Bohemian Rhapsody was released, making Queen popular once more towards the end of 2018 and beginning of 2019. This coincided with the beginning of this research. In an article from that January, according to Spotify AB (2019), *"streams of Queen songs on Spotify surged 333 percent... with a two-week run as the No.1 global artist"*. These numbers ensured that the band Queen was popular enough to be considered well-known. At the same time, this proved that Queen had not been in the top charts for a while, which hopefully would ensure that a minority of user testers might still be somewhat unfamiliar with the music. For the second playlist 'Sunrise', a random artist was chosen from Moodplay based on the

Table 5: Playlist information as of May 13th, 2020 from Andjelkovic, Parra, O'Donovan, & Herrera (2019) and Spotify.[12]

Playlist	Genre	Artists	Mood	Monthly Listeners	Top plays
1) Bohemian	Rock	Queen	Style	32,235,160	1,196,251,378
		David Bowie	Style	14,250,978	201,374,337
		R.E.M.	Sublimity	10,218,700	475,885,229
2) Sunrise	Jazz	Norah Jones	Sublimity	5,361,176	214,593,841
		Ingrid Michaelson	Sublimity	2,849,272	98,359,641
		Inger Marie Gundersen	Sublimity	104,074	2,867,004
3) Get Up	House	Bushwacka!	Vitality	6,342	40,673
		Layo & Bushwacka!	Vitality	117,027	2,532,221
		John Acquaviva	Vitality	8,217	163,011

criteria that it was within a different music genre, mood category, and had significantly less Spotify followers. This lesser known seed artist for the second playlist ended up being 'Norah Jones' and this playlist was what people taking Survey 2 saw. For the third playlist 'Get Up', the same criteria applied. The most unfamiliar artist from all the playlists was 'Bushwacka!' with just over 6,000 followers. They were chosen as the seed artist for this last playlist which was evaluated in Survey 3.

Moodplay has a controllable slider to change the type of recommendations users receive. The standard is set to the LDA-b algorithm which is primarily based off of mood data. At the other end of the slider is t-SNE-a which factors in other aspects such as audio or genre data. It is slightly less dependent on moods for recommendation. To create a single playlist once a seed artist was chosen, the second artist chosen for that playlist was based on Moodplay's recommendations using the LDA-b algorithm. The third and final artist in the playlist was chosen based on recommendations from the t-SNE-a algorithm. Playlists were named after the first song displayed. Other criteria when choosing artists was that all their songs chosen to be in the playlists, were songs in English. When the artists were chosen, playlists were created in Spotify to create the playlist recommendations. For each playlist, 10 of the songs which were recommended by Spotify were chosen at random to be in the recommended songs list. Each recommendation list also contained one song from each of the artists which were in the playlist. A new Spotify account was created in order not to receive recommended songs based on previous music listening history. Recommendations used were also checked against the popularity of the playlist artists to make sure that they were within the same popularity range. Though it was not necessary for them to be within the same genre or mood, most recommendations were, with some exceptions.

[12] https://developer.spotify.com/documentation/web-api/reference/artists/get-artist/

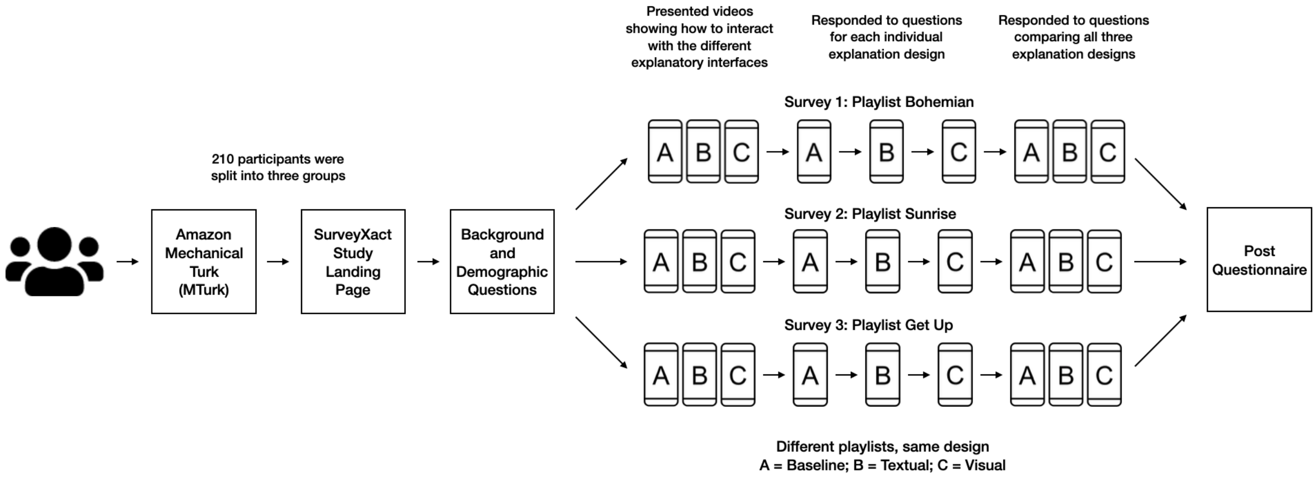

Presented videos showing how to interact with the different explanatory interfaces
Responded to questions for each individual explanation design
Responded to questions comparing all three explanation designs
210 participants were split into three groups
Amazon Mechanical Turk (MTurk)
SurveyXact Study Landing Page
Background and Demographic Questions
Survey 1: Playlist Bohemian
Survey 2: Playlist Sunrise
Survey 3: Playlist Get Up
A
B
C
Post Questionnaire
Different playlists, same design
A = Baseline; B = Textual; C = Visual

Figure 34: MTurk study user task chart.

3.5 Approach to Main Study

In order to answer the research questions, a mixed methods approach was taken to collect both quantitative and qualitative data. Quantitative data can be validated statistically but can be quite generalized, while qualitative data provides more specific information about user needs. While qualitative data better captures the thoughts and feelings of user testers, it may be more difficult to interpret or be too vague. It is important to collect both types of data to avoid any potential limitations or bias through evaluating only one type. All data was collected through a series of surveys, as shown in Figure 34. Quantitative data was gathered through asking participants questions both relating to their demographics and personal experiences, and through questions which required them to scale or rank specific design aspects. Statistical tests were completed on the latter. Results of the qualitative data emerged after completing a sentiment analysis of the comments users wrote when responding to open-ended questions.

A sequential within-subjects manipulation approach was taken when designing the main study. This means there was a kind of combination of within-subjects and between-subjects. This was done as there were two independent variables that were being tested, explanation design and playlist familiarity. Ricci et al. (2015, p. 326) explains that *"in a between-subjects experiment, participants are randomly assigned to one of the experimental conditions"*. In regard to this study, the experimental condition that stayed the same was the interface design, and the one which was changed was the playlist. The reason for it being sequential, displaying the three different interface designs in the same order for each participant, was due to limitations with the SurveyXact software. All questions which benefited from randomization were randomized when possible. Unfortunately, there was not a possibility to randomize page order or main questions, only sub-questions. While all 213 subjects were exposed to the same questions and same interface designs, they were randomly assigned one of the three different musical playlists mentioned earlier.

A statistical analysis was completed with all questions about the three different interface designs. SurveyXact data was exported into Excel to calculate the means, medians, population standard deviations, and population standard errors. A non-paired two-sided Wilcoxon rank sum test with continuity correction was then performed in RStudio to find significant p-values.[13] A Bonferroni correction test was completed on all the p-values in order to counteract the problem of having false positives when conducting multiple comparisons.[14] The reason for choosing a rank sum test over a student t-test is that it is non-parametric which is better when evaluating user ratings such as Likert scale data. All other questions were analyzed using mean scores and are visually represented in graphs created in Excel.

A content analysis was completed on the ambiguous comments participants made when responding to open-ended questions. Answers were analyzed by iteratively grouping comments to identify themes based on the research questions. According to Lazar et al. (2017, p. 301), this is done to frame the scope of the context, a process which "searches for theoretical interpretations that may generate new knowledge". Categories were primarily based on pre-conceived concepts from prior research. This is indicative of a priori coding method, which according to Lazar et al. (2017, p. 301) "involves the use of an established theory or hypothesis to guide the selection of coding categories". At the same time, as this research in the domain of mobile phones is brand new, an emergent coding method of not basing categories on past theories was still kept in mind. So even though the analysis was guided by previous research models, this slight mixture allowed for the ability to also establish new concepts based on new information. All themes and comments presented are representative of multiple respondents, but should still be considered purely subjective and not necessarily representative of society as a whole.

3.5.1 Evaluation Metrics

According to Sutton (2018), as certain people may be sensitive to certain questions, it is better to phrase questions as ambiguously as possible. This is why the background questions for example, used the phrase: *"which category best describes you?"*. The music and technology preferences section and music recommendation preferences section were adapted from questions asked by Millecamp et al. (2018). This section specifically includes two questions about tech-saviness, one phrased positively and one phrased negatively. Similar questions phrased in different manners

[13] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/wilcox.test

[14] https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-9863-7_1213

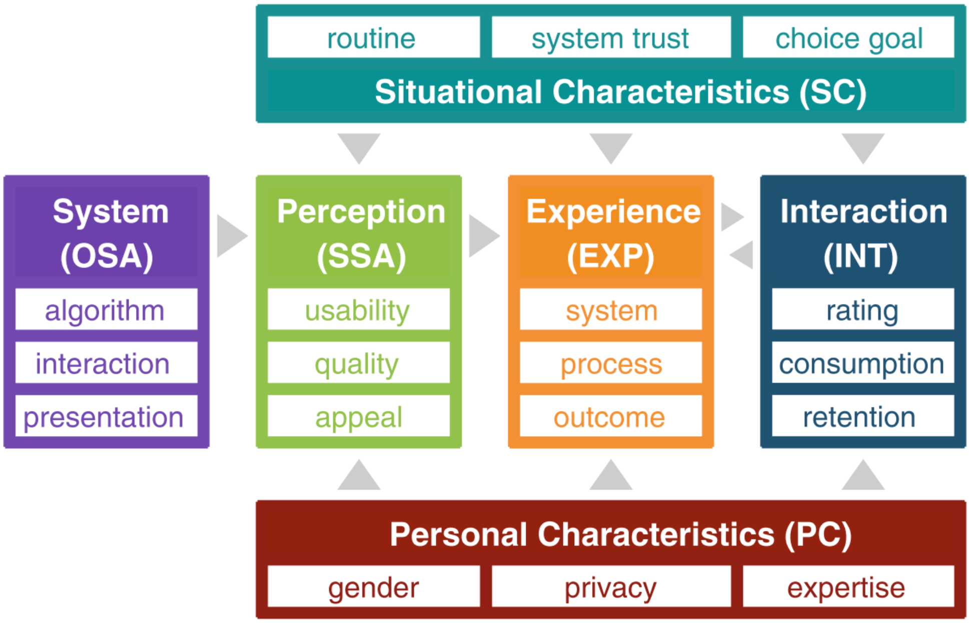


Figure 35: User centric evaluation framework chart from Knijnenburg & Willemsen (2015).

creates multi-way comparisons which can remove a potential bias based on the effects of question polarity. This same approach was used for the Likert scale questions asked in the interface comparison sections. Likert scale questions asked participants to agree or disagree with a statement on a scale of 1 to 5, where 1 equated to 'strongly disagree' and 5 equated to 'strongly agree'. In order to see whether or not participants were paying attention, two 'attention check' questions were hidden amongst other statements asking about the explanatory interface designs. Participants had to choose correct answers for both of the statements *"I will choose the neutral answer three for this question"* and *"The answer to this question is disagree which is number two"*, in order to be included in the statistical results. The end of the survey asked users about their experience with the survey and also included some additional open-ended questions. A copy of Survey 1 in its entirety can be found in Appendix B.

According to Tintarev & Masthoff (2011), the first guideline when designing recommendation explanations is to establish what needs to be explained and why. After defining the explanatory goals, one can choose the appropriate evaluation metrics. A discussion of these follow in the next paragraph. The second guideline is to remember that in order to evaluate a recommendation explanation, the entire system needs to be taken into account. In the case of this thesis, it may be possible to examine aspects of this guideline such as recommendation accuracy and acceptance/satisfaction with the system, but not aspects such as learning rate (the rate at which the RS determines user preferences) or coverage (how wide or narrow the spectrum of items recommended are). Recommendation accuracy is not explicitly evaluated, though a few participants commented about this. In terms of acceptance, Tintarev & Masthoff (2011, p. 494) notes that *"If users are satisfied with a system with an explanation component, it remains unclear whether this is due to: satisfaction with the explanation component, satisfaction with recommendations, or general design and visual appeal"*. When participants were asked about their overall satisfaction with a particular recommendation explanation, the hope is that they responded in terms of its design since they only saw an image of the explanation and did not really have an opportunity to properly evaluate the system as a whole with these explanations. This idea should still hold true even if users were already familiar with Spotify, though Tintarev's note on acceptance is still important to keep in mind when reading through the evaluation results. The third guideline is to look at how explanatory goals are affected by the way the explanation is presented to the user and by the way a user interacts with the explanation. The explanations designed for the prototype created in this thesis gave users a structured overview of the recommendations in order for them to make side-by-side music comparisons. In terms of interaction, preference controllability was not an aspect which was investigated. The final guideline says to keep in mind the relationship between the explanation style and the recommendation algorithm. The recommendations and explanations were independent of one another, as Spotify was used for the former and Moodplay for the latter. As the explanation design was the most important part of this research, this criterion was not explicitly measured, though some participants also commented about this.

All interfaces were designed in respect to explainability, transparency, and satisfaction. Survey questions were primarily based off of the well-known user-centric evaluation frameworks for recommender systems, ResQue and the Knijnenburg Scale. Ricci et al. (2015) highly recommends using these existing methods and even encourages reusing their constructs word for word. Pu et al. (2011, p. 157) explains that ResQue, which stands for the Recommender systems' Quality of user experience, is "aimed at measuring the qualities of the recommended items, the system's usability, usefulness, interface and interaction qualities, users' satisfaction with the systems, and the influence of these qualities on users' behavioral intentions, including their intention to purchase the products recommended to them and return to the system". The constructs from Pu's work which were taken into consideration when creating the survey are interface adequacy, information sufficiency and explicability, perceived usefulness, perceived ease of use, transparency, overall satisfaction, trust, and behavioral intention to use the system. The ResQue framework even includes a construct for explanations, though it is only one question. The Knijnenburg Framework as shown in Figure 35 explains a framework for a user-centric evaluation process for recommender systems. Each category points to another, explaining which other category it may influence or have a direct impact upon as well as how different evaluation metrics overlap and intertwine with one another. Based on these two frameworks, the concepts used to develop the main MTurk study were the presentation of the systems recommendation explanations (OSA), user perception of recommendation explanations (SSA), user experience with recommendation explanations (EXP), user trust in the recommender system (SC), and user trust in technology (PC). Survey questions were formulated based off of the following constructs: Information sufficiency and recommendation presentation (OSA), Interface adequacy and perceived explanation quality (SSA), Perceived Usefulness and use intention and overall satisfaction (EXP), Transparency (SC), and General trust in technology (PC).

4 RESULTS

This Chapter gives an overview of significant findings revealed through user testing. Results from the pre-study questionnaires are discussed in Section 4.1 and Section 4.2. Background information about the people who partook in the main study is disclosed in section 4.3, followed by a statistical analysis of user responses in Section 4.4. Finally, Section 4.5 analyzes the particpant comments to open-ended questions.

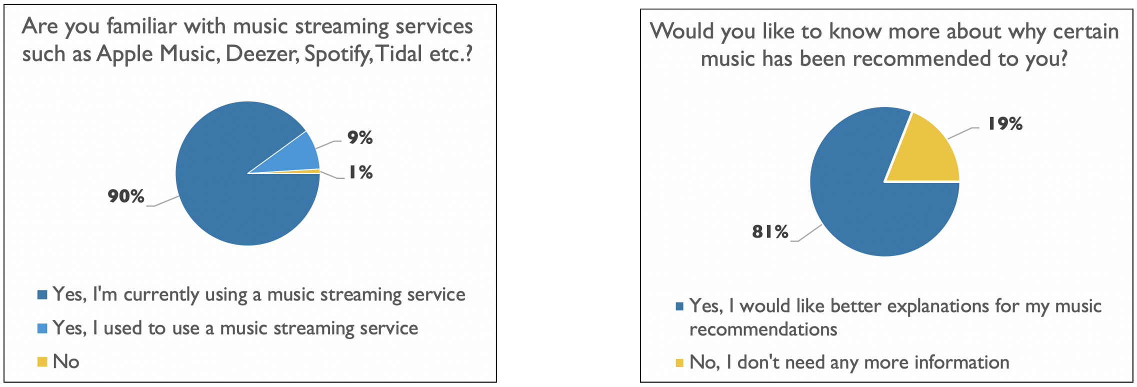
4.1 First Pre-Study Questionnaire

The first pre-study questionnaire presented 100 participants with a brief statement about this thesis and a few examples of textual music recommendation explanations. Results from the first two questions, shown in Figure 36, help to reinforce findings from previous research by verifying high familiarity/usage of music streaming services, in addition to the need for more music recommendation explanations. This establishes grounds for asking RQ1 which also evaluates the value of explanations, but more specifically in a mobile setting. For the third question, participants were asked to choose which of the 14 musical data attributes displayed in Figure 6 they would like to see in a music recommendation explanation. The results, shown in Figure 37, are categorized to shown comparisons between all participants and different participant sub-groups. The y axis represents the number of people who voted for the attribute named on the x axis. In regard to the most popular data types, 71% of participants chose genre information, 60% chose moods, and 57% chose similarity score. TagFlip (Kamalzadeh et al., 2016) for example, also found that the top two ways to classify tags were genre and moods. These results support previous research which stress the importance of exploring the exploitation of affective-information for explanations more in depth, and was used as the basis for asking RQ4. Therefore, the final prototype utilized affective-information, in addition to providing mood-based similarity scores for the recommendation explanations.

4.2 Second Pre-Study Questionnaire

The next task in this research study was to create a questionnaire which would help to determine the design of these recommendation explanations in order to answer RQ2. At this point, it was already decided that the bar graph design from Moodplay's newest web interface was to be used as the visual explanation design. This meant that only one additional questionnaire about textual explanation design needed to be carried out. A little over 100 people were asked to rank the 8

[14] https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/



(a) (b)

Figure 36: First pre-study questionnaire background results (n=100).

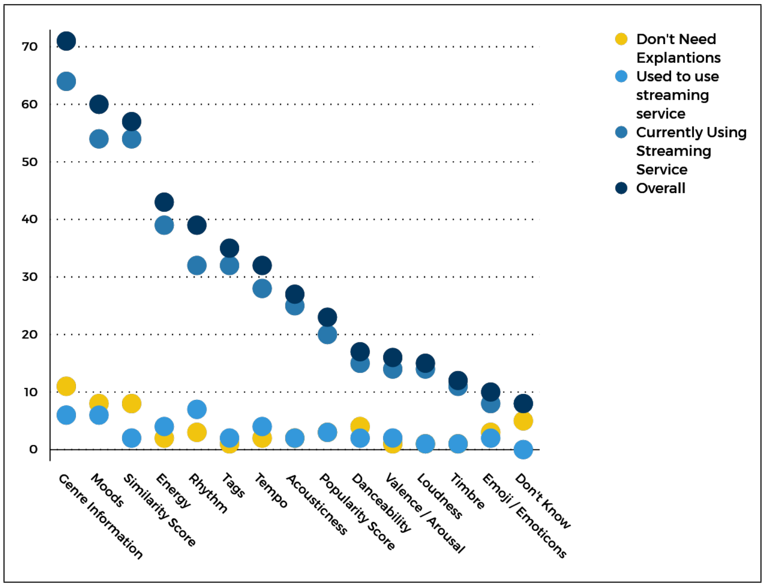


Figure 37: Overall results from the first pre-study questionnaire (n=100).

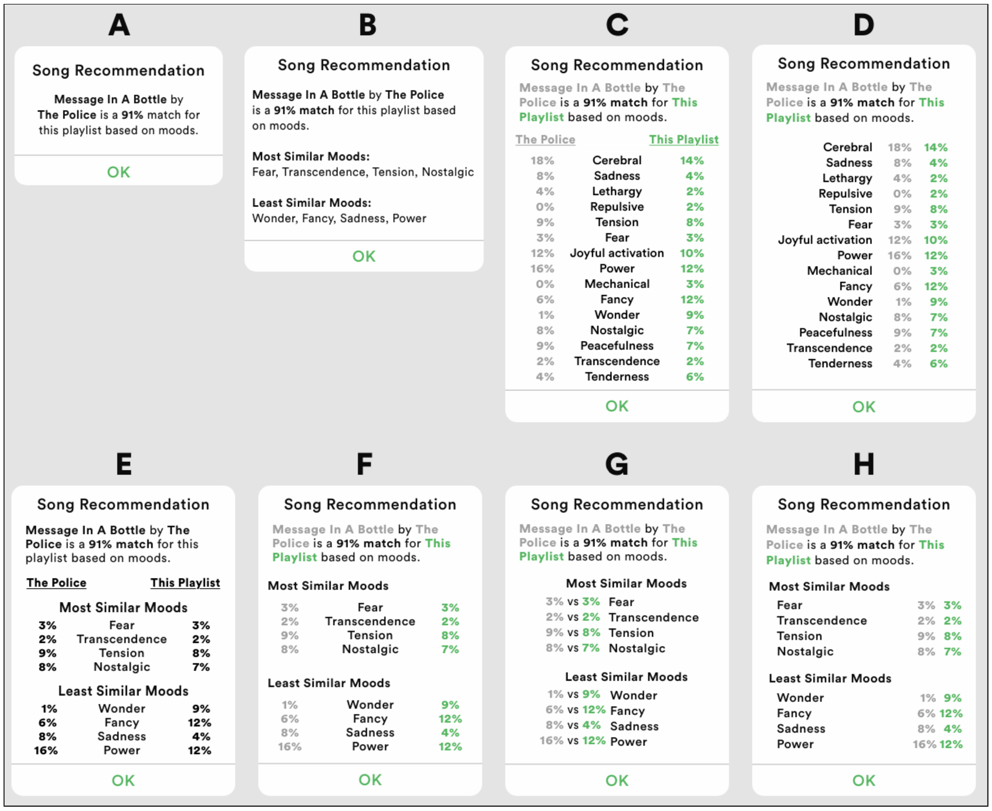
different textual explanation designs see in Figure 38 from their favorite (1) to least favorite (8). Out of the 91 participants who fully completed the questionnaire, only 47 were found to be valid, most likely due to issues with question formatting and validations in SurveyXact. Some only ranked the designs with either a 1 or an 8 while others tried to rank them but repeated one of the numbers. As SurveyXact does not give an error message until after one clicks the 'next' button to submit the form, users did not know in real time that they had made a mistake. Many simply left the questionnaire without correcting their mistakes, but they were still counted as having completed the questionnaire as they did fill out every question. Caution should be taken when interpreting these results, but they can still can provide a rough estimation of what people think of different types of textual explanations. These 47 participants do not accurately represent the general population.

Table 6: Descriptions of musical data attributes adapted from Spotify.[15]

Data	**Definition**
Acousticness	How acoustic or electric a song is, as in whether or not the song has a lot of acoustic analog elements or electric digital elements.
Danceability	How danceable the song is, as in whether or not the song is fitting to dance to.
Energy	The energy level of a song, as in whether the song is energetic, fast, loud, noisy or has a low energy being characterized as slow, sad, sleepy, suspenseful, soft etc.
Emojis/Emoticons	Describing artists based on their top emotional icons.
Genres	Musical categories such as rock, pop, blues etc.
Loudness	The musical dynamic, how loud a song is in terms of decibels (dB).
Moods	The type of emotion a certain song is categorized as being, how a song makes one feel while listening to it.
Rhythm	The musical feel.
Similarity Score	The percentage of how similar two items, such as an artist or album or playlist, are to each other.
Tags	Musical keywords or terms or annotations.
Tempo	Beats per minute (BPM), if a song is fast or slow.
Timbre	Tone color or tone quality.
Valence/Arousal	How musically positive or negative a certain song is, which could be interpreted as a sort of combination of energy (arousal) and moods (valence).

Once again, participants were asked about their familiarity with music streaming services and the need for more music recommendation explanations, in addition to a question about mobile phone music listening habits. Figure 39 shows that almost all participants use their phone to listen to music at least on a weekly basis and that 2/3 would prefer more explanations. The stacked horizontal bar graph in Figure 40 depicts which interface designs users preferred. Percentages shown on the left and right sides of the graph represent the number of negative and positive rankings, respectively, for that particular interface. The average ranking for these interfaces, where a score of 1 is best and 8 is worst, are: B (2.95), A (3.50), C (4.35), F (4.55), H (4.95), D (5.05), G (5.20), E (5.50). Overall, users ranked the designs as follows: B, A, C, F, H, D, G, E, suggesting that even though users would like explanations, they may prefer ones with minimal information.

A provided the least amount of information, not fully explaining how the similarity score was calculated in terms of moods. Design B provided some of that extra information, but still did not include a specific breakdown of these moods in percentages. Based purely on the idea that users prefer minimal information, one would then assume that designs E, F, G, and H would be the next



A

Song Recommendation

Message In A Bottle by **The Police** is a **91%** match for this playlist based on moods.

OK

B

Song Recommendation

Message In A Bottle by **The Police** is a **91% match** for this playlist based on moods.

Most Similar Moods:
Fear, Transcendence, Tension, Nostalgic

Least Similar Moods:
Wonder, Fancy, Sadness, Power

OK

C

Song Recommendation

Message In A Bottle by **The Police** is a **91% match** for **This Playlist** based on moods.

The Police		This Playlist
18%	Cerebral	14%
8%	Sadness	4%
4%	Lethargy	2%
0%	Repulsive	2%
9%	Tension	8%
3%	Fear	3%
12%	Joyful activation	10%
16%	Power	12%
0%	Mechanical	3%
6%	Fancy	12%
1%	Wonder	9%
8%	Nostalgic	7%
9%	Peacefulness	7%
2%	Transcendence	2%
4%	Tenderness	6%

OK

D

Song Recommendation

Message In A Bottle by **The Police** is a **91% match** for **This Playlist** based on moods.

Cerebral	18%	14%
Sadness	8%	4%
Lethargy	4%	2%
Repulsive	0%	2%
Tension	9%	8%
Fear	3%	3%
Joyful activation	12%	10%
Power	16%	12%
Mechanical	0%	3%
Fancy	6%	12%
Wonder	1%	9%
Nostalgic	8%	7%
Peacefulness	9%	7%
Transcendence	2%	2%
Tenderness	4%	6%

OK

E

Song Recommendation

Message In A Bottle by **The Police** is a **91% match** for this playlist based on moods.

The Police		This Playlist
	Most Similar Moods	
3%	Fear	3%
2%	Transcendence	2%
9%	Tension	8%
8%	Nostalgic	7%
	Least Similar Moods	
1%	Wonder	9%
6%	Fancy	12%
8%	Sadness	4%
16%	Power	12%

OK

F

Song Recommendation

Message In A Bottle by **The Police** is a **91% match** for **This Playlist** based on moods.

Most Similar Moods

3%	Fear	3%
2%	Transcendence	2%
9%	Tension	8%
8%	Nostalgic	7%

Least Similar Moods

1%	Wonder	9%
6%	Fancy	12%
8%	Sadness	4%
16%	Power	12%

OK

G

Song Recommendation

Message In A Bottle by **The Police** is a **91% match** for **This Playlist** based on moods.

Most Similar Moods

3% vs 3% Fear
2% vs 2% Transcendence
9% vs 8% Tension
8% vs 7% Nostalgic

Least Similar Moods

1% vs 9% Wonder
6% vs 12% Fancy
8% vs 4% Sadness
16% vs 12% Power

OK

H

Song Recommendation

Message In A Bottle by **The Police** is a **91% match** for **This Playlist** based on moods.

Most Similar Moods

Fear	3%	3%
Transcendence	2%	2%
Tension	9%	8%
Nostalgic	8%	7%

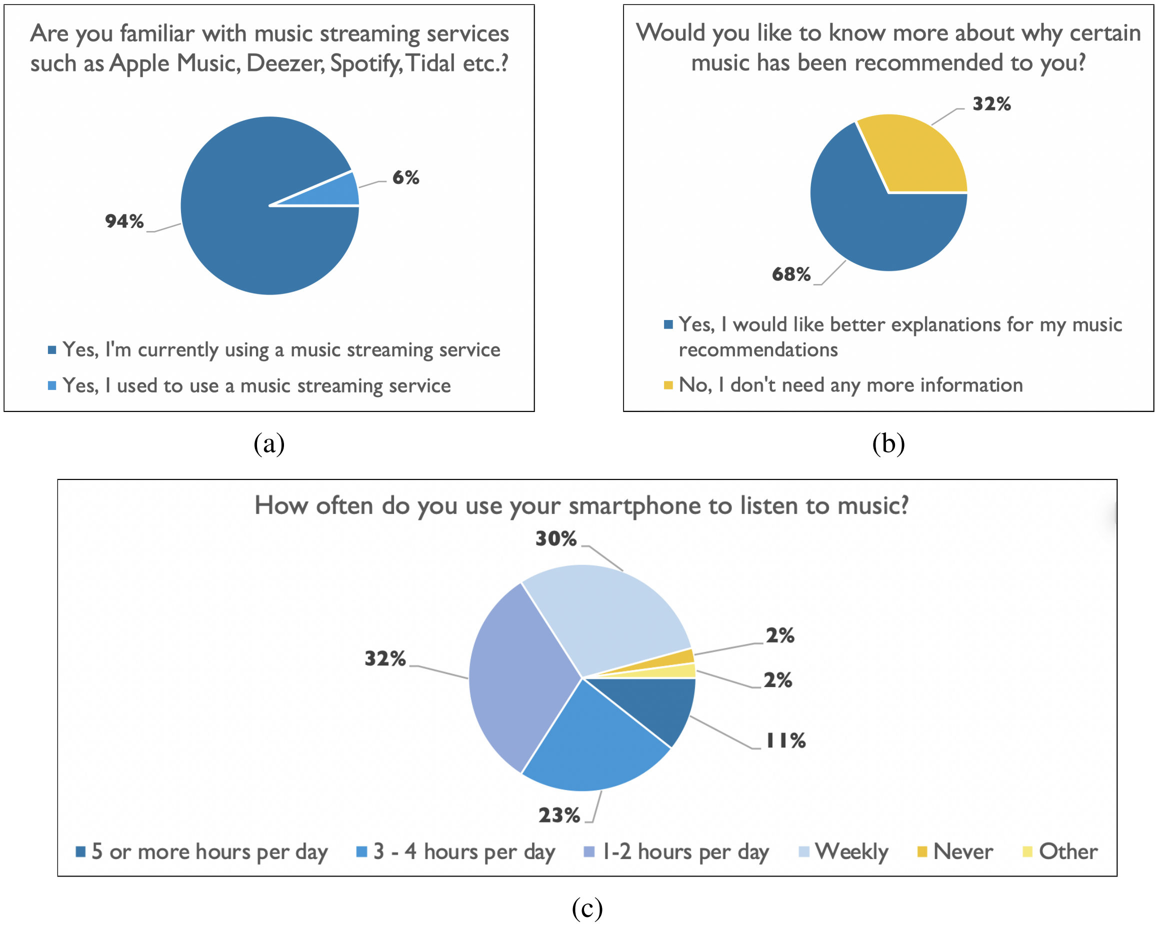
Least Similar Moods

Wonder	1%	9%
Fancy	6%	12%
Sadness	8%	4%
Power	16%	12%

OK

Figure 38: Textual explanation designs from the second pre-study questionnaire.

popular interfaces, as apposed to designs C and D. However, the third most popular explanation design was C, which contained the most amount of information. More people also gave design A a rating of 8 than they did to design C. This means that more users strongly disliked design A than strongly disliked design C, even though 15% more participants gave design C lower rankings overall. Design F was the fourth most popular design, and both F and C followed a similar layout with the percentages being on either side of the mood labels, rather than being side by side. It's possible that graphic design played a role in how participants determined which explanation design they preferred. Out of the 21 people who gave explicit feedback, 6 people mentioned that they preferred colorful text, and 11 mentioned that anything other than A and B contained too many details. With E in last place, it can be assumed when it comes to the more explanatory interfaces, that users prefer colorful designs as apposed to only plain black text. The results of this study impacted the decision to use design C as the textual explanation design for the prototype, in addition to impacting the layout for the visual explanation. This was due to its third place standing and since it matched Moodplay's visual explanation design in terms of amount of information shown.



(a)

(b)

(c)

Figure 39: Second pre-study questionnaire background results (n=47).

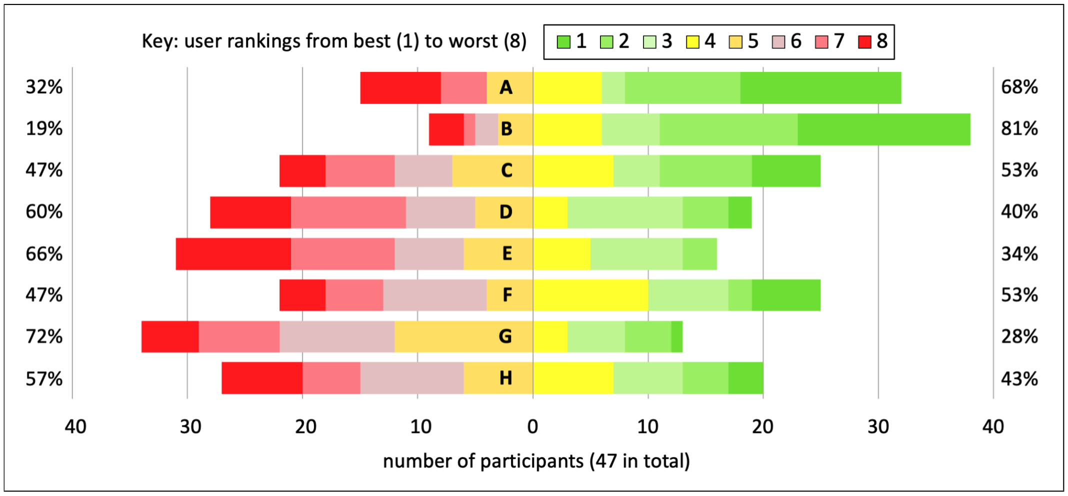


Figure 40: Overall results from the second pre-study questionnaire (n=47).

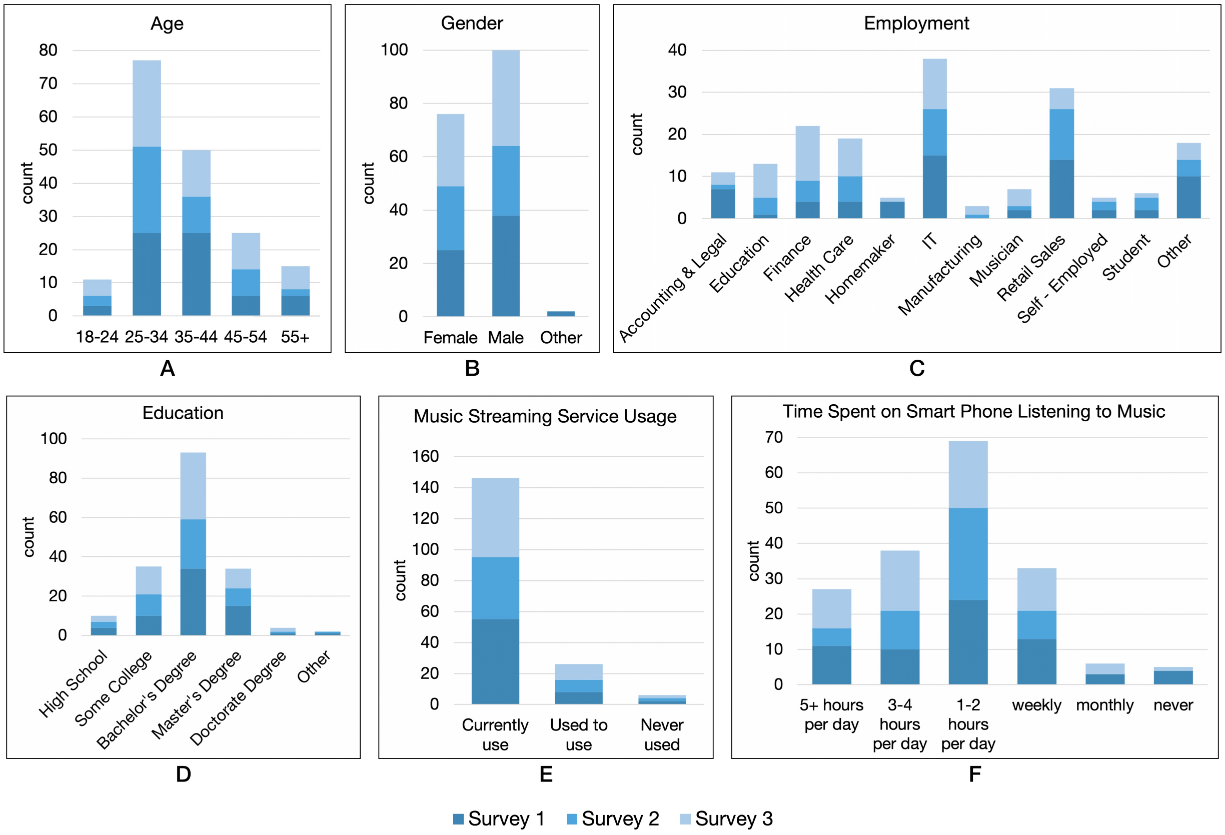


Figure 41: Demographics and music listening habits of MTurk participants (n=178).

4.3 MTurk Participants

Demographic information and music listening habits of all 178 participants can be found in Figure 41. Most participants were between the ages of 25 and 34. 56% were male. Not surprisingly, more IT workers responded to the MTurk announcement than any other employment type. Just over half of all participants hold bachelor's degrees. 146 participants currently use a music streaming service. Surveys 1 and 2 had 8 participants each that used to use a music streaming service, and there were 10 in Survey 3. 6 participants, 2 from each Survey, stated that they had never used a music streaming service before. 39% of participants stated that they spend 1 to 2 hours per day listening to music on their phones, 21% listen for 3 to 4 hours, 19% listen weekly, and 15% listen for more than 5 hours per day.

An overview of confidence and comfort levels with technology for all 178 participants can be found in Table 7 and Figure 42. An overwhelming majority of participants responded positively to statement A. However when responding to statement B, the reverse coding of this question showed that user responses did not always match with those from the previous statement. Table 7 shows that there were no statistical differences between the responses of these two questions for

Table 7: Digital technology confidence levels comparison. (Likert scale 1-5, higher values indicate more agreement with the statement; Significance values are based on comparisons between the two statements where ** = significant at p<0.01).

Statement		**Survey 1 (n=65)**	**Survey 2 (n=50)**	**Survey 3 (n=63)**	**All (n=178)**
A: I am confident about my ability to use digital technologies	Mean	4.38	4.26	4.08**	4.24**
	SD	0.60	0.77	0.67	0.69
	SE	0.07	0.11	0.08	0.05
B: The thought of using unfamiliar digital technology is comfortable[a]	Mean	3.95	3.72	3.46**	3.71**
	SD	1.26	1.28	1.22	1.27
	SE	0.16	0.18	0.15	0.10

[a] This was reverse coded so that the high end of the scale, scores of 4 and 5, indicate agreement with the positive version of the statement.

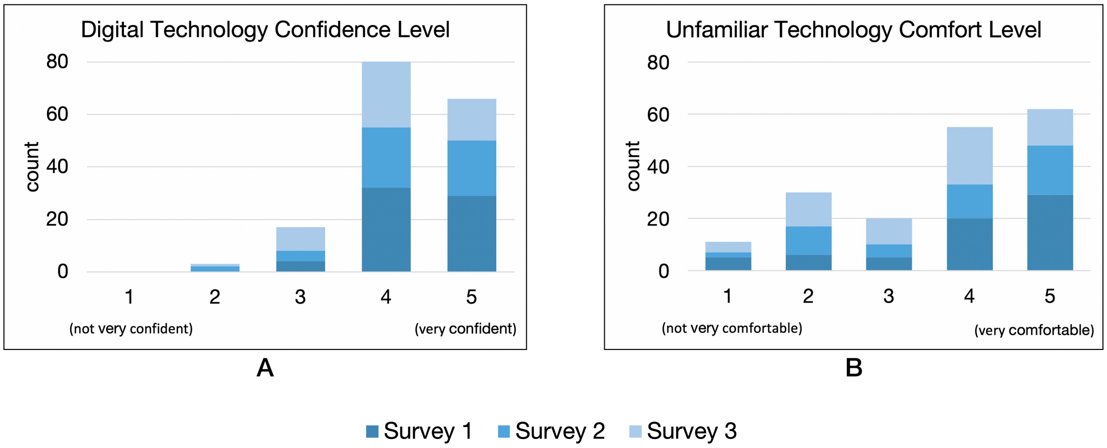


Figure 42: Digital technology preferences of MTurk participants. (n=178; Supporting graphs for Table 7).

Survey 1 and Survey 2, however there is a significance of p<0.01 for both Survey 3 and overall responses. The standard deviation (SD) is almost double for the negatively phrased statement B than for A. There is insufficient information as to why there was a wider variation in responses. While difficult to accurately ascertain to what extend participants are tech-savvy, Figure 43 shows that the majority currently use a music streaming service. In addition, the majority also perceive themselves to be familiar with both Spotify and recommender systems in general.

Pertinent to RQ1, to get a better idea of how music taste may affect user responses to the musicians listed in the different playlists, participants were asked to choose their top 3 favorite genres. Figure 43 reveals that around 72% of participants ended up choosing more than 3. For the top three genres, in total, 85 participants chose pop as one of their favorite music genres, 76 chose rock, and 74 chose hip hop / rap. Of the participants from Survey 1 who evaluated the rock playlist Bohemian, 29 out of 65 people included rock in their list of favorite genres. Only 10 of the

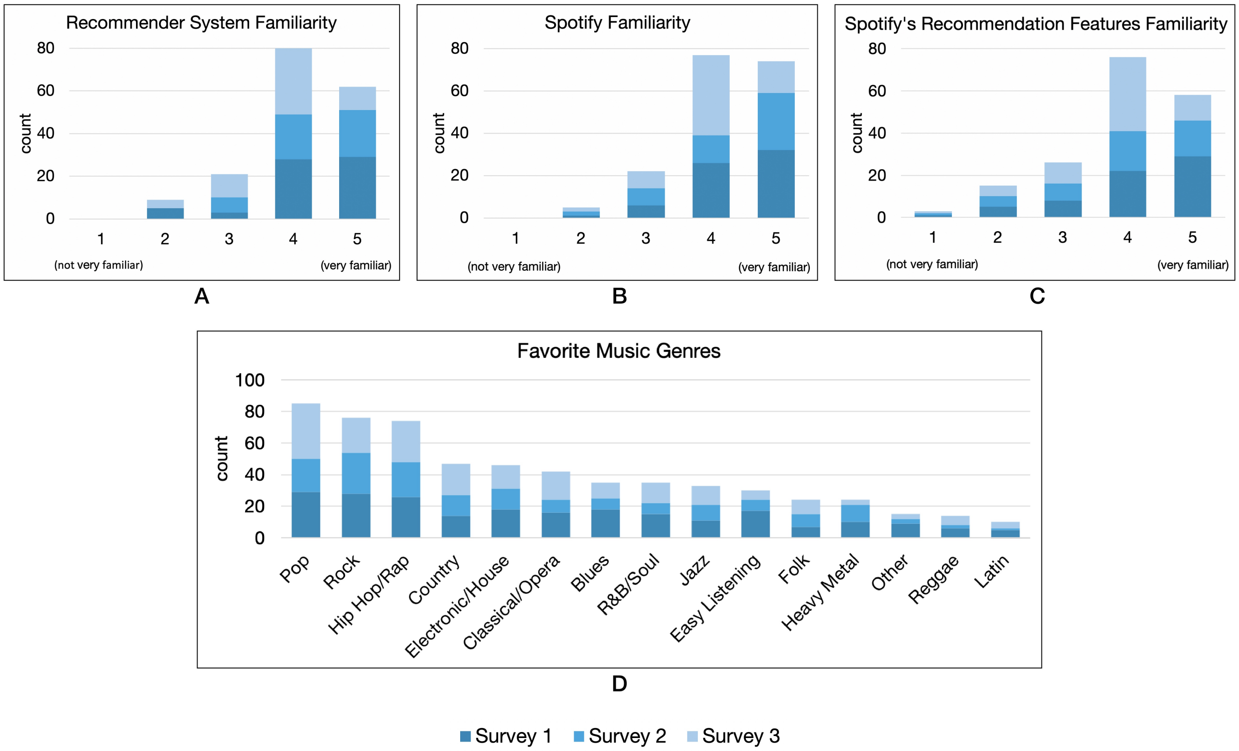


Figure 43: Music and recommender system preferences of MTurk participants (n=178).

55 people who evaluated the jazz playlist Sunrise in Survey 2 chose jazz as one of their favorite genres. In total out of all 178 participants, 33 favored the jazz genre. For Survey 3, which included the house genre based playlist Get Up, only 15 of the 63 participants listed electronic / house as a favorite genre. It could be inferred that 45% of participants may have already been familiar with the artists and songs in Bohemian, 18% were already familiar with Sunrise, and 24% were already familiar with Get Up. Looking ahead, Table 11 reveals that participants claimed to be most familiar with with the artists and songs in the playlist Bohemian, and a bit more familiar with the Sunrise playlist than of those in the Get Up playlist.

If a participant responded that they either currently use or used to use a music streaming service, they proceeded to answer questions about music recommendations and explanations. The overall results from the 172 participants who completed this section are shown in Figure 44. These findings help to answer RQ1 by looking at how much recommendations and their explanations are valuable to users and what type of explanation data users value most. The 26 participants who used to use music streaming services had slightly lower averages overall, higher variance (SD) and more than double the standard error in their scores than the 146 people who currently listen to music streaming services. Results suggest that users are quite aware of their music recommendations and interact with them often. Users would like more explanations for their recommendations, in

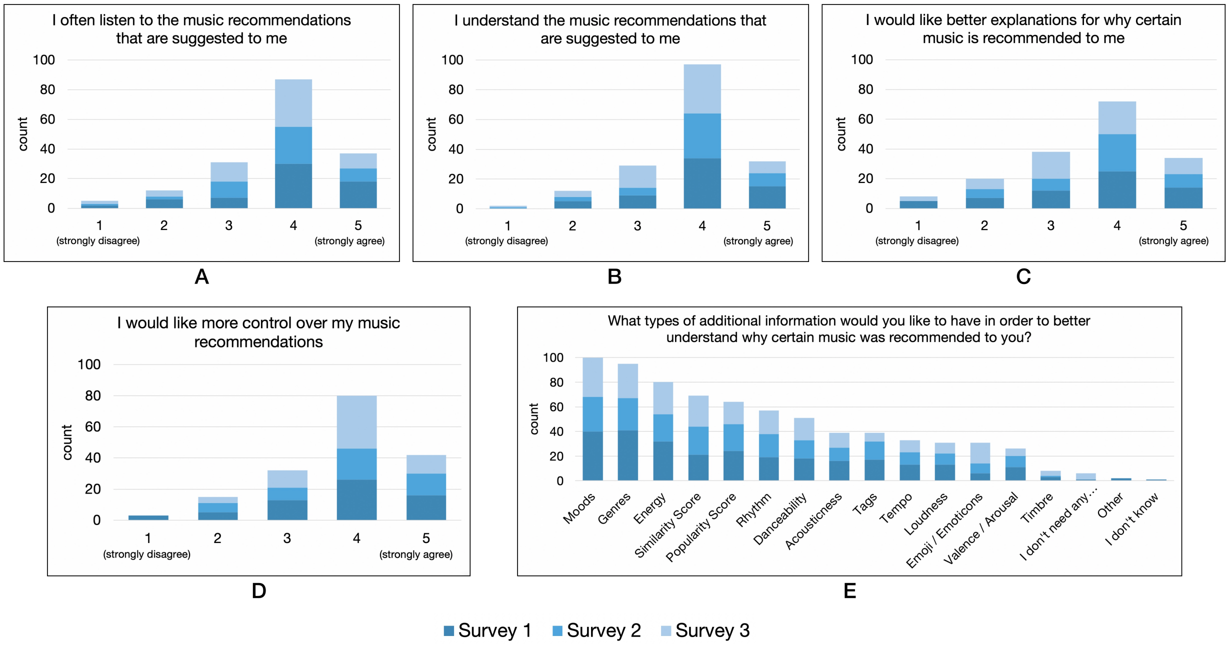


Figure 44: Music recommendation and explanation preferences of MTurk participants (n=172).

line with previous results, albeit teetering on the lower end of the positive side of the scale. More users claimed they already understand their recommendations than those who would like better explanations. It is difficult to posture whether or not this data alone positively contradicts the notion that if a user already understands their recommendations, there is not a strong need for explanations, but it does lean in that direction. Interestingly, results also show that people seem to want more control over their recommendations than want better recommendation explanations. In terms of what type of information users would like to see in music recommendation explanations, 58% of participants chose moods, 55% chose genres, 47% chose energy, and 40% chose similarity score. The rankings of moods, genres, and similarity score are comparable to what the previous two pre-studies discovered.

4.4 MTurk Interface Comparisons

For each of the three explanation designs shown in Figure 33, participants were asked to answer the 13 question statements shown in Table 8 in a random order. The numbers used in Tables 9, 10 and 11 refer to these same question statements. Numbers in bold represent the best mean score for that statement, but are not necessarily statistically significant. For Tables 9 and 10, the baseline does not display any stars, but can be inferred from the star counts in the other two columns.

S1 tries to find a relationship between music familiarity and the need for explanations which speaks to RQ1, *"To what extent are explanations in mobile music recommender applications* *valued by users, and how does music familiarity affect that?"*. S2 and S3 tackle the topic of comprehensibility by evaluating what Pu et al. (2011) categorizes as interface adequacy. They fall under Knijnenburg's evaluation category subjective system aspects which deal with user perception of recommendation explanations (Knijnenburg & Willemsen, 2015). S4 and S5 deal the issue of information overload, which falls under the evaluation category of information sufficiency. S2 through S5 work toward answering RQ2, *"How do users evaluate the different design elements of music recommendation explanations in a mobile setting?"*. S6 and S7 consider the value of the recommendation explanations, as in whether or not the the information presented is useful and whether or not it helps the user understand the recommendation. Note that while the table uses the general terms 'song' and 'artist' in S6, each Survey asked participants about a specific recommendation such as 'Message In A Bottle' by 'The Police'. These statements also examine perceived usefulness and an aspect of system transparency which may have implications for RQ4 *"To what extent do affect-based explanations influence a user's perception of the system and its music recommendations?"*. S8 and S9 help to answer RQ4. S8 looks at system trust which Knijnenburg considers to be a situational characteristic. S9 appraises Knijnenburg's process experience variable cognitive effort, which is a factor in user decision making. S10 and S11 measure user evaluations of how adequate an explanation is in terms of design which answer RQ2. S12 explores users' intentions to use the system, which Knijnenburg considers to be a way of evaluating interaction, even though the study itself was not interactive. It is important to know whether or not users are willing to take advantage of recommendation explanations. This supports RQ1 as there is, for example, little value in a system someone likes but would never use. S13 looks at overall satisfaction which aids in answering RQ4.

4.4.1 The value of explanations - statistics (RQ1)

Research question one looks to uncover the extent to which users would like explanations for music recommendations to be included in music applications on their mobile devices and whether or not the value of explanations is changed by a user's music familiarity. Table 10 shows the correlation between music familiarity and how participants perceived the explanation designs. As previously stated, artist popularity was taken into consideration when building playlists and the results of S1 in this table reflect these considerations. Generally speaking, results of S1 appear to show that participants were most familiar with music in the Bohemian playlist from Survey 1, moderately familiar with the music in the Sunrise playlist from Survey 2, and least familiar with the music in the Get Up playlist from Survey 3. Statistically speaking, there was not a big difference between Sunrise and Get Up, except some significance ($p<0.05$) for the visual explanation comparison

Table 8: Statements responded to by users during the interface comparison section.

Number	Statement
S1	I am familiar with these artists / songs.
S2	The design layout of the explanation is clear and easy to understand.
S3[a]	The explanation is unnecessarily complex.
S4	The explanation contains the right amount of information.
S5[a]	The explanation is too detailed.
S6	I understand why ‘song’ by ‘artist’ is recommended.
S7[a]	The explanation does not provide useful information.
S8	The explanation increased my trust in the system.
S9[a]	It took me a long time to understand the explanation.
S10	The explanation is attractive and pleasing to look at.
S11[a]	I do not like the design of the explanation.
S12	I would like to use a system which uses this type of explanation.
S13	Overall, I am satisfied with the recommendation explanation.

a/ These were reverse coded so that the high end of the scale, scores of 4 and 5, indicate agreement with the positive version of the statement.

(Sunrise M = 3.38, SE = 0.17; Get Up M = 2.75, SE = 0.16). There were however statistical significances ($p<0.001$) for the baseline comparisons of Bohemian (M = 4.20, SE = 0.12) and Sunrise (M = 3.16, SE = 0.16) and between Bohemian and Get Up (M = 2.67, SE = 0.16), the textual comparisons of Bohemian (M = 4.02, SE = 0.14) and Sunrise (M = 3.18, SE = 0.16) and between Bohemian and Get Up (M = 2.92, SE = 0.17), and the visual comparison between Bohemian (M = 3.89, SE = 0.13) and Get Up (M = 2.75, SE = 0.16). Participants were clearly much more familiar with the artists in Bohemian, though it is difficult to ascertain to what extent participants were more familiar with Sunrise than Get Up. It is difficult to concretely attribute results of S4 to music familiarity, but it should be nonetheless mentioned that Sunrise participants found that the visual explanation contained the right amount of information much more so than the participants of Bohemian ($p<0.01$). There were significant ($p<0.05$) results between Bohemian and Get Up in terms of the evaluating the textual explanation design for S6-S9. Results were also significant ($p<0.01$) for S9 between Bohemian and Sunrise. In more detail, data from S9 suggests that Survey 1 participants took less time to understand the textual explanation than participants of Survey 2 and Survey 3 due to being more familiar with the songs. While difficult to interpret, there is a trend that shows that more familiarity with a playlist may lead to better understanding, better perceived usefulness, increased trust, and less cognitive effort to understand textual explanations.

One goal of this study was to see if explanations would aid in a user's ability to understand their recommendations.Not surprisingly, the overall results of S7 from Table 9 indicate that the baseline provided users with less useful information than both explanatory designs ($p<0.01$). Overall participant responses to S12 reveal that they would prefer to use a system which

Table 9: Overall MTurk explanation design comparison results (mean±SE) with significance ratings. (n=178; Likert scale 1-5, higher values indicate more agreement with the statement; Significance values are based on comparisons between the baseline and textual explanation designs (*), baseline and visual explanation designs (∗), textual and visual explanation designs (∗), where * = significant at $p<0.05$, ** = significant at $p<0.01$, *** = significant at $p<0.001$ with Bonferroni correction).

Statement	Baseline (n=178)	Text (n=178)	Visual (n=178)
S2	3.59 ±0.08	3.83 ±0.08	**3.87 ±0.08** ∗
S3[a]	**3.70 ±0.09**	3.48 ±0.10	3.69 ±0.10
S4	3.24 ±0.09	3.71 ±0.08 ∗∗∗	**3.75 ±0.08** ∗∗∗
S5[a]	**3.61 ±0.10**	3.11 ±0.10 ∗∗	3.31 ±0.10
S6	3.11 ±0.09	**3.93 ±0.07** ∗∗∗	3.90 ±0.07 ∗∗∗
S7[a]	3.13 ±0.10	**3.61 ±0.09** ∗∗	3.57 ±0.10 ∗∗
S8	3.24 ±0.09	3.70 ±0.08 ∗∗∗	**3.76 ±0.07** ∗∗∗
S9[a]	3.42 ±0.09	**3.58 ±0.10**	3.46 ±0.10
S10	3.46 ±0.08	3.48 ±0.09 ∗	**3.83 ±0.08** ∗∗∗
S11[a]	3.35 ±0.09	3.43 ±0.10	**3.53 ±0.10**
S12	3.30 ±0.09	3.63 ±0.07 ∗∗	**3.69 ±0.08** ∗∗
S13	3.45 ±0.09	**3.79 ±0.10** ∗	3.78 ±0.08 ∗

a/ These were reverse coded so that the high end of the scale, scores of 4 and 5, indicate agreement with the positive version of the statement.

implements either textual (M = 3.63, SE = 0.07) or visual (M = 3.69, SE = 0.08) explanatory designs rather than the baseline (M = 3.30, SE = 0.09) system without explanations ($p<0.01$). There were considerable statistical significances ($p<0.001$) overall for S6 when the baseline (M = 3.11 SE = 0.09) was compared to both the textual (M = 3.93, SE = 0.07) and visual (M = 3.90, SE = 0.07) designs. On an individual survey level, similar results for S6 can be found in Table 10. S6 results for all individual surveys have significance levels of at least $p<0.01$. The smallest difference in Likert scale rating averages between the baseline and explanatory interfaces, cumulative for all three individual surveys, is 0.73. This means there is at least almost 1 full Likert scale point between the baseline mean results and the explanatory interface mean results for S6. In all these comparisons, both the mean values and the levels of statistical significance show that the addition of explanations proved to be very informative. In specific regard to music familiarity, Table 11 shows a statistical significance ($p<0.05$) for textual explanations between the playlist deemed most familiar, Bohemian (M = 4.05, SE = 0.13), and the playlist deemed least familiar, Get Up (M = 3.78, SE = 0.10). No other comparisons have statistical significance for S6. This data suggests that textual explanations may be valued more when the user is already familiar with the music recommendation.

4.4.2 Designing explanations for mobiles - statistics (RQ2)

Research question two addresses how to design informative music recommendation explanations for mobile devices and how users evaluate these design elements. In terms of layout, the visual design was slightly clearer and easier to understand than the baseline. Even though Table 10 shows no statistical significance on an individual survey level for S2, overall results in Table 9 show a significance ($p<0.05$) between the baseline (M = 3.59, SE = 0.08) and the visual design (M = 3.87, SE = 0.08). None of the results for S3 were found to be statistically significant. Overall, results of S4 show a consensus ($p<0.001$) that both the textual (M = 3.71, SE = 0.08) and visual (M = 3.75, SE = 0.08) designs contained the right amount of information, as apposed to the baseline (M = 3.24, SE = 0.09). In terms of individual playlists, Bohemian participants found the textual design (M = 3.75, SE = 0.15) to contain the correct amount of information ($p<0.05$) as apposed to the baseline (M = 3.18, SE = 0.16). Get Up participants had results of the same significance level between the baseline (M = 3.22, SE = 0.14) and the textual (M = 3.71, SE = 0.11) designs. However, both Sunrise and Get Up showed more significance ($p<0.001$) for the baseline-visual comparisons (Sunrise Baseline M = 3.32, SE = 0.15; Sunrise Visual M = 4.08, SE = 0.12; Get Up Baseline M = 3.22, SE = 0.14; Get Up Visual M = 3.86, SE = 0.12). From the results of S5 shown in both Table 9 and Table 10, all the reverse coded mean values for the textual design were the lowest, meaning it was seen as the most detailed. The baseline unsurprisingly had the highest mean values when reverse coded for all results. Significance for baseline-textual comparisons was indicated in both the overall results ($p<0.01$), and also interestingly on an individual level for the Bohemian playlist ($p<0.05$), even though the majority of other results for the Bohemian playlist point towards the textual design as being most favorable in general. The visual design (M = 3.83, SE = 0.08) was considered more attractive than both the baseline (M = 3.46, SE = 0.08, $p<0.01$) and the textual design (M = 3.48, SE = 0.09, $p<0.05$), as seen in the overall results of S10. There was no statistical significance for S11 in any of the comparisons.

4.4.3 Textual versus visual explanations - statistics (RQ3)

Research question three discusses user preferences for textual and visual music recommendation explanations in a mobile setting, which can be answered by comparing the results of almost all statements for the textual and visual designs. Unfortunately, the overall findings in Table 9 reveal almost no statistical significances for the textual-visual comparisons. For significant overall results when compared to the baseline, the textual interface had slightly higher means for S6, S7, S9, and S13, and the visual interface had slightly higher means for S2, S4, S8, S10, S11, and S12. The results of S10 were the only significant ($p<0.05$) findings to be found between the textual (M = 3.48, SE = 0.09) and visual (M = 3.83, SE = 0.08) explanation designs in this table. This

Table 10: MTurk explanation design comparison results (mean±SE) with significance ratings for each individual playlist. (Likert scale 1-5, higher values indicate more agreement with the statement; Significance values are based on comparisons between the baseline and textual explanation designs (∗), baseline and visual explanation designs (∗), textual and visual explanation designs (∗), where * = significant at p<0.05, ** = significant at p<0.01, *** = significant at p<0.001 with Bonferroni correction).

	Survey playlist 1: Bohemian (n=65)			Survey playlist 2: Sunrise (n=50)			Survey playlist 3: Get Up (n=63)		
	Baseline	**Text**	**Visual**	**Baseline**	**Text**	**Visual**	**Baseline**	**Text**	**Visual**
S2	3.65 ±0.14	**3.94 ±0.13**	3.75 ±0.16	3.64 ±0.16	3.86 ±0.16	**4.12 ±0.13**	3.49 ±0.14	3.68 ±0.13	**3.78 ±0.13**
S3 [a]	**3.88 ±0.15**	3.71 ±0.16	3.75 ±0.16	**3.74 ±0.17**	3.38 ±0.21	3.72 ±0.18	3.48 ±0.14	3.33 ±0.16	**3.60 ±0.16**
S4	3.18 ±0.16	**3.75 ±0.15** ∗	3.40 ±0.15	3.32 ±0.15	3.64 ±0.14	**4.08 ±0.12** ∗∗∗	3.22 ±0.14	3.71 ±0.11 ∗	**3.86 ±0.12** ∗∗∗
S5 [a]	**3.80 ±0.15**	3.15 ±0.18 ∗	3.35 ±0.17	**3.58 ±0.19**	2.94 ±0.20	3.34 ±0.19	**3.44 ±0.17**	3.19 ±0.16	3.25 ±0.15
S6	3.08 ±0.16	**4.05 ±0.13** ∗∗∗	3.78 ±0.14 ∗∗	3.16 ±0.15	3.96 ±0.13 ∗∗∗	**4.02 ±0.10** ∗∗∗	3.10 ±0.16	3.78 ±0.10 ∗∗	**3.94 ±0.11** ∗∗∗
S7 [a]	3.43 ±0.16	**3.91 ±0.14**	3.75 ±0.14	3.04 ±0.18	3.58 ±0.19	**3.72 ±0.19** ∗	2.89 ±0.16	**3.33 ±0.15**	3.25 ±0.17 ∗
S8	3.17 ±0.15	**3.88 ±0.12** ∗∗∗	3.66 ±0.12 ∗	3.34 ±0.17	3.72 ±0.16	**3.88 ±0.13**	3.22 ±0.14	3.49 ±0.11	**3.76 ±0.11**
S9 [a]	3.52 ±0.16	**4.00 ±0.15** ∗	3.40 ±0.16 ∗	3.56 ±0.15	3.16 ±0.19	**3.62 ±0.19**	3.21 ±0.14	**3.49 ±0.14**	3.38 ±0.16
S10	3.46 ±0.14	3.71 ±0.14	**3.82 ±0.14**	3.56 ±0.14	3.38 ±0.18	**3.86 ±0.16**	3.38 ±0.14	3.32 ±0.16	**3.83 ±0.13**
S11 [a]	3.54 ±0.16	**3.77 ±0.16**	3.29 ±0.18	3.32 ±0.17	3.24 ±0.20	**3.80 ±0.16**	3.19 ±0.16	3.27 ±0.16	**3.57 ±0.15**
S12	3.26 ±0.15	**3.75 ±0.13** ∗	3.58 ±0.15	3.46 ±0.14	3.52 ±0.18	**3.70 ±0.15**	3.22 ±0.15	3.59 ±0.12	**3.78 ±0.12**
S13	3.38 ±0.16	**3.86 ±0.12**	3.60 ±0.15	3.44 ±0.15	3.76 ±0.16	**4.02 ±0.14** ∗	3.52 ±0.15	3.73 ±0.11	**3.75 ±0.12** ∗

a/ These were reverse coded so that the high end of the scale, scores of 4 and 5, indicate agreement with the positive version of the statement.

would imply, that overall, the visual design is more attractive than the textual one. However, Table 10, which compares designs on an individual survey level, shows no statistical significance for S10 even though all the averages for the visual design were the highest. The only statistical significance (p<0.05) found in this table was in the reverse coded statement of S9 for participants of Survey 1. Results show that it took them the least amount of time to understand the textual explanation. Participants of Survey 1 rated the textual design (M = 4.00, SE = 0.15) higher than the visual design (M = 3.40, SE = 0.16). The overall average for the textual design (M = 3.58, SE = 0.10) is also slightly higher than for the visual design (M = 3.46, SE = 0.10), though with no significance. There may be a slight trend in the data that participants of Survey 1 preferred the textual interface over the visual interface. Looking at the averages alone, participants of Survey 2 and Survey 3 seem to prefer the visual interface the most, though without any statistical significance it is impossible to make any concrete conclusions. Explicit feedback from the following comment analysis in Section 4.5 provides additional insight into these results.

4.4.4 User perceptions of mood explanations - statistics (RQ4)

The final research question examines to what extent user perceptions increase or decrease when affect-based explanations are utilized in a music recommender system when compared to the baseline. Table 9 shows that overall, participants perceived the system as being more trustworthy due to the addition of affect-based explanations. Textual (M = 3.70, SE = 0.08) and visual (M = 3.76, SE = 0.07) designs were significantly (p<0.001) more trustworthy than the baseline (M = 3.24, SE = 0.09). As previously mentioned, results from S8 in Table 11 show that Bohemian

Table 11: MTurk results (mean±SE) with significance ratings comparing the effects of playlist familiarity. (Likert scale 1-5, higher values indicate more agreement with the statement; Significance values are based on comparisons between the bohemian and sunrise playlists (*), bohemian and get up playlists (*), sunrise and get up playlists (*), where * = significant at p<0.05, ** = significant at p<0.01, *** = significant at p<0.001 with Bonferroni correction).

	Baseline Explanation (n=178)			Textual Explanation (n=178)			Visual Explanation (n=178)		
	Bohemian (n=65)	**Sunrise (n=50)**	**Get Up (n=63)**	**Bohemian (n=65)**	**Sunrise (n=50)**	**Get Up(n=63)**	**Bohemian (n=65)**	**Sunrise (n=50)**	**Get Up (n=63)**
S1	**4.20 ±0.12** ******	3.16 ±0.16 ***	2.67 ±0.16 ***	**4.02 ±0.14** ******	3.18 ±0.16 ***	2.92 ±0.17 ***	**3.89 ±0.13** ***	3.38 ±0.17 *	2.75 ±0.16 ****
S2	**3.65 ±0.14**	3.64 ±0.16	3.49 ±0.14	**3.94 ±0.13**	3.86 ±0.16	3.68 ±0.13	3.75 ±0.16	**4.12 ±0.13**	3.78 ±0.13
S3 [a]	**3.88 ±0.15**	3.74 ±0.17	3.48 ±0.14	**3.71 ±0.16**	3.38 ±0.21	3.33 ±0.16	**3.75 ±0.16**	3.72 ±0.18	3.60 ±0.16
S4	3.18 ±0.16	3.32 ±0.15	3.22 ±0.14	3.75 ±0.15	3.64 ±0.14	3.71 ±0.11	3.40 ±0.15 **	**4.08 ±0.12** **	3.86 ±0.12
S5 [a]	**3.80 ±0.15**	3.58 ±0.19	3.44 ±0.17	3.15 ±0.18	2.94 ±0.20	3.19 ±0.16	3.35 ±0.17	3.34 ±0.19	3.25 ±0.15
S6	3.08 ±0.16	**3.16 ±0.15**	3.10 ±0.16	**4.05 ±0.13** *	3.96 ±0.13	3.78 ±0.10 *	3.78 ±0.14	**4.02 ±0.10**	3.94 ±0.11
S7 [a]	**3.43 ±0.16**	3.04 ±0.18	2.89 ±0.16	**3.91 ±0.14** *	3.58 ±0.19	3.33 ±0.15 *	**3.75 ±0.14**	3.72 ±0.19	3.25 ±0.17
S8	3.17 ±0.15	**3.34 ±0.17**	3.22 ±0.14	**3.88 ±0.12** *	3.72 ±0.16	3.49 ±0.11 *	3.66 ±0.12	**3.88 ±0.13**	3.76 ±0.11
S9 [a]	3.52 ±0.16	**3.56 ±0.15**	3.21 ±0.14	**4.00 ±0.15** ***	3.16 ±0.19 **	3.49 ±0.14 *	3.40 ±0.16	**3.62** ±0.19	3.38 ±0.16
S10	3.46 ±0.14	**3.56 ±0.14**	3.38 ±0.14	**3.71 ±0.14**	3.38 ±0.18	3.32 ±0.16	3.82 ±0.14	**3.86 ±0.16**	3.83 ±0.13
S11 [a]	**3.54 ±0.16**	3.32 ±0.17	3.19 ±0.16	**3.77 ±0.16**	3.24 ±0.20	3.27 ±0.16	3.29 ±0.18	**3.80 ±0.16**	3.57 ±0.15
S12	3.26 ±0.15	**3.46 ±0.14**	3.22 ±0.15	**3.75 ±0.13**	3.52 ±0.18	3.59 ±0.12	3.58 ±0.15	3.70 ±0.15	**3.78 ±0.12**
S13	3.38 ±0.16	3.44 ±0.15	**3.52 ±0.15**	**3.86 ±0.12**	3.76 ±0.16	3.73 ±0.11	3.60 ±0.15	**4.02 ±0.14**	3.75 ±0.12

a/ These were reverse coded so that the high end of the scale, scores of 4 and 5, indicate agreement with the positive version of the statement.

participants found the textual explanation to be more trustworthy than Get Up participants (p<0.05). Table 10 reveals that Bohemian participants trust both the textual design (M = 3.88, SE = 0.12, p<0.001) and the visual design (M = 3.66, SE = 0.12, p<0.05) more than the baseline (M = 3.17, SE = 0.15). In terms of cognitive effort, reverse coded responses for S9 from the same table reveal with significance (p<0.05) that it took Bohemian participants less time to understand the textual explanations (M = 4.00, SE = 0.15) than the visual (M = 3.40, SE = 0.16) ones. The same holds true for the overall responses of S9, albeit marginally with with no statistical significance. People answered S13 with positive significance overall (p<0.05), informing that the addition of mood-based explanations lead to higher perceived user satisfaction. Looking at the mean values alone without statistical significance in Table 10, participants of Survey 1 were most satisfied with the textual explanation design, but also more satisfied with the visual explanations than the baseline. In terms of statistically significant results, both Survey 2 and 3 participants were also more satisfied with the explanatory interfaces, but unlike Survey 1 show significance (p<0.05) for the baseline-visual comparison.

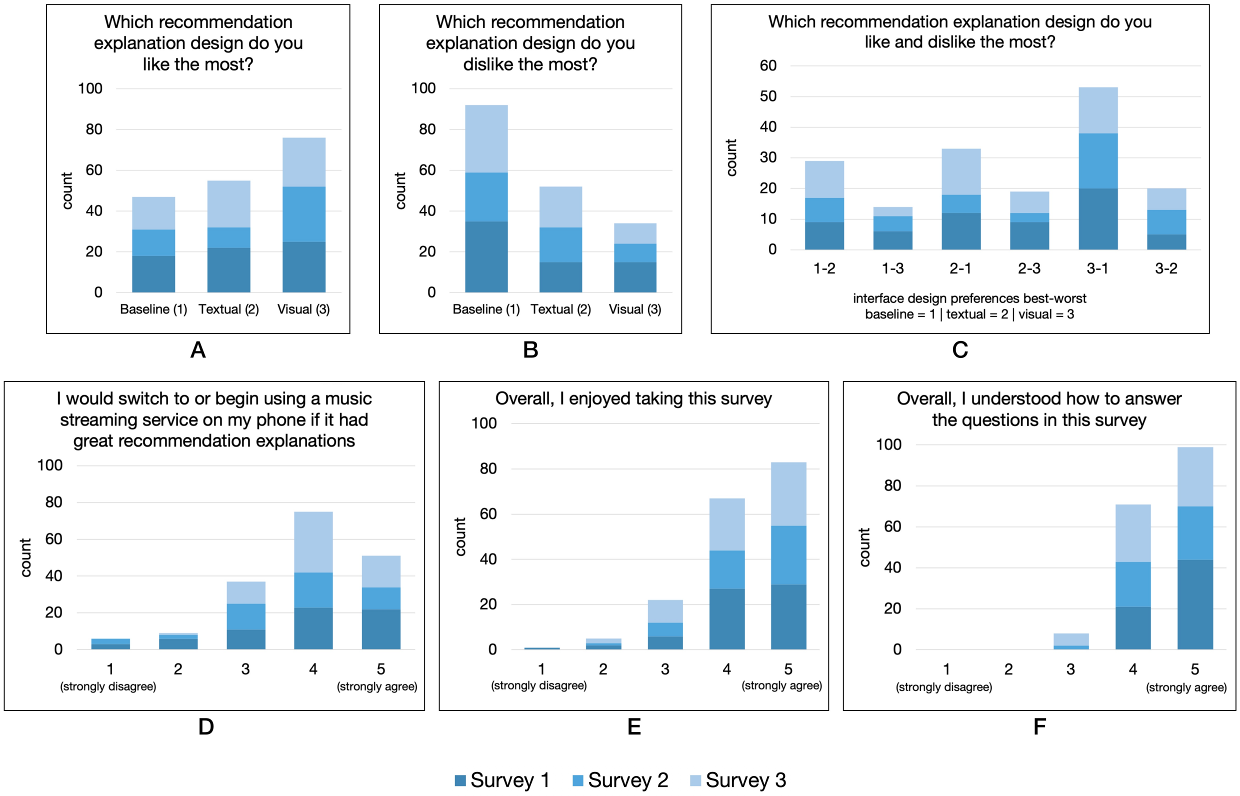


Figure 45: MTurk post-survey results (n=178).

4.5 MTurk Participants' Interface Preferences and Comment Analysis

After comparing interfaces, participants specified their interface preferences, discussed their explanation design preferences in general, and then expressed their opinions about using mood data. Verbatim comments from all 178 participants were categorized and dissected in order to better understand how people perceived the music recommendation explanations. The themes and specific sentiments within, are referred to as constructs and codes respectively in Table 14 and Table 15. Sentiments which were not mentioned by at least 10 people were excluded. Participants are designated by their Survey number S1, S2, or S3, alongside their participant number P1, P2 etc. So, the first participant from the first Survey would be referred to as (S1P1). Comments below are grouped by the type of the interface preferred by the user, with an additional section concerning explanations in general.

Table 12: MTurk explanation design preferences based on gender. (Note: 2 participants associated as other genders and are not included in this table).

Preference	Design	Women (n=76)	Men (n=100)
Liked Most	Baseline (1)	22%	30%
	Text (2)	36%	28%
	Visual (3)	42%	42%
Disliked Most	Baseline (1)	53%	50%
	Text (2)	28%	31%
	Visual (3)	20%	19%

Table 13: MTurk explanation design preferences based on age group.

Preference	Design	18-24 (n=11)	25-34 (n=77)	35-44 (n=50)	45-54 (n=25)	55+ (n=15)
Liked Most	Baseline (1)	11	36	0	0	0
	Text (2)	0	41	14	0	0
	Visual (3)	0	0	36	25	15
Disliked Most	Baseline (1)	3	25	39	14	11
	Text (2)	5	26	7	10	4
	Visual (3)	3	26	4	1	0

4.5.1 Overall explanation design preferences

Overall results from Figure 45 show that 26% of all participants preferred the baseline, 31% preferred the textual design, and 43% preferred the visual design. Interesting, the likes and dislikes of interfaces based on gender were fairly similar for both men and women, as described in Table 12. Even more interesting, Table 13 shows that younger participants preferred the baseline, middle aged participants preferred the textual design, while older participants preferred the visual design. There may be a correlation between design preferences and age, though a statistical analysis of this was not completed. People who took Survey 1 would rather have textual explanations over none, and visual explanations over text though only marginally. Over half of the people who took Survey 2 stated that they preferred the visual design over the other two, yet more were in favor of the baseline than the textual. For Survey 3, more people preferred the explanatory interface designs much more than the baseline, however they seem to like the textual and visual designs almost equally. In terms of which interface design participants liked least, there was a bit more consensus. Just over half of all participants, 52%, disliked the baseline the most. Graph C shows the relationship between liked and disliked explanation designs where the majority of people who preferred the baseline, disliked the textual design most, and that the majority of people who preferred either explanatory design, disliked the baseline the most. However, there does not seem to be one clear cut discernible pattern between a user's likes and

Table 14: Comment analysis for preferred explanation design. (Number of frequently mentioned terms per explanation design).

Construct	**Code**	**Baseline (n=47)**	**Textual (n=55)**	**Visual (n=76)**	**Total (n=178)**
Data	Graphic comparisons	-	-	15	15
	Precise percentages	-	16	3	19
	Right amount of information	2	12	4	18
Design	Aesthetically pleasing	6	5	35	46
	Clean layout	17	12	15	44
	Colors	-	2	19	21
	Detailed	-	7	6	13
Efficiency	Straight to the point	4	6	20	30
Usability	Easy to understand	17	12	11	40

dislikes. When it comes to the survey in general, participants overwhelmingly enjoyed completing the surveys and understood how to do so. In regard to RQ1, results from the post-survey question *"I would switch to or begin using a music streaming service on my phone if it had great recommendation explanations"* also reveal that the majority of people seem to desire better music recommendation explanations than what their music streaming services currently provide (M = 3.88, SE = 0.07, SD = 0.99). Overall, the findings seem to suggest that users find recommendation explanations to be important.

4.5.2 Preferred baseline design

47 participants who preferred the baseline interface without explanations mentioned that it was due to its clean layout and felt the system was more capable of being understood by users than those which included explanations. Feedback from these participants seem to indicate that they are more comfortable using familiar technology as opposed to trying something new, especially when some felt that the explanatory interfaces contained too much useless information. One person explained how:

> *"Every recommendation doesn't need to disclose that it was chosen based on 15 parameters"*. Referring to his dislike of the textual interface, he goes on to say, *"I want clean and simple. I don't want something that would take an engineering degree to decipher. I'm trying to listen to music, not design a jet engine"*. (S1P62)

Although not included in the tables as less than 10 people mentioned it, 4 users who preferred the baseline wrote that familiarity with the recommendations made them less likely to need the explanations, although this sentiment holds true for a handful of participants who actually preferred the explanatory interfaces as well. One female participant stated:

> *"I think it provides too much information that people don't really need. Most people understand what mood they get from different types of music already. Most people will be familiar with the songs that are recommended anyway"*. (S1P27)

4.5.3 Preferred textual design

The 55 people who liked the textual explanation design the most perceived it as being specific, precise, and detailed. Ironically, these aspects of this design were also the reason why 20 people disliked this interface. Focusing on those who preferred this design, 12 of them mentioned how they liked how the comparison breakdown of the similarity score was done with numbers. The balanced and concise information is easy for them to understand and the amount of specific detail makes the design look the most complete. One participant who disliked the baseline, as it lacked an explanation button, described his preference for the textual interface as:

> *"It gives you the analytical breakdown on why it is recommended. It doesn't waste your time with glitz and glamour since it's info you went out of your way to inquire about I like it straightforward"*. (S2P40)

16 people explicitly wrote that they chose the textual explanation design due to the use of percentages and numbers as they are more precise than graphs. Two female participants explained:

> *"Interface 3 is aesthetically pleasing but the bar graph being used is not helpful, it's a bit confusing. Since it doesn't have the numerical data to go along with it, I am clueless as to what the bars truly mean. I understand the length is supposed to represent something but the something I don't understand. Too confusing and too much assumptions going into it"*. This scientific researcher continues by saying that, *"personally I like when there is numerical data to guide my thoughts. I am able to differentiate easily between different elements. Seeing the percentages, allows me to actually fully understand why the recommendation was made, I don't have to assume nor do I need to look for more information as everything I need is giving to me"*. (S1P52)

> *"I like the amount of information it is giving you with data and facts but making it very clear of what it means but the design is friendly and welcoming... I don't like the graphic design it looks too industry type and not relatable for people to want to learn and understand of trust in whats being said"*. (S3P33)

Table 15: Comment analysis for least preferred explanation design. (Number of frequently mentioned terms per explanation design).

Construct	**Code**	**Baseline (n=92)**	**Textual (n=52)**	**Visual (n=34)**	**Total (n=178)**
Data	Too little information	31	1	2	34
	Too much information	-	13	3	16
Design	Unattractive	26	4	4	34
	Unorganized layout	-	4	6	10
Transparency	Not explanatory enough	33	1	-	34
Usability	Difficult to understand	7	20	12	39

4.5.4 Preferred visual design

Out of the 76 participants who preferred the visual design, the most important explicit feedback that emerged were comments about the colorfulness, efficiency, and easy to read layout of the visual interface. 20 people found the visual design to be the most straightforward, as opposed to only 4 and 6 people who preferred the baseline and textual interfaces respectively. Three participants, who disliked the baseline the most as it provides no explanations, wrote:

> *"Many details can be absorbed in a glance compared to the others"*. (S1P39)

> *"I like interface 3 the most because it balances providing extra information for why the system recommended the songs while also presenting this information is a nice appealing and easy to understand way. The visuals of the bar graph are superior to the statistics in my opinion which are dry and hard to look at"*. (S2P1)

> *"You can see why it was chosen just by looking at the bars, no reading or comparing numbers"*. Yet, she also specified that maybe the explanations include *"too many categories"* and she *"wasn't sure about what some of them meant"*. (S3P59)

35 people declared that they preferred the visual design as it was the most aesthetically pleasing, and 19 mentioned how they liked the use of colors. At the same time, there were underlying sentiments that the interface could be improved with the addition of numbers or by being able to customize different explanations. For example:

> *"It is simply the most aesthetically appealing and the bar graphs really help with engagement... [but] I like the designs and percentages. I'm a numbers guy, so seeing the statistics really gets me involved. I like how precise the numbers get as well"*. He disliked interface 1 because *"It is too plain and disengaging"*. (S1P16)

Table 16: MTurk participant opinions on the helpfulness of affective data. (Question: *Were the mood recommendation explanations helpful?).*

Answer	**Bohemian (n=65)**	**Sunrise (n=50)**	**Get Up (n=63)**	**Total (n=178)**
Yes	45	33	38	116
Somewhat	7	5	7	19
No	8	8	7	23
Unclear	5	4	11	20

> *"It's the most visually appealing and although I appreciate the numbers (percentages), it's much, much faster to process the comparisons using the graphic bars in Interface 3"*. However, she also construes that explanations might not always be necessary, stating that *"because of my familiarity with the artists, I can easily see why they're recommended, but in an instance where I wasn't, [the baseline] wouldn't tell me much"*. (S1P26)

> *"I really like the colors that are used for the different moods on the graphs, it really gives the recommendation explanation some personality, since a lot of people use colors to express moods. I also like the bar graphs for the different moods as, at a glance, it is easy to see which moods represent the song and playlist easily. This also makes it easier to compare the different moods which to me is what I am doing when I look at the recommendations"*. He goes on to say that he would like *"a customization section that allows you to set the list of moods that are used to determine the playlists"*. (S3P21)

> *"I like how colorful it is, and I like that I can easily compare the two sides without looking at numbers and having to concentrate on them. This is super easy. While I found it a little difficult to interpret at first (something about the bars both signaling an increase even though they're pointed in the opposite direction), it's such an attractive visual. It doesn't feel overwhelming to look at these soft colors on a black/grey background"*. (S3P31)

4.5.5 General perceptions of mood-based explanations

Specifically in terms of utilizing affective information, Table 16 shows that 116 of 178 participants explicitly wrote that the mood explanations were helpful. 17 participants acknowledged that they either often want to build playlists based on moods, or listen to songs based on their current mood. Two interesting participant responses to the question *"What do you think about using moods for recommendation explanations?"* are:

> *"I think the mood based recommendations are helpful for trying to figure out which music can be appropriate in certain contexts. Of course feelings are subjective so it's not going to be right all the time but in general it is useful to have in my opinion"*. (S2P1)

> *"Yes! My playlist often emphasizes a mood I am in or wish to get in, and so that is extremely useful and I like it and would use something like that regularly to try out new music"*. (S2P6)

Regardless of the similarity metric being used, it could be argued of all algorithms that none of them will ever be 100% perfect. Therefore, some users may be reluctant to adopt explanations. 11 participants mentioned that while the idea of mood-based explanations is interesting, explanations may not be necessary so they may or may not use them. For example:

> *"I think in certain circumstances it could be beneficial. For me though, it is just more stuff to click through potentially. I just want to get down to the point and get music"*. (S1P9)

> *"It is something I might or might not use. It is not something I miss having"*. (S3P42)

Overall however, most participant comments positively mimicked those of the statistical results which revealed that participants would like more recommendation explanations and would switch to or being using a new system if it had great recommendation explanations. After completing this study, many users are more aware of recommendation explanations and most seem to perceive a system as being more valuable if it includes explanations. For example:

> *"I would use recommendations a lot more often since I can see why they were chosen"*. (S1P23)

> *"Well, if this system were implemented, I would pay much more attention to recommendations! It's a royal pain when you don't know WHY something has been recommended to you, and makes you wary of trying new music simply because the system only says "you'll like it, trust me" (after having recommended things that are not at all like what you're listening to by choice!)"*. (S1P31)

> *"In general I like how they increase my understanding of why the system recommends the things it does"*. (S2P1)

> *"I now think that recommendation explanations are even more important. I see how an insight into the algorithms can help you identify your preferences in a more consistent manner, too"*. (S3P31)

> *"Using moods for recommendations meets the needs of the user"*. (S3P58)

5 SUMMARY AND CONCLUSION

This final Chapter provides a summary of the main findings, revealing the implications of this study for future research. Section 5.1 discusses the overall findings of the research study in relation to the research questions. Section 5.2 points out potential drawbacks with the research methodology, constraints related to the time-frame of this thesis, in addition to other issues related to survey questions and results. It also highlights which aspects of this study researchers should make further inquires into and concludes this thesis.

5.1 Discussion

The main goal of the presented study was to perform a comparative user evaluation of two affect-based explanatory interface designs against a non-explanatory baseline. Moodplay's approach to mood-based explanations were explored further by applying their latest design to the realm of mobile devices. First, the need for explanations in the mobile domain was established through a systematic literature review and 2 pre-studies. These findings also chartered the way forward on how to design such explanations, which lead to creating the Spotify lookalike baseline, textual, and visual interfaces. 210 Amazon Mechanical Turkers, divided into three groups by playlist type, answered an online questionnaire comparing the three different interfaces; the results of which were compiled from the 178 people who passed the attention checks. In the following, the conclusions drawn from this research are summarized by research question.

5.1.1 The value of explanations - discussion (RQ1)

The first research question asked, *"To what extent are explanations in mobile music recommender applications valued by users, and how does music familiarity affect that?"*. The premise for asking this question was based on the primarily positive feedback from previous studies, displayed in Table 3, regarding the importance of recommendation explanations. These findings were looked at from a UX perspective in order to design music recommendation explanations for mobile devices, where there is a gap in the present state-of-the-art. The research completed in this thesis was able to substantiate the claim that users would like better music recommendation explanations as evident by looking back at Figures 36, 39, and 44. At the same time, results in Figure 44 also indicate that while explanations would be nice to have, the majority of users expressed that they already understand their music recommendations and are actually slightly more interested in being able to control them. Figure 45 shows that the majority of users stated they would switch to or begin using a music streaming service on their mobile phone if it had great music recommendation explanations. As expected, Table 9 depicts a large statistical significance in favor of the explanatory interfaces over the baseline interface.

Many of the participants who preferred the baseline commented that they did not need explanations as they were already familiar with the music being recommended to them, or assumed that they would be. There was a tendency for Survey 1 participants, whom evaluated the Bohemian playlist containing the most well know artists, to prefer the textual interface design. It could be speculated that if this type of person were to ever utilize explanations, they would only do so to collect more in-depth information to obtain new knowledge. This may be the reason they preferred the textual design, which provided users with a precise percentage breakdown of mood data. Some of the participants expressed that when music recommendations are unfamiliar, they want a quick way of deciding if the recommendation is right for them. That's why they preferred the visual design. While interesting, there does not seem to be enough evidence of statistical significance to make any strong definitive conclusions about the relationship between music familiarity and explanation utilization.

5.1.2 Designing explanations for mobiles - discussion (RQ2)

The second research question asked, *"How do users evaluate the different design elements of music recommendation explanations in a mobile setting?"*. The comment analysis from Section 4.5 paints a picture of what explanation features users are most concerned with. According to Table 14 and Table 15, the main themes surrounding these important factors were data (type and amount of information), design (look of the explanation), efficiency (cognitive load), transparency (understanding how and why something is recommended), and usability (ease of use). In terms of information types, these studies uncovered that users would like to know more about musical moods and genres. There were mixed opinions about information quantity, though users preferred explanations over none, so it is in the best interest of the designer to allow users to regulate this aspect of explanations if possible. Table 9 shows that mean scores were reasonably high for S2 and S4, and the non-coded means for S3 and S5 were reasonably low, showing that the amount of information and level of detail of both the textual and visual designs were acceptable. In both the second pre-survey and the MTurk study, users mentioned that they value the level of detail and the use of color in explanation designs. Although Holm (2012) found that it is difficult to visualize genres through only colors, this study may have seen more positive results due to the addition of labels and since the colors were based on moods. While there is a significance between the baseline and visual in terms of layout, almost equal amounts of participants with different design preferences commented that they chose that particular design as their favorite based on its clean layout. The findings highlight that it is difficult, if not impossible, to create one explanation design which would please all users. Designers must balance user preferences and explanatory goals, as user perceptions and the outcomes of the goals can be greatly affected by the explanation design.

5.1.3 Textual versus visual explanations - discussion (RQ3)

The third research question asked, *"To what extent do users prefer these explanations to be either textual or visual?"*. User feedback shows that attractiveness is important. Table 9 shows that the most aesthetically pleasing design was the visual. There were no other significant overall results for any textual-visual comparisons. Many of the participants who preferred the textual explanation adamantly stated a sentiment similar to *'I'm just more of a numbers person'*, while those who preferred the visual explanation expressed the exact opposite sentiment. Statistically speaking, numbers show that people found the visual explanation to be slightly less detailed than the textual one. Textual explanations were generally preferred by users who wanted raw explicit detailed information. The visual design was described as being simpler or less complex than the textual and generally preferred by users who wanted a quick way to compare recommendations. A handful of people actually suggested that the visual design should include more details, such as percentages, more in line with the information shown in the textual interface. The main conclusion to take away from these results is that humans are very diverse and process information in different way, and as such, have different preferences on what design they like. This is in line with the findings from He et al. (2016).

5.1.4 User perceptions of mood explanations - discussion (RQ4)

The fourth research question asked, *"To what extent do affect-based explanations influence a users perception of the system and its music recommendations?"*. Explanations can be seen a precursor to achieving an overall better user experience. Both explanatory interfaces did significantly better than the baseline in terms of ease of use, transparency, trust, and satisfaction. Additionally, the 47 MTurk participants who did not like the explanatory interfaces had a lower overall satisfaction with their preferred interface, the baseline, than those who chose an explanatory interface. Participants also mentioned that explanation efficiency is important to them, however both explanatory designs prevailed similarly over the baseline with no clear-cut winner in this category. Statistics from Table 10 show that it took participants of Survey 1 less time to understand the design of the textual explanation than the visual explanation, however there were no other statistical significances between or within Surveys. Specifically regarding affective information, very few participants mentioned either that they did not understand how the moods were matched to the music, or that as mood classifications are not accurate and open to interpretation they should not be used as a recommendation explanation metric. Generally, people feel that there is a strong connection between music and moods and therefore consider moods to be a good way to represent recommendation explanations.

5.1.5 Key research question conclusions

The answers to each research question can be summed up into 4 short statements:

RQ1: Users perceive the addition of explanations as an improvement to current mobile music applications which have little to no explanations, though music familiarity may slightly affect their perceptions.

RQ2: Users' design evaluations of mobile music recommendation explanations vary due to differences in personal preferences.

RQ3: No significant difference between textual and visual explanations was found.

RQ4: Moods are an acceptable way to explain music recommendations and the use of affective information may help systems become more efficient, transparent, trustworthy, and satisfactory.

5.2 Limitations and Future Research

Several limitations should be acknowledged when interpreting the results of this research. Order bias may have occurred in both the second pre-survey and the main Mturk study. In the second pre-survey, all designs were shown together side by side, so some may have ranked them partially based off of the order in which they were placed on the page. The main study was meant to be a controlled lab-study where 24 new participants fully tested 3 final high fidelity prototypes in random order with a latin-square design, but this was not possible due to the corona virus outbreak. Given having 3 conditions, the 3 different interfaces, and 24 people, 4 people would be randomly assigned to each one of the following test orders: 123, 132, 213, 231, 312, 321. In other words, 4 people would complete the tasks first with interface A then B then C, another 4 would complete the tasks first with interface A then C then B, etc. There could potentially be a positive bias towards the explanatory interfaces as the baseline was always shown first, which could also be exacerbated by the fact that users were aware of the research hypothesis, that music recommender systems do not have enough explanations, before starting the survey. The fact that the research study was not interactive is also a limitation as it was unable to properly analyze the system in terms of usability. One user specifically mentioned that it was difficult to answer questions properly without having experience using the systems. Originally, users were going to be asked to fill out a SUS questionnaire, but that did not fit with the MTurk online testing model. In addition, not all users watched the videos presented. User feedback shows that some of these participants did not completely understand the ‘WHY?’ button or the pop-up interaction. It is difficult to interpret user feedback regarding system use intention when none of the participants actually used the application. In general, more user testing needs to be conducted interactively in order to determine the best way for explanations to be implemented practically in a mobile music application. There may have also some confusion around recommendation generation and explanation design. Further studies may benefit from explicitly explaining to users that all recommendation explanations are generated in the same fashion, just like Herlocker et al. (2000) did.

This research study has proved that music recommendation explanations for mobile devices is a subject area which should be further explored. This should also be an area of interest to major music streaming service companies as including explanations could potentially be considered a competitive advantage as the main user study in this thesis showed that the majority of users would be willing to switch providers in order to receive better recommendation explanations. One impediment to implementing explanations in commercial music streaming services is that customers are already satisfied with the status quo, as millions of people use music streaming services regardless of how pleased they are with the system. Many participants mentioned that previously, they were unaware that music recommendations could be more explanatory. The textual and visual explanations tested in this thesis look like promising designs which could be seamlessly integrated into an existing music mobile application. Further research should also look at including both a textual and a visual explanation option in a modern day music steaming service application for users to choose between. It was not within the scope of this thesis to explore the relationship between control over music recommendations and the need for explicit music recommendation explanations. Control comes in a variety of forms and has been experimented with, to different extents, in many of the previous works referenced in this thesis. While controllability may not have been a main objective, Lillie (2008), Goto & Goto (2009), Gretarsson et al. (2010), Guo et al. (2011), Saito & Itoh (2011), Bogdanov et al. (2013), Lehtiniemi & Holm (2013), Kamalzadeh et al. (2016), Andjelkovic, Parra, & O’Donovan (2019), Millecamp et al. (2019), and Vavrille et al. (n.d.), all show at least one example where users were provided with some form of control over their music recommendations. As mobile devices have small screens, allowing users to fine tune their recommendations or provide them with personalized recommendations may help with spacial issues. It should be noted that a small screen does not necessarily correlate with users needing or wanting less explanations. When conducting future research, consider the following when designing new music recommendation explanations for mobile devices as not one explanation fits all. **Who** is using the system? Users have different preferences. Consider personalization. **What** kind of explanations? Moods, genres, and similarity score are all viable data-types for music recommendation explanations. Colorful designs, whether textual or visual, fair better than black and white. **Where** to place explanations? Normative or pragmatic. It's better to let the user choose when they want to see detailed explanations, rather than being shown on screen, so they don't occupy too much screen space. **When** to show explanations? Context. Users may not always need to be informed. **Why** explanations? Explanations should match explanatory goals. **How much** information to display? Too much data leads information overload. Allow for controllability to reduce cognitive load.

5.3 Open Science

This thesis participates in open science by sharing the results of this work freely and openly with the public in order to make this study reproducible. This master's thesis will be published in the University of Bergen's Open Research Archive (BORA)[16]. The results from this study can be found in a Github repository[17]. Along with the raw data, a snippet of R code used to calculate the p-values in RStudio is also included along with an example csv file. For possible access to the Moodplay Dataset, it is advised to speak with the authors of Andjelkovic, Parra, & O'Donovan (2019).

[16] http://bora.uib.no/

[17] https://github.com/akbobrow/RecSysThesis

References

Åman, P., & Liikkanen, L. A. (2010). A survey of music recommendation aids. In *Ceur workshop proceedings* (Vol. 633, pp. 25–28).

Andjelkovic, I., Parra, D., & O'Donovan, J. (2016). Moodplay: Interactive Mood-based Music Discovery and Recommendation. In *Proceedings of the 2016 conference on user modeling adaptation and personalization - umap '16* (pp. 275–279). New York, New York, USA: ACM Press. doi: 10.1145/2930238.2930280

Andjelkovic, I., Parra, D., & O'Donovan, J. (2019, 1). Moodplay: Interactive music recommendation based on Artists' mood similarity. *International Journal of Human Computer Studies*, *121*, 142–159. doi: 10.1016/j.ijhcs.2018.04.004

Andjelkovic, I., Parra, D., O'Donovan, J., & Herrera, R. (2019). *Moodplay.* Retrieved from http://moodplay.pythonanywhere.com/

Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E., & Herrera, P. (2013, 1). Semantic audio content-based music recommendation and visualization based on user preference examples. *Information Processing & Management*, *49*(1), 13–33. doi: 10.1016/J.IPM.2012.06.004

Bostandjiev, S., O'Donovan, J., & Höllerer, T. (2012). TasteWeights: A visual interactive hybrid recommender system. *Proceedings of the 2012 Conference on Recommender Systems*, 35–42. doi: 10.1145/2365952.2365964

Budiu, R. (2018, 5). *Between-Subjects vs. Within-Subjects Study Design.* Nielsen Norman Group. Retrieved from https://www.nngroup.com/articles/between-within-subjects/

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi: 10.1177/1745691610393980

Burke, R. D. (2007). Hybrid Web Recommender Systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web. lecture notes in computer science* (Vol. 4321, pp. 377–408). Heidelberg: Springer.

Cardoso, B., Sedrakyan, G., Gutiérrez, F., Parra, D., Brusilovsky, P., & Verbert, K. (2018). IntersectionExplorer, a multi-perspective approach for exploring recommendations. *International Journal of Human Computer Studies*, *121*(May 2017), 73–92. doi: 10.1016/j.ijhcs.2018.04.008

Chen, Y. (2015). *Social Interface and Interaction Design for Group Recommender Systems* (Unpublished doctoral dissertation). Swiss Federal Institute of Technology in Lausanne, Lausanne.

Chen, Y., Ma, X., Cerezo, A., & Pu, P. (2014). Empatheticons: Designing Emotion Awareness Tools for Group Recommenders. In (pp. 1–8). doi: 10.1145/2662253.2662269

Chen, Y., & Pu, P. (2013). CoFeel: Using emotions to enhance social interaction in group recommender systems. *Workshop on Tools and Technology for Emotion-Awareness in Computer Mediated Collaboration and Learning*, 3–4. Retrieved from http://www2.immersion.com/developers

Dresch, A., Lacerda, D. P., & Antunes Jr, J. A. V. (2015). *Design Science Research: A Method for Science and Technology Advancement* (3rd ed.). Heidelberg. doi: 10.1201/b16768-27

Eiband, M., Schneider, H., & Buschek, D. (2018). Normative vs Pragmatic: Two Perspectives on the Design of Explanations in Intelligent Systems. In *Proceedings of the 23rd international conference on intelligent user interfaces - iui '18.* Munich.

Goto, M., & Goto, T. (2009). Musicream: Integrated Music-Listening Interface for Active, Flexible, and Unexpected Encounters with Musical Pieces. *Journal of Information Processing*, *17*, 292–305. doi: 10.2197/ipsjjip.17.292

Gretarsson, B., O'Donovan, J., Bostandjiev, S., Hall, C., & Höllerer, T. (2010). SmallWorlds: Visualizing social recommendations. *Computer Graphics Forum*, *29*(3), 833–842. doi: 10.1111/j.1467-8659.2009.01679.x

Guo, J., Zhang, X. L., Wu, L., Gou, L., & You, F. (2011). SFViz: Interest-based Friends Exploration and Recommendation in Social Networks. In (pp. 1–10). doi: 10.1145/2016656.2016671

He, C., Parra, D., & Verbert, K. (2016). Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, *56*, 9–27. doi: 10.1016/j.eswa.2016.02.013

Herlocker, J. L. (2000). *Understanding and improving automated collaborative filtering systems* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). *Explaining Collaborative Filtering Recommendations* (Tech. Rep.). Minneapolis: University of Minnesota.

Hevner, A. R. (2007). Scandinavian Journal of Information Systems A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems © Scandinavian Journal of Information Systems*, *19*(192), 87–92. Retrieved from http://aisel.aisnet.org/sjishttp://aisel.aisnet.org/sjis/vol19/iss2/4

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *Design Science in IS Research MIS Quarterly*, *28*(1), 75–105.

Holm, J. (2012). *Visualizing Music Collections Based on Metadata: Concepts, User Studies and Design Implications* (Unpublished doctoral dissertation). Tampere University of Technology.

Holm, J., Lehtiniemi, A., & Eronen, A. (2010). Evaluating an avatar-based user interface for discovering new music. In *Proceedings of the 9th international conference on mobile and ubiquitous multimedia - mum '10* (pp. 1–10). doi: 10.1145/1899475.1899484

Hu, X. (2019). Evaluating mobile music services in China: An exploration in user experience. *Journal of Information Science*, *45*(1), 16–28. doi: 10.1177/0165551518762070

IFPI. (2018). *Music Consumer Insight Report* (Tech. Rep.). Retrieved from https://www.ifpi.org/downloads/Music-Consumer-Insight-Report-2018.pdf

Iqbal, M. (2020, 5). *Spotify Usage and Revenue Statistics (2020).* Retrieved from https://www.businessofapps.com/data/spotify-statistics/

Jannach, D., Kamehkhosh, I., & Bonnin, G. (2019). Music Recommendations. In S. Berkovsky, I. Cantador, & D. Tikk (Eds.), *Collaborative recommendations: Algorithms, practical challenges and applications* (pp. 481–518). Covent Garden, London: World Scientific Publishing Co. Pte. Ltd.

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2011). *Recommender Systems - An Introduction*. Cambridge, United Kingdom: Cambridge University Press.

Kamalzadeh, M., Kralj, C., Möller, T., & Sedlmair, M. (2016). TagFlip: Active Mobile Music Discovery with Social Tags. In (pp. 19–30). doi: 10.1145/2856767.2856780

Kitchenham, B., & Brereton, P. (2013, 12). A systematic review of systematic review process research in software engineering. *Information and Software Technology*, *55*(12), 2049–2075. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0950584913001560 doi: 10.1016/j.infsof.2013.07.010

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3* (Vol. 45; Tech. Rep. No. 4). doi: 10.1145/1134285.1134500

Knijnenburg, B. P., & Willemsen, M. C. (2015). Evaluating Recommender Systems with User Experiments. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (Second ed., pp. 309–352). Boston, MA: Springer US. doi: 10.1007/978-1-4899-7637-6{_}9

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 441–504. Retrieved from https://link.springer.com/content/pdf/10.1007%2Fs11257-011-9118-4.pdf doi: 10.1007/s11257-011-9118-4

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013). Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the sigchi conference on human factors in computing systems - chi '13* (pp. 207–216). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=2470654.2470684 doi: 10.1145/2470654.2470684

Kouki, P., Getoor, L., Pujara, J., Schaffer, J., & O'Donovan, J. (2017). User Preferences for Hybrid Explanations. In (pp. 84–88). doi: 10.1145/3109859.3109915

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research Methods in Human-Computer Interaction* (Second ed.). Cambridge, MA: Elsevier.

Lehtiniemi, A., & Holm, J. (2011). Easy Access to Recommendation Playlists: Selecting Music by Exploring Preview Clips in Album Cover Space. *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, 94–99. doi: 10.1145/2107596.2107607

Lehtiniemi, A., & Holm, J. (2013). Designing for Music Discovery: Evaluation and Comparison of Five Music Player Prototypes. *Journal of New Music Research*, *42*(3), 283–302. doi: 10.1080/09298215.2013.796997

Lillie, A. S. (2008). *MusicBox: Navigating the space of your music* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Millecamp, M., Htun, N. N., Conati, C., & Verbert, K. (2019). To explain or not to explain. In *Proceedings of the 24th international conference on intelligent user interfaces - iui '19* (pp. 397–407). New York, New York, USA: ACM Press. doi: 10.1145/3301275.3302313

Millecamp, M., Htun, N. N., Jin, Y., & Verbert, K. (2018). Controlling Spotify Recommendations. In *Proceedings of the 26th conference on user modeling, adaptation and personalization - umap '18* (Vol. 18, pp. 101–109). New York, New York, USA: ACM Press. doi: 10.1145/3209219.3209223

Nave, G., Minxha, J., Greenberg, D. M., Kosinski, M., Stillwell, D. J., & Rentfrow, P. J. (2018). Musical Preferences Predict Personality: Evidence From Active Listening and Facebook Likes. *Psychological Science*. doi: 10.1177/0956797618761659

Nunes, I., & Jannach, D. (2017, 12). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, *27*(3-5), 393–444. doi: 10.1007/s11257-017-9195-0

Parra, D., & Brusilovsky, P. (2015). User-controllable personalization: A case study with SetFusion. *International Journal of Human Computer Studies*, *78*, 43–67. doi: 10.1016/j.ijhcs.2015.01.007

Parra, D., Brusilovsky, P., & Trattner, C. (2014). See what you want to see. In *Proceedings of the 19th international conference on intelligent user interfaces - iui '14* (Vol. 11, pp. 235–240). New York, New York, USA: ACM Press. doi: 10.1145/2557500.2557542

Pu, P., Chen, L., & Hu, R. (2011). A user-centric evaluation framework for recommender systems. In (p. 157). doi: 10.1145/2043932.2043962

Rentfrow, P. J., Goldberg, L. R., Stillwell, D. J., Kosinski, M., Gosling, S. D., & Levitin, D. J. (2012). The Song Remains the Same: A Replication and Extension of the MUSIC Model. *Music Perception*, *30*(2), 161–185. doi: 10.1525/mp.2012.30.2.161

Rentfrow, P. J., & Gosling, S. D. (2007). The content and validity of music-genre stereotypes among college students. *Music and Psychology Research*, *352*(35).

Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender Systems Handbook* (2nd ed.). Boston, MA: Springer US. doi: 10.1007/978-1-4899-7637-6

Saito, Y., & Itoh, T. (2011). MusiCube: A Visual Music Recommendation System featuring Interactive Evolutionary Computing. In (pp. 1–6). doi: 10.1145/2016656.2016661

Schedl, M., Zamani, H., Chen, C. W., Deldjoo, Y., & Elahi, M. (2018). Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, *7*(2), 95–116. Retrieved from https://doi.org/10.1007/s13735-018-0154-2 doi: 10.1007/s13735-018-0154-2

Simon, H. A. (1996). *The Sciences of the Artificial* (3rd Editio ed.). London, England: MIT Press.

Spotify AB. (2019, 1). *Queen Will, Queen Will, Rock You (and Your Kids, and Their Kids . . .).* Retrieved from https://newsroom.spotify.com/2019-01-03/queen-will-queen-will-rock-you-and-your-kids-and-their-kids/

Spotify AB. (2020, 4). *Spotify Company Info.* Retrieved from https://newsroom.spotify.com/company-info/

Sutton, G. W. (2018). *Creating Surveys, Evaluating Programs and Reading Research.* Sunflower Press.

Swearingen, K., & Sinha, R. (2001). Beyond Algorithms : An HCI Perspective on Recommender Systems. In *Acm sigir 2001 workshop on recommender systems (2001)* (pp. 1–11).

Tintarev, N., & Masthoff, J. (2007). *A Survey of Explanations in Recommender Systems* (Tech. Rep.). doi: 10.1.1.418.9237

Tintarev, N., & Masthoff, J. (2011). Designing and Evaluating Explanations for Recommender Systems. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 479–510). Boston, MA: Springer US. doi: 10.1007/978-0-387-85820-3{_}15

Tintarev, N., & Masthoff, J. (2015). Explaining Recommendations: Design and Evaluation. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (Second ed., pp. 353–382). Boston, MA: Springer US. doi: 10.1007/978-1-4899-7637-6{_}10

Tsai, C.-H., & Brusilovsky, P. (2018). Beyond the Ranked List: User-Driven Exploration and Diversification of Social Recommendation. In *Proceedings of the 2018 conference on human information interaction & retrieval - iui '18* (pp. 239–250). New York, New York, USA: ACM Press. doi: 10.1145/3172944.3172959

Tsai, C.-H., & Brusilovsky, P. (2019). Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th international conference on intelligent user interfaces - iui '19* (pp. 391–396). doi: 10.1145/3301275.3302318

Tsai, C.-H., Rahdari, B., & Brusilovsky, P. (2019). Exploring user-controlled hybrid recommendation in conference contexts. In *Proceedings of the 24th international conference on intelligent user interfaces - iui '19.*

Vavrille, F., Castaignet, V., & Moreau, V. (n.d.). *Musicovery - Home.* Retrieved from https://www.facebook.com/Musicovery-101689336552492/

Verbert, K., Parra, D., Brusilovsky, P., & Duval, E. (2013). Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on intelligent user interfaces - iui '13* (p. 351). New York, New York, USA: ACM Press. doi: 10.1145/2449396.2449442

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions Evoked by the Sound of Music: Characterization, Classification, and Measurement. *Emotion*, *8*(4), 494–521. doi: 10.1037/1528-3542.8.4.494

Zhang, Y., & Chen, X. (2018, 4). Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends R in Information Retrieval*, 1–87.

Appendix

A Literature Review Additional Information

The academic digital databases used were ACM digital Library[18], Elsevier[19], Google Scholar[20], Oria[21], and ScienceDirect[22]. The scope of search terms included: *data visualization*, *mobile user interface design*, *mood representation*, *music information retrieval*, *music information visualization*, *music recommender systems*, *music stereotypes*, *recommendation explanations*, *recommender system evaluation guidelines*, *usability heuristics*, *user-centric evaluation*, *user experience*. Papers were organized into categories by associated keywords in the reference management software Mendeley[23] with the possibility of fitting into multiple categories. When a relevant article was found, any preceding or subsequent paper was also read in order to understand how the authors improved upon their original findings. If another paper was directly quoted in a relevant article, that was also read to verify the claims they made. As explained by Kitchenham & Brereton (2013), this is know as snowballing, or citation-based search. Older articles found via snowballing which relate to the generalities of explanations, mood, or recommender systems for example, were deemed beneficial, while older articles relating to visualizations or interface design were seen as outdated. Many researchers also referenced commercial streaming services or non-academic music recommender systems research. Therefore, these types of works were also allowed to be examined if they were already mentioned in academic works in order to comprehend the current state-of-art in this industry as a whole. Two commercial streaming service assessments which were included in this thesis but not included in Table 3 are Hu (2019) and Vavrille et al. (n.d.).

There will always be limitations to completing a literature review as it is in often the very nature of research to be conducted differently. Almost none of the papers read were repeated experiments, but new ones with new findings, making it hard to do proper comparisons between papers. Table 3 only presents the key aspects of research which are most relevant to this thesis. There are numerous components to every recommender system and as such, each research paper discussed focused on different aspects and different goals to be achieved. Some papers tested a new interface, some compared this new interface to a baseline, and others compared several different interfaces or visualization styles. Additional factors such as different domain or different

[18] https://dl.acm.org/

[19] https://www.elsevier.com/

[20] https://scholar.google.com/

[21] https://uib.oria.no/

[22] https://www.sciencedirect.com/

[23] https://www.mendeley.com/

user testing techniques also have to be taken into consideration when drawing comparisons. During the course of completing this background research, a cornucopia of other quasi related data was uncovered which was not mentioned in Section 2. Certain academic and non-academic works were flagged as semi-relevant due to them being either too old and therefore no longer pertinent, there being no user studies or user studies with too few participants to have a statistical significance, or since their evaluation and results were not considered to be useful enough. The other music recommendation systems which were also taken into consideration when developing this study, though excluded from this thesis, are: BBC labs Radio Waves[24], Cardoso et al.'s *"MOODetector: A Prototype Software Tool for Mood-based Playlist Generation"*[25], Chen and Butz's *"MusicSim: Integrating Audio Analysis and User Feedback in an Interactive Music Browsing UI"*[26], Dalhuijsen's MusicalNodes thesis *"A graphical representation of digital music libraries using relational and absolute data to rediscover your collection"*[27], Esswein et al.'s *"geMsearch: Personalized Explorative Music Search"*[28], Gulik and Vignoli's *"Visual Playlist Generation on the Artist Map"*[29], Hides et al.'s *"Efficacy and Outcomes of a Music-Based Emotion Regulation Mobile App in Distressed Young People: Randomized Controlled Trial"*[30], Hilliges et al.'s *"AudioRadar: A Metaphorical Visualization for the Navigation of Large Music Collections"*[31], Jentsch's graduation project *"The Consumption of Music in the Era of iTunes"*[32], Jin et al.'s *"Effects of personal characteristics in control-oriented user interfaces for music recommender systems"*[33], Kangasraasio et al.'s *"Improving Controllability and Predictability of Interactive Recommendation Interfaces for Exploratory Search"*[34], Katarya et al.'s *"An Interactive Interface for Instilling Trust and providing Diverse Recommendations"*[35], Lehtiniemi and Ojala's *"Evaluating MoodPic – A Concept for Collaborative Mood Music Playlist Creation"*[36], Lehtiniemi and Ojala's *"Using Adaptive Avatars for Visualizing Recent Music"*[37],

[24] https://www.bbc.co.uk/blogs/radiolabs/2009/10/bbc_radio_waves_visualising_mu.shtml

[25] https://www.researchgate.net/publication/257307898

[26] https://dl.acm.org/doi/pdf/10.1145/1502650.1502713

[27] https://www.studiowith.nl/uploads/MusicalNodes_-_LisaDalhuijsen_20110620.pdf

[28] http://ceur-ws.org/Vol-2068/milc2.pdf

[29] http://ismir2005.ismir.net/proceedings/2011.pdf

[30] https://mhealth.jmir.org/2019/1/e11482/

[31] https://link.springer.com/chapter/10.1007/11795018_8

[32] http://www.formater.de/

[33] http://doi.org/10.1007/s11257-019-09247-2

[34] https://dl.acm.org/doi/pdf/10.1145/2678025.2701371

[35] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7001463

[36] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6676547

[37] https://dl.acm.org/doi/pdf/10.1145/2663806.2663820

Moodbar[38], Mufin (MUsic FINder) player[39], MusicMap[40], Nakano et al.'s *"PlaylistPlayer: An Interface Using Multiple Criteria to Change the Playback Order of a Music Playlist"*[41], Onyro's TuneGlue[42], Pampalk and Goto's *"MusicRainbow: A New User Interface to Discover Artists Using Audio-based Similarity and Web-based Labeling"*[43], Pampalk and Goto's *"Musicsun: A New Approach to Artist Recommendation"*[44], Vad et al.'s *"Design and Evaluation of a Probabilistic Music Projection Interface"*[45], Vallandingham's Track Tag Love[46], Welten's master's thesis *"Personalized Organization of Music on Mobile Devices"*[47], and Yang et al.'s *"Mr. Emo: Music retrieval in the emotion plane"*[48]. Going into detail about these systems was not within the scope of this paper. Further research should explore the systems mentioned above. The study by Hides et al. is of particular interest as it relates directly to mood based music recommendations in a mobile setting, though has no direct ties to recommendation explanations. Unfortunately the paper was published after the literature review had already been completed and there was not enough time to include it.

[38] https://github.com/exaile/moodbar

[39] https://smartphones.gadgethacks.com/news/mapping-your-music-collection-with-mufin-player-for-android-0128924/

[40] https://www.gnod.com/

[41] https://staff.aist.go.jp/m.goto/PAPER/IUI2016nakano.pdf

[42] https://onyro.com/tuneglue-audiomaptunegluenet

[43] https://staff.aist.go.jp/m.goto/MusicRainbow/pam_ismir06_music_rainbow.pdf

[44] http://pampalk.at/publications/pampalk_ismir07_music_sun.pdf

[45] http://www.dcs.gla.ac.uk/~rod/publications/VadBolMur15.pdf

[46] http://vallandingham.me/msd/vis/

[47] https://pdfs.semanticscholar.org/2308/7805bbbede75bbcd933c66dc6849df40d04d.pdf

[48] https://www.researchgate.net/publication/221573849

B SurveyXact Surveys

The three surveys executed on Amazon Mechanical Turk were created through SurveyXact. All surveys were identical apart from the playlists shown. The following includes the entire questionnaire from the first MTurk survey, links to the YouTube videos which show the design interactions for each interface, and additional pictures from surveys two and three.

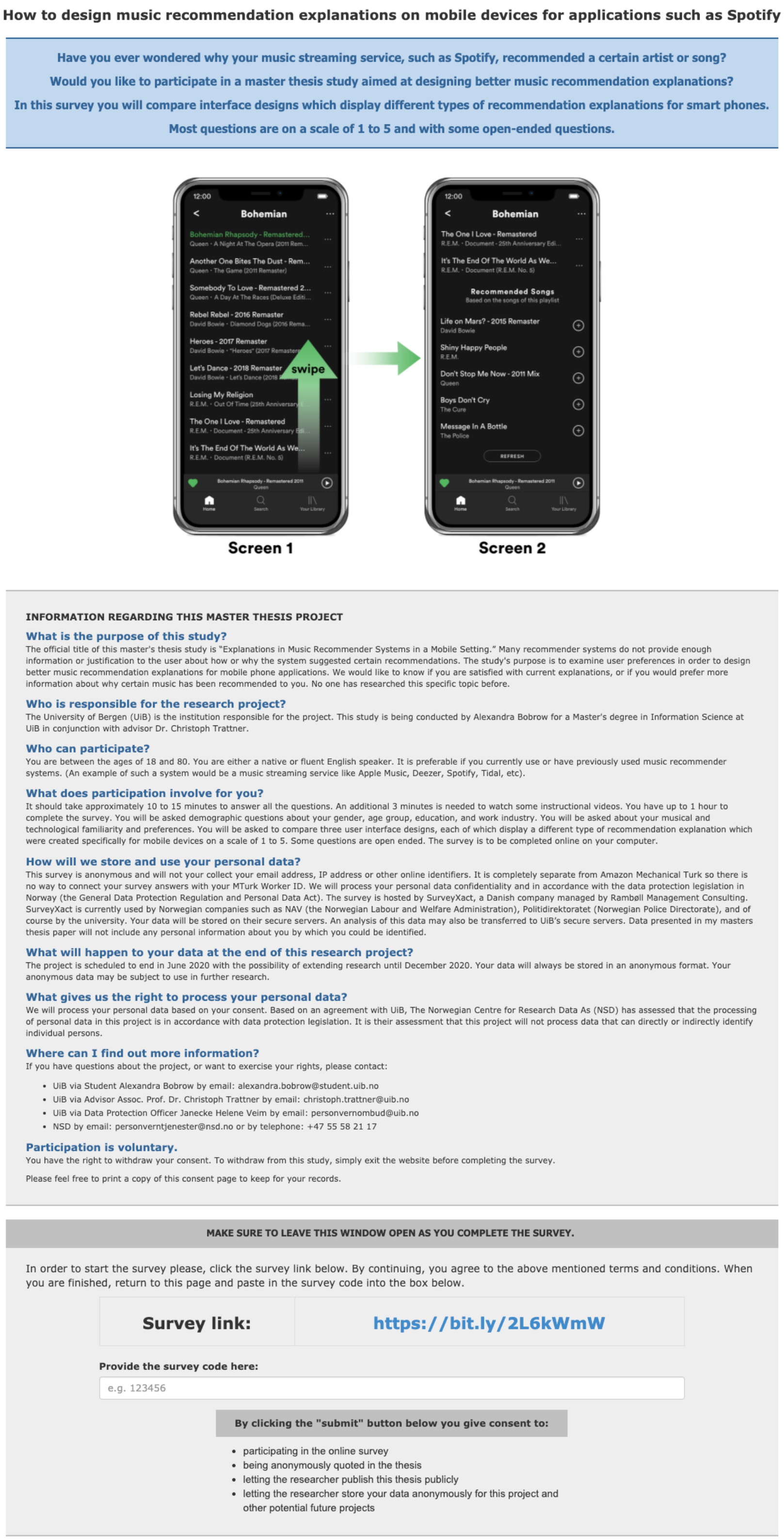
B.1 Links to YouTube Interface Videos

Survey 1 Bohemian Playlist
Baseline:
https://www.youtube.com/watch?v=mvMtqFIvrhk&list=UUX0748cn_BQsmuaMAhYEUiA
Text:
https://www.youtube.com/watch?v=HVNdb4L28Zk&list=UUX0748cn_BQsmuaMAhYEUiA
Visual:
https://www.youtube.com/watch?v=rtUCigccCBk&list=UUX0748cn_BQsmuaMAhYEUiA

Survey 2 Sunrise Playlist
Baseline:
https://www.youtube.com/watch?v=buox2jI1WOw&list=UUX0748cn_BQsmuaMAhYEUiA
Text:
https://www.youtube.com/watch?v=LzwIK8O_K0g&list=UUX0748cn_BQsmuaMAhYEUiA
Visual:
https://www.youtube.com/watch?v=2TM7K9iPEPM&list=UUX0748cn_BQsmuaMAhYEUiA

Survey 3 Get Up Playlist
Baseline:
https://www.youtube.com/watch?v=o-ZFtr8DH2o&list=UUX0748cn_BQsmuaMAhYEUiA
Text:
https://www.youtube.com/watch?v=dWxhkcCZ33Y&list=UUX0748cn_BQsmuaMAhYEUiA
Visual:
https://www.youtube.com/watch?v=9PnWnZ7Uo5A&list=UUX0748cn_BQsmuaMAhYEUiA

B.2 MTurk Survey Task Announcement



How to design music recommendation explanations on mobile devices for applications such as Spotify

Have you ever wondered why your music streaming service, such as Spotify, recommended a certain artist or song?
Would you like to participate in a master thesis study aimed at designing better music recommendation explanations?
In this survey you will compare interface designs which display different types of recommendation explanations for smart phones.
Most questions are on a scale of 1 to 5 and with some open-ended questions.

12:00
Bohemian
Bohemian Rhapsody - Remastered...
Queen · A Night At The Opera (2011 Rem...
Another One Bites The Dust - Rem...
Queen · The Game (2011 Remaster)
Somebody To Love - Remastered 2...
Queen · A Day At The Races (Deluxe Editi...
Rebel Rebel - 2016 Remaster
David Bowie · Diamond Dogs (2016 Rema...
Heroes - 2017 Remaster
David Bowie · "Heroes" (2017 Remaster...
Let's Dance - 2018 Remaster
David Bowie · Let's Dance (2018 ...
Losing My Religion
R.E.M. · Out Of Time (25th Anniversary E...
The One I Love - Remastered
R.E.M. · Document - 25th Anniversary Edi...
It's The End Of The World As We...
R.E.M. · Document (R.E.M. No. 5)
Bohemian Rhapsody - Remastered 2011
Queen

Screen 1

12:00
Bohemian
The One I Love - Remastered
R.E.M. · Document - 25th Anniversary Edi...
It's The End Of The World As We...
R.E.M. · Document (R.E.M. No. 5)
Recommended Songs
Based on the songs of this playlist
Life on Mars? - 2015 Remaster
David Bowie
Shiny Happy People
R.E.M.
Don't Stop Me Now - 2011 Mix
Queen
Boys Don't Cry
The Cure
Message In A Bottle
The Police
Bohemian Rhapsody - Remastered 2011
Queen

Screen 2

INFORMATION REGARDING THIS MASTER THESIS PROJECT

What is the purpose of this study?

The official title of this master's thesis study is "Explanations in Music Recommender Systems in a Mobile Setting." Many recommender systems do not provide enough information or justification to the user about how or why the system suggested certain recommendations. The study's purpose is to examine user preferences in order to design better music recommendation explanations for mobile phone applications. We would like to know if you are satisfied with current explanations, or if you would prefer more information about why certain music has been recommended to you. No one has researched this specific topic before.

Who is responsible for the research project?

The University of Bergen (UiB) is the institution responsible for the project. This study is being conducted by Alexandra Bobrow for a Master's degree in Information Science at UiB in conjunction with advisor Dr. Christoph Trattner.

Who can participate?

You are between the ages of 18 and 80. You are either a native or fluent English speaker. It is preferable if you currently use or have previously used music recommender systems. (An example of such a system would be a music streaming service like Apple Music, Deezer, Spotify, Tidal, etc).

What does participation involve for you?

It should take approximately 10 to 15 minutes to answer all the questions. An additional 3 minutes is needed to watch some instructional videos. You have up to 1 hour to complete the survey. You will be asked demographic questions about your gender, age group, education, and work industry. You will be asked about your musical and technological familiarity and preferences. You will be asked to compare three user interface designs, each of which display a different type of recommendation explanation which were created specifically for mobile devices on a scale of 1 to 5. Some questions are open ended. The survey is to be completed online on your computer.

How will we store and use your personal data?

This survey is anonymous and will not your collect your email address, IP address or other online identifiers. It is completely separate from Amazon Mechanical Turk so there is no way to connect your survey answers with your MTurk Worker ID. We will process your personal data confidentiality and in accordance with the data protection legislation in Norway (the General Data Protection Regulation and Personal Data Act). The survey is hosted by SurveyXact, a Danish company managed by Rambøll Management Consulting. SurveyXact is currently used by Norwegian companies such as NAV (the Norwegian Labour and Welfare Administration), Politidirektoratet (Norwegian Police Directorate), and of course by the university. Your data will be stored on their secure servers. An analysis of this data may also be transferred to UiB's secure servers. Data presented in my masters thesis paper will not include any personal information about you by which you could be identified.

What will happen to your data at the end of this research project?

The project is scheduled to end in June 2020 with the possibility of extending research until December 2020. Your data will always be stored in an anonymous format. Your anonymous data may be subject to use in further research.

What gives us the right to process your personal data?

We will process your personal data based on your consent. Based on an agreement with UiB, The Norwegian Centre for Research Data As (NSD) has assessed that the processing of personal data in this project is in accordance with data protection legislation. It is their assessment that this project will not process data that can directly or indirectly identify individual persons.

Where can I find out more information?

If you have questions about the project, or want to exercise your rights, please contact:

- UiB via Student Alexandra Bobrow by email: alexandra.bobrow@student.uib.no
- UiB via Advisor Assoc. Prof. Dr. Christoph Trattner by email: christoph.trattner@uib.no
- UiB via Data Protection Officer Janecke Helene Veim by email: personvernombud@uib.no
- NSD by email: personverntjenester@nsd.no or by telephone: +47 55 58 21 17

Participation is voluntary.

You have the right to withdraw your consent. To withdraw from this study, simply exit the website before completing the survey.

Please feel free to print a copy of this consent page to keep for your records.

MAKE SURE TO LEAVE THIS WINDOW OPEN AS YOU COMPLETE THE SURVEY.

In order to start the survey please, click the survey link below. By continuing, you agree to the above mentioned terms and conditions. When you are finished, return to this page and paste in the survey code into the box below.

Survey link:	**https://bit.ly/2L6kWmW**

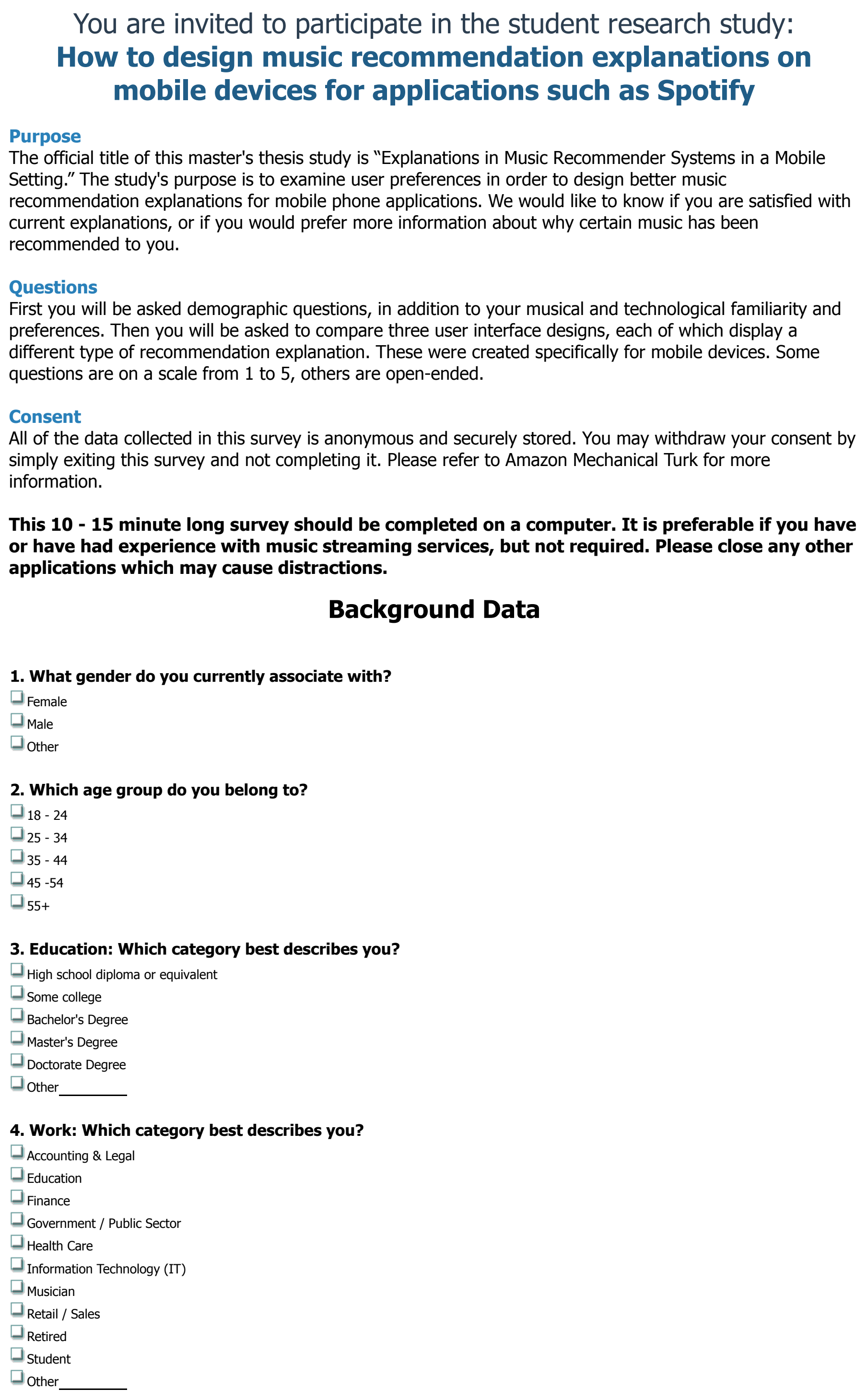
Provide the survey code here:

e.g. 123456

By clicking the "submit" button below you give consent to:

- participating in the online survey
- being anonymously quoted in the thesis
- letting the researcher publish this thesis publicly
- letting the researcher store your data anonymously for this project and other potential future projects

B.3 MTurk Survey 1 Example (Bohemian Playlist)



You are invited to participate in the student research study:
How to design music recommendation explanations on mobile devices for applications such as Spotify

Purpose

The official title of this master's thesis study is "Explanations in Music Recommender Systems in a Mobile Setting." The study's purpose is to examine user preferences in order to design better music recommendation explanations for mobile phone applications. We would like to know if you are satisfied with current explanations, or if you would prefer more information about why certain music has been recommended to you.

Questions

First you will be asked demographic questions, in addition to your musical and technological familiarity and preferences. Then you will be asked to compare three user interface designs, each of which display a different type of recommendation explanation. These were created specifically for mobile devices. Some questions are on a scale from 1 to 5, others are open-ended.

Consent

All of the data collected in this survey is anonymous and securely stored. You may withdraw your consent by simply exiting this survey and not completing it. Please refer to Amazon Mechanical Turk for more information.

This 10 - 15 minute long survey should be completed on a computer. It is preferable if you have or have had experience with music streaming services, but not required. Please close any other applications which may cause distractions.

Background Data

1. What gender do you currently associate with?

□ Female
□ Male
□ Other

2. Which age group do you belong to?

□ 18 - 24
□ 25 - 34
□ 35 - 44
□ 45 -54
□ 55+

3. Education: Which category best describes you?

□ High school diploma or equivalent
□ Some college
□ Bachelor's Degree
□ Master's Degree
□ Doctorate Degree
□ Other_________

4. Work: Which category best describes you?

□ Accounting & Legal
□ Education
□ Finance
□ Government / Public Sector
□ Health Care
□ Information Technology (IT)
□ Musician
□ Retail / Sales
□ Retired
□ Student
□ Other_________

Music and Technology Preferences

1. Please choose whether you agree or disagree with the following statements on a scale of 1 - 5, where 1 means strongly disagree and 5 means strongly agree.

	(Strongly disagree)	(Disagree)	(Neutral)	(Agree)	(Strongly agree)
	1	2	3	4	5
I am familiar with recommender systems.	□	□	□	□	□
The thought of using unfamiliar digital technology is uncomfortable.	□	□	□	□	□
I am familiar with Spotify.	□	□	□	□	□
I am confident about my ability to use digital technologies.	□	□	□	□	□
I am familiar with Spotify's recommendation features.	□	□	□	□	□

2. Do you currently use a music streaming service such as Apple Music, Deezer, Spotify, Tidal etc.?

□ Yes, I am currently using a music streaming service.

□ No, but I used to use a music streaming service.

□ No, I have never used a music streaming service before.

3. How often do you use your smartphone to listen to music?

□ 5 or more hours per day

□ 3 - 4 hours per day

□ 1 - 2 hours per day

□ Weekly

□ Monthly

□ Never

□ Other_________

4. What are your favorite music genres? You may choose up to 3.

□ Blues

□ Classical / Opera

□ Country

□ Dance / Electronic / House

□ Easy Listening

□ Folk

□ Heavy Metal

□ Hip Hop / Rap

□ Jazz

□ Latin

□ Pop

□ Reggae

□ R&B / Soul

□ Rock

□ Other_________

Music Recommendation Preferences

You previously stated that you currently use a music streaming service. Please answer these questions based on your experience with your current music streaming service.

Please choose whether you agree or disagree with the following statements on a scale of 1 - 5, where 1 means strongly disagree and 5 means strongly agree.

	(Strongly disagree) 1	(Disagree) 2	(Neutral) 3	(Agree) 4	(Strongly agree) 5
I would like more control over my music recommendations.	□	□	□	□	□
I understand the music recommendations that are suggested to me.	□	□	□	□	□
I would like better explanations for why certain music is recommended to me.	□	□	□	□	□
I often listen to the music recommendations that are suggested to me.	□	□	□	□	□

You previously stated that you used to use a music streaming service. Please answer these questions based on your previous experience with music streaming service to the best of your recollection.

Please choose whether you agree or disagree with the following statements on a scale of 1 - 5, where 1 means strongly disagree and 5 means strongly agree.

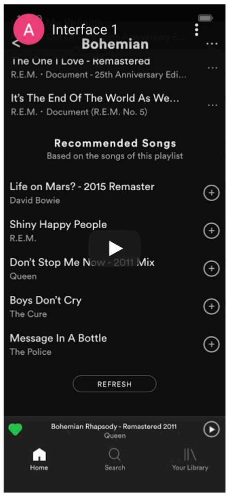
	(Strongly disagree) 1	(Disagree) 2	(Neutral) 3	(Agree) 4	(Strongly agree) 5
I understood the music recommendations that were suggested to me.	□	□	□	□	□
I would have liked more control over my music recommendations.	□	□	□	□	□
I often used to listen to the music recommendations that were suggested to me.	□	□	□	□	□
I would have liked better explanations for why certain music was recommended to me.	□	□	□	□	□

Recommendation Explanation Data: What types of additional information would you like to have in order to better understand why certain music was recommended to you? Choose as many as you'd like.

□ Acousticness (how acoustic or electronic the music is, acoustic/analog vs electric/digital instruments)
□ Danceability (how suitable the music is for dancing)
□ Energy (how energetic / fast / loud a song is)
□ Emoji / Emoticons (describes music based on emotional icons)
□ Genre information (rock, pop, blues...)
□ Loudness (musical dynamic, the overall loudness of music in decibels (dB))
□ Moods (what state of mind the music will put you in, how it makes you feel)
□ Popularity Score (how popular the music is)
□ Rhythm (musical feel)
□ Similarity Score (percentage of how alike different music is to each other)
□ Tags (musical keywords, terms, or annotations)
□ Tempo (beats per minute, if the music is fast or slow)
□ Timbre (tone color or tone quality)
□ Valence / Arousal (how positive or negative the music is)
□ I don't need any more explanations
□ I don't know
□ Other

Interface Comparison Section

The next few questions ask you to analyze different interface designs. All images are screenshots taken from a mobile phone application. This survey is not interactive and only contains static images. You are not able to interact with the images in any way. The following three videos show how one would interact with the mobile application. This should you give a better idea of how the application works. You only need to watch the first 30-45 seconds of each video.



Interface 1

Bohemian

The One I Love - Remastered
R.E.M. · Document - 25th Anniversary Edi...

It's The End Of The World As We...
R.E.M. · Document (R.E.M. No. 5)

Recommended Songs
Based on the songs of this playlist

Life on Mars? - 2015 Remaster
David Bowie

Shiny Happy People
R.E.M.

Don't Stop Me Now - 2011 Mix
Queen

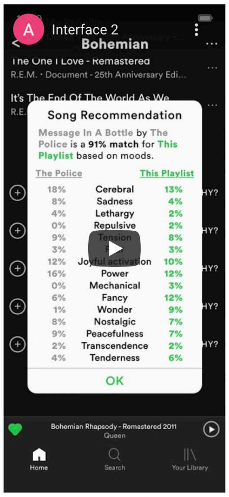
Boys Don't Cry
The Cure

Message In A Bottle
The Police

REFRESH

Bohemian Rhapsody - Remastered 2011
Queen

Interface 1



Interface 2

Bohemian

Song Recommendation

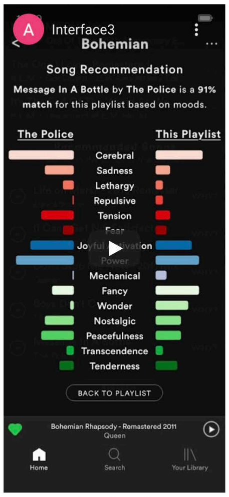
Message In A Bottle by The Police is a 91% match for This Playlist based on moods.

The Police		This Playlist
18%	Cerebral	13%
8%	Sadness	4%
4%	Lethargy	2%
0%	Repulsive	2%
9%	Tension	8%
3%	Fear	3%
12%	Joyful activation	10%
16%	Power	12%
0%	Mechanical	3%
6%	Fancy	12%
1%	Wonder	9%
8%	Nostalgic	7%
9%	Peacefulness	7%
2%	Transcendence	2%
4%	Tenderness	6%

OK

Bohemian Rhapsody - Remastered 2011
Queen

Interface 2



Interface3

Bohemian

Song Recommendation

Message In A Bottle by The Police is a 91% match for this playlist based on moods.

BACK TO PLAYLIST

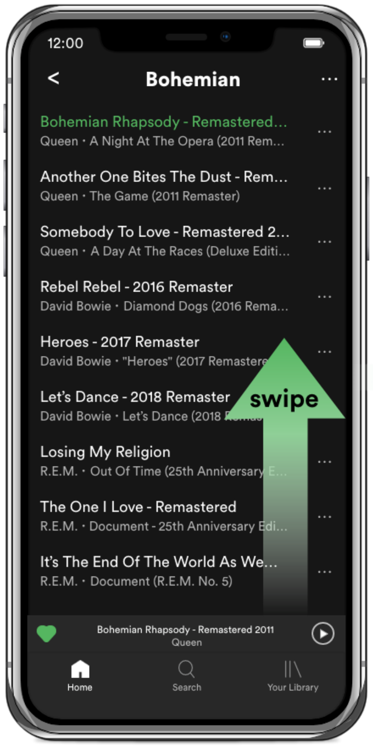
Bohemian Rhapsody - Remastered 2011
Queen

Interface 3

INTERFACE 1

This is an example of a present-day music recommender system, recommending songs for a playlist.

Now imagine yourself listening to a playlist you created called Bohemian on your smart phone.
You would like to add more songs to this playlist.
You notice the system provides a list of recommended songs at the bottom of the playlist.
You also see an explanation for why these songs were recommended.



12:00

< Bohemian ...

Bohemian Rhapsody - Remastered...
Queen · A Night At The Opera (2011 Rem...

Another One Bites The Dust - Rem...
Queen · The Game (2011 Remaster)

Somebody To Love - Remastered 2...
Queen · A Day At The Races (Deluxe Editi...

Rebel Rebel - 2016 Remaster
David Bowie · Diamond Dogs (2016 Rema...

Heroes - 2017 Remaster
David Bowie · "Heroes" (2017 Remastere...

Let's Dance - 2018 Remaster
David Bowie · Let's Dance (2018 Remas...

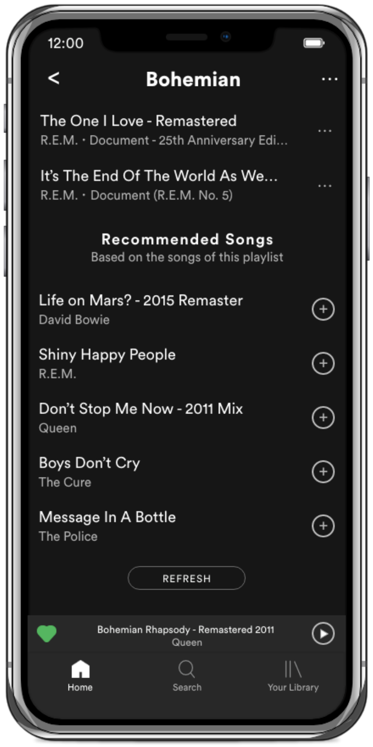
Losing My Religion
R.E.M. · Out Of Time (25th Anniversary E...

The One I Love - Remastered
R.E.M. · Document - 25th Anniversary Edi...

It's The End Of The World As We...
R.E.M. · Document (R.E.M. No. 5)

Bohemian Rhapsody - Remastered 2011
Queen

Screen 1



12:00

< Bohemian ...

The One I Love - Remastered
R.E.M. · Document - 25th Anniversary Edi...

It's The End Of The World As We...
R.E.M. · Document (R.E.M. No. 5)

Recommended Songs
Based on the songs of this playlist

Life on Mars? - 2015 Remaster
David Bowie

Shiny Happy People
R.E.M.

Don't Stop Me Now - 2011 Mix
Queen

Boys Don't Cry
The Cure

Message In A Bottle
The Police

Bohemian Rhapsody - Remastered 2011
Queen

Screen 2

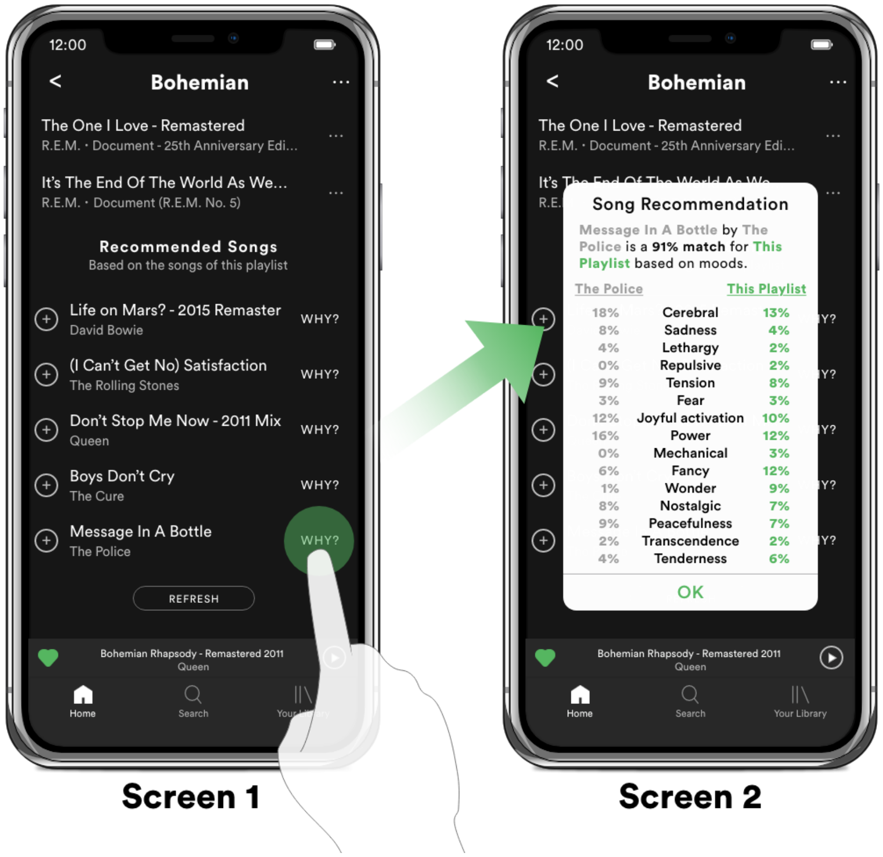
Please choose whether you agree or disagree with the following statements on a scale of 1 - 5, where 1 means strongly disagree and 5 means strongly agree.

	(Strongly disagree) 1	(Disagree) 2	(Neutral) 3	(Agree) 4	(Strongly agree) 5
I do not like the design of the explanation	□	□	□	□	□
The explanation is too detailed	□	□	□	□	□
I understand why 'Message In A Bottle' by 'The Police' is recommended	□	□	□	□	□
The explanation increased my trust in the system	□	□	□	□	□
The explanation is attractive and pleasing to look at	□	□	□	□	□
The explanation is unnecessarily complex	□	□	□	□	□
I would like to use a system which uses this type of explanation	□	□	□	□	□
It took me a long time to understand the explanation	□	□	□	□	□
I am familiar with these artists / songs	□	□	□	□	□
The explanation does not provide useful information	□	□	□	□	□
The explanation contains the right amount of information	□	□	□	□	□
The design layout of the explanation is clear and easy to understand	□	□	□	□	□
Overall, I am satisfied with the recommendation explanation	□	□	□	□	□

INTERFACE 2

This is another version of the same recommender system, recommending songs for a playlist.

Now imagine yourself listening to a playlist you created called Bohemian on your smart phone.
You would like to add more songs to this playlist.
You notice the system provides a list of recommended songs at the bottom of the playlist.
You also see an explanation for why these songs were recommended.



The One I Love - Remastered
R.E.M. • Document - 25th Anniversary Edi...

It's The End Of The World As We...
R.E.M. • Document (R.E.M. No. 5)

Recommended Songs
Based on the songs of this playlist

Life on Mars? - 2015 Remaster
David Bowie

(I Can't Get No) Satisfaction
The Rolling Stones

Don't Stop Me Now - 2011 Mix
Queen

Boys Don't Cry
The Cure

Message In A Bottle
The Police

Song Recommendation

Message In A Bottle by The Police is a 91% match for This Playlist based on moods.

The Police		This Playlist
18%	Cerebral	13%
8%	Sadness	4%
4%	Lethargy	2%
0%	Repulsive	2%
9%	Tension	8%
3%	Fear	3%
12%	Joyful activation	10%
16%	Power	12%
0%	Mechanical	3%
6%	Fancy	12%
1%	Wonder	9%
8%	Nostalgic	7%
9%	Peacefulness	7%
2%	Transcendence	2%
4%	Tenderness	6%

Screen 1 Screen 2

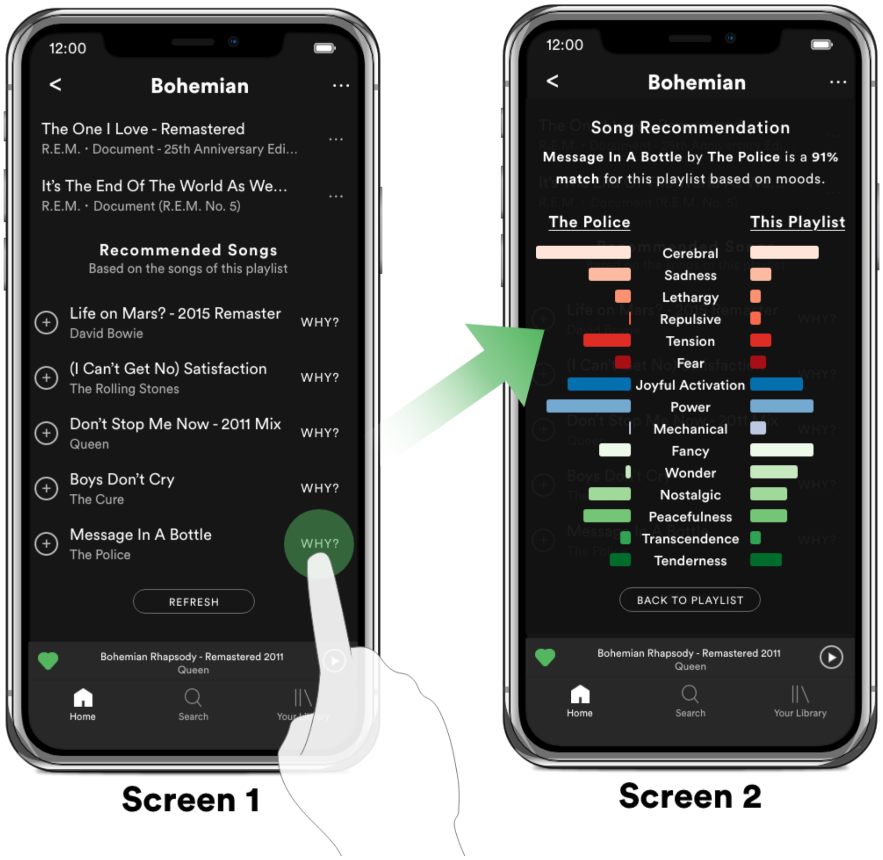
Please choose whether you agree or disagree with the following statements on a scale of 1 - 5, where 1 means strongly disagree and 5 means strongly agree.

	(Strongly disagree) 1	(Disagree) 2	(Neutral) 3	(Agree) 4	(Strongly agree) 5
The design layout of the explanation is clear and easy to understand	□	□	□	□	□
I will choose the neutral answer number three for this question	□	□	□	□	□
The explanation does not provide useful information	□	□	□	□	□
I do not like the design of the explanation	□	□	□	□	□
The explanation is unnecessarily complex	□	□	□	□	□
The explanation increased my trust in the system	□	□	□	□	□
The explanation is attractive and pleasing to look at	□	□	□	□	□
I would like to use a system which uses this type of explanation	□	□	□	□	□
The explanation contains the right amount of information	□	□	□	□	□
It took me a long time to understand the explanation	□	□	□	□	□
I understand why 'Message In A Bottle' by 'The Police' is recommended	□	□	□	□	□
The explanation is too detailed	□	□	□	□	□
I am familiar with these artists / songs	□	□	□	□	□
Overall, I am satisfied with the recommendation explanation	□	□	□	□	□

INTERFACE 3

This is another version of the same recommender system, recommending songs for a playlist.

Now imagine yourself listening to a playlist you created called Bohemian on your smart phone.
You would like to add more songs to this playlist.
You notice the system provides a list of recommended songs at the bottom of the playlist.
You also see an explanation for why these songs were recommended.



12:00

< Bohemian ···

The One I Love - Remastered
R.E.M. · Document - 25th Anniversary Edi...

It's The End Of The World As We...
R.E.M. · Document (R.E.M. No. 5)

Recommended Songs
Based on the songs of this playlist

Life on Mars? - 2015 Remaster
David Bowie
WHY?

(I Can't Get No) Satisfaction
The Rolling Stones
WHY?

Don't Stop Me Now - 2011 Mix
Queen
WHY?

Boys Don't Cry
The Cure
WHY?

Message In A Bottle
The Police
WHY?

REFRESH

Bohemian Rhapsody - Remastered 2011
Queen

Home Search Your Library

Screen 1

12:00

< Bohemian ···

Song Recommendation

Message In A Bottle by The Police is a 91% match for this playlist based on moods.

The Police This Playlist

Cerebral
Sadness
Lethargy
Repulsive
Tension
Fear
Joyful Activation
Power
Mechanical
Fancy
Wonder
Nostalgic
Peacefulness
Transcendence
Tenderness

BACK TO PLAYLIST

Bohemian Rhapsody - Remastered 2011
Queen

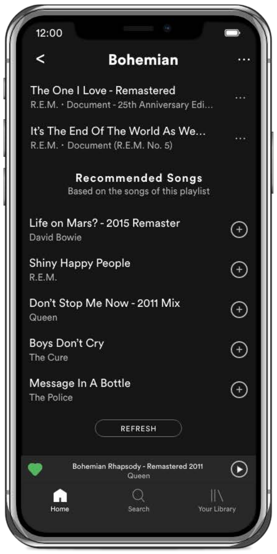
Home Search Your Library

Screen 2

Please choose whether you agree or disagree with the following statements on a scale of 1 - 5, where 1 means strongly disagree and 5 means strongly agree.

	(Strongly disagree) 1	(Disagree) 2	(Neutral) 3	(Agree) 4	(Strongly agree) 5
The explanation is too detailed	□	□	□	□	□
The design layout of the explanation is clear and easy to understand	□	□	□	□	□
I am familiar with these artists / songs	□	□	□	□	□
The explanation increased my trust in the system	□	□	□	□	□
I do not like the design of the explanation	□	□	□	□	□
The explanation does not provide useful information	□	□	□	□	□
I would like to use a system which uses this type of explanation	□	□	□	□	□
The answer to this question is disagree which is number two	□	□	□	□	□
I understand why 'Message In A Bottle' by 'The Police' is recommended	□	□	□	□	□
The explanation contains the right amount of information	□	□	□	□	□
The explanation is attractive and pleasing to look at	□	□	□	□	□
The explanation is unnecessarily complex	□	□	□	□	□
It took me a long time to understand the explanation	□	□	□	□	□
Overall, I am satisfied with the recommendation explanation	□	□	□	□	□

Please compare the different recommendation explanation designs shown in these three interfaces.



12:00

< **Bohemian** ···

The One I Love - Remastered
R.E.M. · Document - 25th Anniversary Edi... ···

It's The End Of The World As We...
R.E.M. · Document (R.E.M. No. 5) ···

Recommended Songs
Based on the songs of this playlist

Life on Mars? - 2015 Remaster
David Bowie ⊕

Shiny Happy People
R.E.M. ⊕

Don't Stop Me Now - 2011 Mix
Queen ⊕

Boys Don't Cry
The Cure ⊕

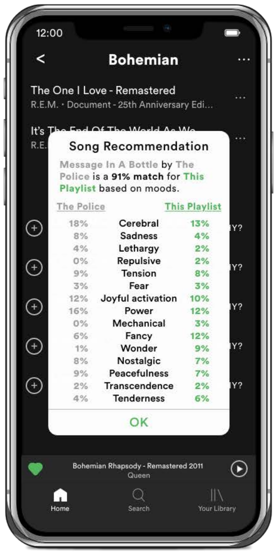
Message In A Bottle
The Police ⊕

REFRESH

Bohemian Rhapsody - Remastered 2011
Queen

Home Search Your Library

Interface 1



12:00

< **Bohemian** ···

The One I Love - Remastered
R.E.M. · Document - 25th Anniversary Edi... ···

It's The End Of The World As We...
R.E.

Song Recommendation

Message In A Bottle by The Police is a **91% match** for This Playlist based on moods.

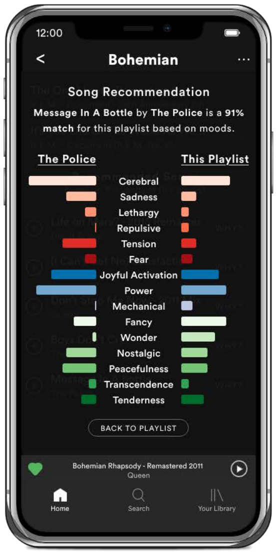
The Police		This Playlist
18%	Cerebral	13%
8%	Sadness	4%
4%	Lethargy	2%
0%	Repulsive	2%
9%	Tension	8%
3%	Fear	3%
12%	Joyful activation	10%
16%	Power	12%
0%	Mechanical	3%
6%	Fancy	12%
1%	Wonder	9%
8%	Nostalgic	7%
9%	Peacefulness	7%
2%	Transcendence	2%
4%	Tenderness	6%

OK

Bohemian Rhapsody - Remastered 2011
Queen

Home Search Your Library

Interface 2



12:00

< **Bohemian** ···

Song Recommendation

Message In A Bottle by **The Police** is a **91% match** for this playlist based on moods.

The Police **This Playlist**

BACK TO PLAYLIST

Bohemian Rhapsody - Remastered 2011
Queen

Home Search Your Library

Interface 3

1. Which recommendation explanation design do you like the most?

Interface 1 Interface 2 Interface 3

□ □ □

2. Why do you like this recommendation explanation design the most?

3. Which recommendationexplanation design do you dislike the most?

Interface 1 Interface 2 Interface 3

□ □ □

4. Why do you dislike this recommendation explanation design the most?

5. In general, what do you like most about the recommendation explanations shown above and what do you like the least?

6. Are there any design changes you would make to any of the interfaces shown above? (optional)

The mood data provided in Interface 2 and Interface 3 explains what state of mind the music will put you in after listening to it, or rather how the music will make you feel.

7. Were the mood-based recommendation explanations helpful?

8. In general, now that you have seen this survey, what do you think about using moods for recommendation explanations?

Post-Survey Section

1. Please state whether you agree or disagree with the following statements on a scale of 1-5, where 1 means strongly disagree and 5 means strongly agree.

	(Strongly disagree) 1	(Disagree) 2	(Neutral) 3	(Agree) 4	(Strongly agree) 5
Overall, I understood how to answer the questions in this survey.	□	□	□	□	□
Overall, I enjoyed taking this survey.	□	□	□	□	□
I would switch to or begin using a music streaming service on my phone if it had great recommendation explanations.	□	□	□	□	□

2. You were asked in the pre-survey whether or not music recommendation explanations are important to you. Has your opinion changed after participating in this study?

3. Do you have any other suggestions on how to improve recommendation explanations for music applications on mobile phones? (optional)

4. Please feel free to add any addition comments here. (optional)

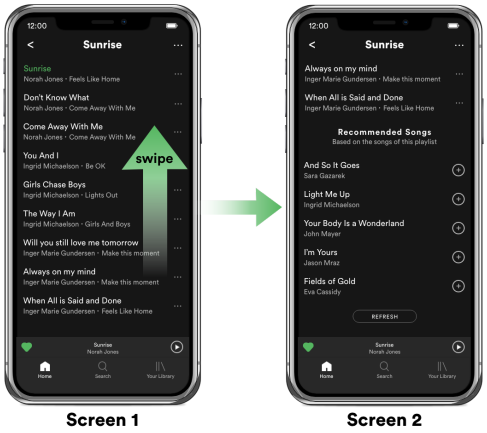
Thank you for participating in this student research project on how to design music recommendation explanations on mobile devices for applications such as Spotify

Please enter the following code into Amazon Mechanical Turk to get credit:

(code was shown here)

Upon clicking Finish you will be redirected to Amazon Mechanical Turk's homepage.

B.4 Pictures from Survey 2 (Sunrise playlist)



12:00

< Sunrise ...

Sunrise
Norah Jones · Feels Like Home

Don't Know What
Norah Jones · Come Away With Me

Come Away With Me
Norah Jones · Come Away With Me

You And I
Ingrid Michaelson · Be OK

Girls Chase Boys
Ingrid Michaelson · Lights Out

The Way I Am
Ingrid Michaelson · Girls And Boys

Will you still love me tomorrow
Inger Marie Gundersen · Make this moment

Always on my mind
Inger Marie Gundersen · Make this moment

When All is Said and Done
Inger Marie Gundersen · Feels Like Home

12:00

< Sunrise ...

Always on my mind
Inger Marie Gundersen · Make this moment

When All is Said and Done
Inger Marie Gundersen · Feels Like Home

Recommended Songs
Based on the songs of this playlist

And So It Goes
Sara Gazarek

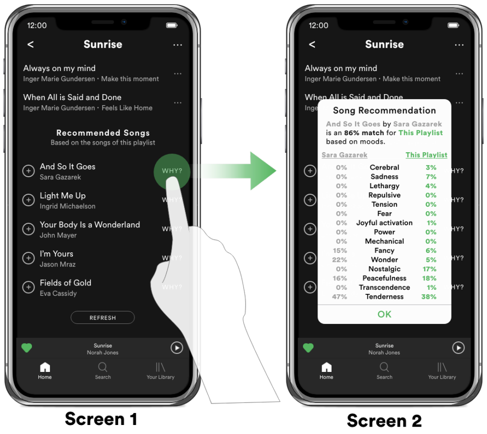
Light Me Up
Ingrid Michaelson

Your Body Is a Wonderland
John Mayer

I'm Yours
Jason Mraz

Fields of Gold
Eva Cassidy

(a) Baseline - Sunrise.



12:00

< Sunrise ...

Always on my mind
Inger Marie Gundersen · Make this moment

When All is Said and Done
Inger Marie Gundersen · Feels Like Home

Recommended Songs
Based on the songs of this playlist

And So It Goes
Sara Gazarek

Light Me Up
Ingrid Michaelson

Your Body Is a Wonderland
John Mayer

I'm Yours
Jason Mraz

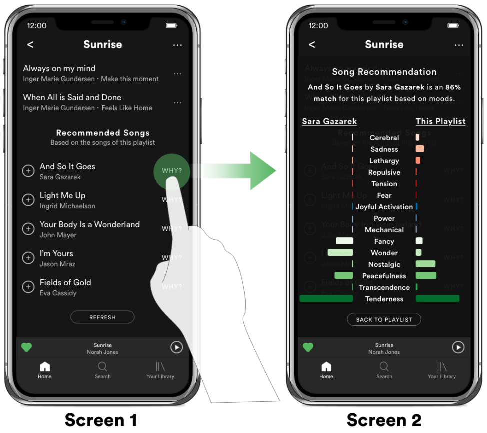
Fields of Gold
Eva Cassidy

Song Recommendation

And So It Goes by Sara Gazarek is an **86% match** for This Playlist based on moods.

Sara Gazarek		This Playlist
0%	Cerebral	3%
0%	Sadness	7%
0%	Lethargy	4%
0%	Repulsive	0%
0%	Tension	0%
0%	Fear	0%
0%	Joyful activation	1%
0%	Power	0%
0%	Mechanical	0%
15%	Fancy	6%
22%	Wonder	5%
0%	Nostalgic	17%
16%	Peacefulness	18%
0%	Transcendence	1%
47%	Tenderness	38%

(b) Textual Explanatory Interface - Sunrise.



12:00

< Sunrise ...

Always on my mind
Inger Marie Gundersen · Make this moment

When All is Said and Done
Inger Marie Gundersen · Feels Like Home

Recommended Songs
Based on the songs of this playlist

And So It Goes
Sara Gazarek

Light Me Up
Ingrid Michaelson

Your Body Is a Wonderland
John Mayer

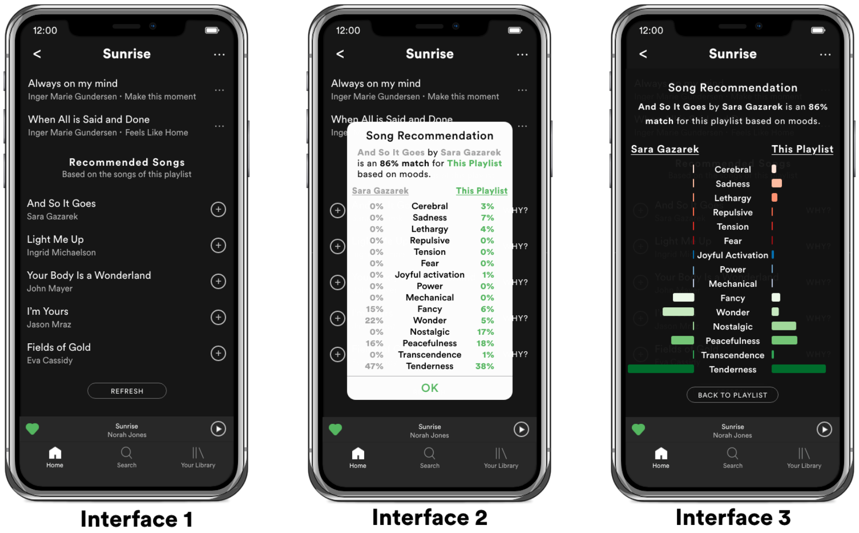
I'm Yours
Jason Mraz

Fields of Gold
Eva Cassidy

Song Recommendation

And So It Goes by **Sara Gazarek** is an **86% match** for this playlist based on moods.

(c) Visual Explanatory Interface - Sunrise.



Interface 1

12:00

< Sunrise ...

Always on my mind
Inger Marie Gundersen · Make this moment

When All is Said and Done
Inger Marie Gundersen · Feels Like Home

Recommended Songs
Based on the songs of this playlist

And So It Goes
Sara Gazarek

Light Me Up
Ingrid Michaelson

Your Body Is a Wonderland
John Mayer

I'm Yours
Jason Mraz

Fields of Gold
Eva Cassidy

REFRESH

Sunrise
Norah Jones

Home Search Your Library

Interface 2

12:00

< Sunrise ...

Always on my mind
Inger Marie Gundersen · Make this moment

When All is Said and Done

Song Recommendation

And So It Goes by Sara Gazarek is an **86% match** for This Playlist based on moods.

Sara Gazarek		This Playlist
0%	Cerebral	3%
0%	Sadness	7%
0%	Lethargy	4%
0%	Repulsive	0%
0%	Tension	0%
0%	Fear	0%
0%	Joyful activation	1%
0%	Power	0%
0%	Mechanical	0%
15%	Fancy	6%
22%	Wonder	5%
0%	Nostalgic	17%
16%	Peacefulness	18%
0%	Transcendence	1%
47%	Tenderness	38%

OK

Sunrise
Norah Jones

Home Search Your Library

Interface 3

12:00

< Sunrise ...

Song Recommendation

And So It Goes by **Sara Gazarek** is an **86% match** for this playlist based on moods.

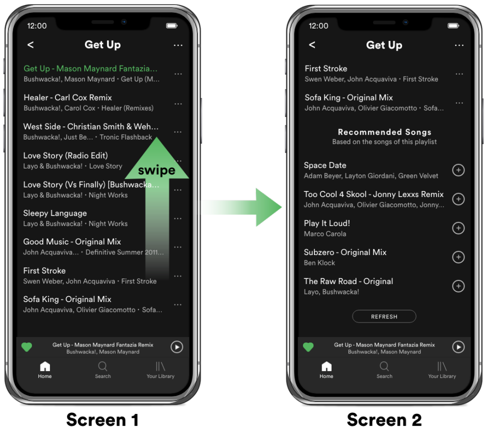
BACK TO PLAYLIST

Sunrise
Norah Jones

Home Search Your Library

Explanatory Interface Comparison - Sunrise.

B.5 Pictures from Survey 3 (Get Up playlist)



12:00

< Get Up …

Get Up - Mason Maynard Fantazia…
Bushwacka!, Mason Maynard · Get Up (M…

Healer - Carl Cox Remix
Bushwacka!, Carol Cox · Healer (Remixes)

West Side - Christian Smith & Weh…
Bushwacka!, Just Be… · Tronic Flashback

Love Story (Radio Edit)
Layo & Bushwacka! · Love Story

Love Story (Vs Finally) [Bushwacka…
Layo & Bushwacka! · Night Works

Sleepy Language
Layo & Bushwacka! · Night Works

Good Music - Original Mix
John Acquaviva… · Definitive Summer 2011…

First Stroke
Swen Weber, John Acquaviva · First Stroke

Sofa King - Original Mix
John Acquaviva, Olivier Giacomotto · Sofa…

Get Up - Mason Maynard Fantazia Remix
Bushwacka!, Mason Maynard

Home Search Your Library

12:00

< Get Up …

First Stroke
Swen Weber, John Acquaviva · First Stroke

Sofa King - Original Mix
John Acquaviva, Olivier Giacomotto · Sofa…

Recommended Songs
Based on the songs of this playlist

Space Date
Adam Beyer, Layton Giordani, Green Velvet

Too Cool 4 Skool - Jonny Lexxs Remix
John Acquaviva, Olivier Giacomotto, Jonny…

Play It Loud!
Marco Carola

Subzero - Original Mix
Ben Klock

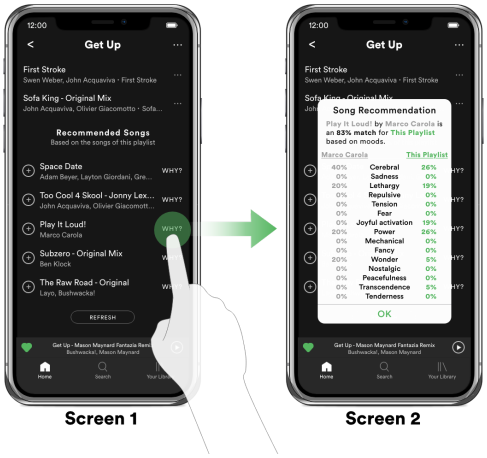
The Raw Road - Original
Layo, Bushwacka!

REFRESH

Get Up - Mason Maynard Fantazia Remix
Bushwacka!, Mason Maynard

Home Search Your Library

(a) Baseline - Get Up.



12:00

< Get Up …

First Stroke
Swen Weber, John Acquaviva · First Stroke

Sofa King - Original Mix
John Acquaviva, Olivier Giacomotto · Sofa…

Recommended Songs
Based on the songs of this playlist

⊕ Space Date
Adam Beyer, Layton Giordani, Gre… WHY?

⊕ Too Cool 4 Skool - Jonny Lex…
John Acquaviva, Olivier Giacomott… WHY?

⊕ Play It Loud!
Marco Carola WHY?

⊕ Subzero - Original Mix
Ben Klock WHY?

⊕ The Raw Road - Original
Layo, Bushwacka! WHY?

REFRESH

Get Up - Mason Maynard Fantazia Remix
Bushwacka!, Mason Maynard

Home Search Your Library

12:00

< Get Up …

Song Recommendation

Play It Loud! by Marco Carola is an **83% match** for This Playlist based on moods.

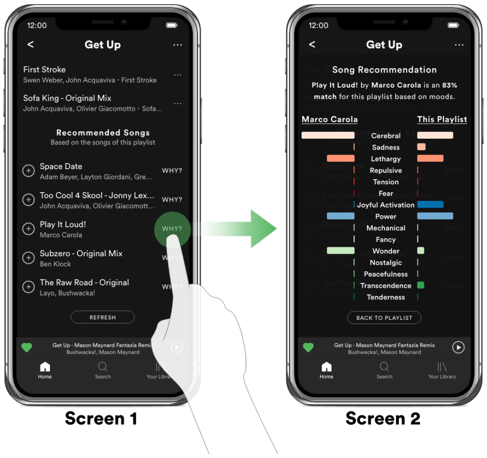
Marco Carola		This Playlist
40%	Cerebral	26%
0%	Sadness	0%
20%	Lethargy	19%
0%	Repulsive	0%
0%	Tension	0%
0%	Fear	0%
0%	Joyful activation	19%
20%	Power	26%
0%	Mechanical	0%
0%	Fancy	0%
20%	Wonder	5%
0%	Nostalgic	0%
0%	Peacefulness	0%
0%	Transcendence	5%
0%	Tenderness	0%

OK

Get Up - Mason Maynard Fantazia Remix
Bushwacka!, Mason Maynard

Home Search Your Library

(b) Textual Explanatory Interface - Get Up.



12:00

< Get Up …

First Stroke
Swen Weber, John Acquaviva · First Stroke

Sofa King - Original Mix
John Acquaviva, Olivier Giacomotto · Sofa…

Recommended Songs
Based on the songs of this playlist

⊕ Space Date
Adam Beyer, Layton Giordani, Gre… WHY?

⊕ Too Cool 4 Skool - Jonny Lex…
John Acquaviva, Olivier Giacomott… WHY?

⊕ Play It Loud!
Marco Carola WHY?

⊕ Subzero - Original Mix
Ben Klock WHY?

⊕ The Raw Road - Original
Layo, Bushwacka! WHY?

REFRESH

Get Up - Mason Maynard Fantazia Remix
Bushwacka!, Mason Maynard

Home Search Your Library

12:00

< Get Up …

Song Recommendation

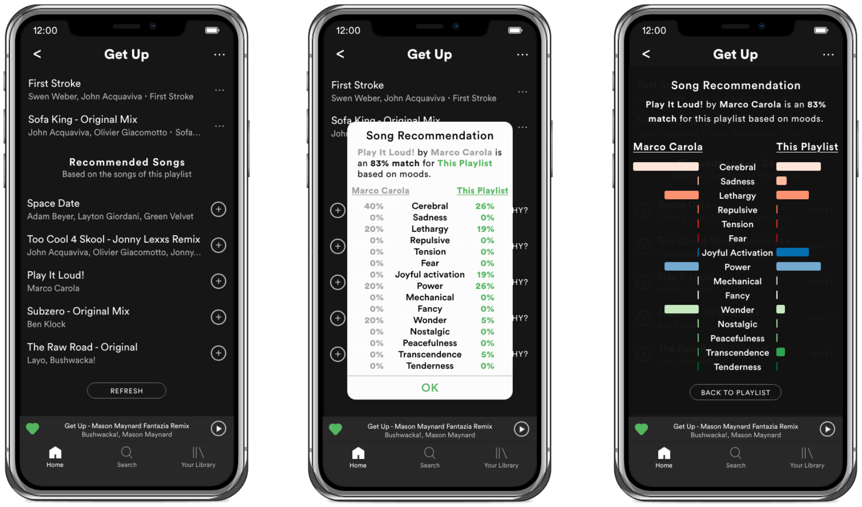
Play It Loud! by **Marco Carola** is an **83% match** for this playlist based on moods.

BACK TO PLAYLIST

Get Up - Mason Maynard Fantazia Remix
Bushwacka!, Mason Maynard

Home Search Your Library

(c) Visual Explanatory Interface - Get Up.



12:00

< Get Up ···

First Stroke
Swen Weber, John Acquaviva · First Stroke

Sofa King - Original Mix
John Acquaviva, Olivier Giacomotto · Sofa...

Recommended Songs
Based on the songs of this playlist

Space Date
Adam Beyer, Layton Giordani, Green Velvet

Too Cool 4 Skool - Jonny Lexxs Remix
John Acquaviva, Olivier Giacomotto, Jonny...

Play It Loud!
Marco Carola

Subzero - Original Mix
Ben Klock

The Raw Road - Original
Layo, Bushwacka!

REFRESH

Get Up - Mason Maynard Fantazia Remix
Bushwacka!, Mason Maynard

Home Search Your Library

Interface 1

12:00

< Get Up ···

First Stroke
Swen Weber, John Acquaviva · First Stroke

Song Recommendation

Play It Loud! by Marco Carola is an **83% match** for This Playlist based on moods.

Marco Carola		This Playlist
40%	Cerebral	26%
0%	Sadness	0%
20%	Lethargy	19%
0%	Repulsive	0%
0%	Tension	0%
0%	Fear	0%
0%	Joyful activation	19%
20%	Power	26%
0%	Mechanical	0%
0%	Fancy	0%
20%	Wonder	5%
0%	Nostalgic	0%
0%	Peacefulness	0%
0%	Transcendence	5%
0%	Tenderness	0%

OK

Get Up - Mason Maynard Fantazia Remix
Bushwacka!, Mason Maynard

Home Search Your Library

Interface 2

12:00

< Get Up ···

Song Recommendation

Play It Loud! by **Marco Carola** is an **83% match** for this playlist based on moods.

BACK TO PLAYLIST

Get Up - Mason Maynard Fantazia Remix
Bushwacka!, Mason Maynard

Home Search Your Library

Interface 3

Explanatory Interface Comparison - Get Up.